Handbook of Particle Detection and Imaging

Claus Grupen and Irène Buvat (Eds.)

Handbook of Particle Detection and Imaging

With 558 Figures and 72 Tables

Springer

Editors
Claus Grupen
Department of Physics
Siegen University
Emmy-Noether-Campus
Walter-Flex-Str. 3
57072 Siegen
Germany

Irène Buvat
IMNC-UMR 8165 CNRS
Paris 7 and Paris 11 Universities
Orsay Campus
Building 440
91406 Orsay Cedex
France

Library of Congress Control Number: 2011934762

ISBN 978-3-642-13270-4
DOI 10.1007/978-3-642-13271-1
This publication is also available as:
Electronic publication under ISBN 978-3-642-13271-1
Print and electronic bundle under ISBN 978-3-642-14621-3

Springer Heidelberg Dordrecht London New York

© Springer-Verlag Berlin Heidelberg 2012

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilm or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer. Violations are liable to prosecution under the German Copyright Law.
The use of general descriptive names, registered names, trademarks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

Springer is part of Springer Science+Business Media (www.springer.com)

Editor: Claus Ascheron, Heidelberg, Germany
Development Editor: Sylvia Blago, Heidelberg, Germany
Production/Cover Design: SPI Publishing, Pondicherry, India

Printed on acid-free paper

Acknowledgments

Claus Grupen would like to thank Professor Douglas McGregor, Ph.D., Kansas State University, for his contributions to the handbook in the initial phase of preliminary conception and planning. The professional help of Dr. Tilo Stroh, Siegen University, in questions of proofreading, layout and LaTeX handling is acknowledged.

Irène Buvat would like to thank all her students for never-ending stimulating discussions that partly motivated the involvement in co-editing this book. She is also deeply and variously indebted to Dr. Stephen Bacharach and Dr. Robert Di Paola who taught her so much.

Preface

Sophisticated instrumentations and imaging devices have become powerful tools in the modern world of technology. Advances in electronics, fast data processing, image reconstruction, and pattern recognition, just to name a few, have enabled the development of very elaborate investigation techniques that are now used in many different fields of science and many domains of applications. A large number of technological advances were originally developed in particle physics, but then spread to astrophysics, medicine, biology, materials science, art, archaeology, and many other application domains.

The field of imaging has known incredible advances in the recent decades. With atomic force microscopes or scanning force microscopes, very-high-resolution images can now be obtained. Resolutions on the order of fractions of a nanometer, more than 1000 times better than the optical diffraction limit have been reached. With the scanning tunneling microscope, atoms can even be dragged along and positioned to build atomic-scale artificial structures. Increasingly higher resolutions require more and more storage space. A Triumph calculator from the fifties in the last century using ferrite cores had a storage capacity of 32 bits! The Large Hadron Collider (LHC) at CERN will produce roughly 15 petabytes (15 million gigabytes) of data annually. The discovery of a giant magnetoresistance by the 2007 Nobel laureates Peter Grünberg and Albert Fert enabled a breakthrough in gigabyte hard disk drives, and so this technique allows massive number crunching and can cope with huge data files.

Also the image quality and image reconstruction accuracy have substantially increased. For a decent photo one needs more than 100 million photons in the optical range. In observational astronomy the sensitivity of instruments has improved over a period of 380 years – from Galileo's telescope to the Hubble Space Telescope – by a factor of 100 millions. For the discovery of an X-ray source in our galaxy, approximately 100 photons are sufficient. Using the imaging atmospheric Cherenkov technique, the discovery of a gamma-ray source can be claimed if more than about 10 photons – with little background – come from the same point in the sky. These advances also come about because high-resolution pixel detectors are available. In the 1960s the best charge-sensitive amplifiers used for the readout of semiconductor counters had a noise level equivalent to that of 1000 electrons. Nowadays one can unambiguously count single electrons, because the noise level has decreased by a factor of 1000.

Imaging also plays a major role in medical diagnosis. The old X-ray technique has been improved and has been made more sensitive by using image intensifiers. Antiparticles, discovered in cosmic rays in the thirties of the last century are now routinely used in positron emission tomography (PET) for cancer diagnosis and therapy monitoring. Also γ rays are used in gamma cameras, scintigraphy, single photon emission computed tomography (SPECT), and PET, with all sorts of diagnostic indications, including suspicion of heart or brain disease. The operation of Compton telescopes, known from astroparticle physics experiments, has now found its way into medical diagnosis as Compton cameras, although image reconstruction from Compton cameras remains a major challenge.

Beyond diagnostic imaging, nuclear techniques have also entered the domain of medical therapy. The strong ionization of charged particles at the end of their range (Bragg peak) has initiated new methods in cancer treatment (particle therapy) with substantial advantages over

therapy with cobalt-60 γ rays. Apart from γ rays and charged particles, also neutrons have important applications in therapy.

On the detector side, a lot of progress has also been achieved. Early detectors, like the Wilson cloud chamber, provided lots of details about charged particles and their interactions. One drawback was a poor time resolution or repetition time. If you can only record one event per minute, such a detector is not suited for accelerator experiments. In the LHC, protons collide every 25 nanoseconds, resulting in a possible event rate of 40 million per second. Modern detectors cannot only measure the spatial coordinates, the energy and momentum of a particle, but they can also determine the identity of the particle. Particle identification is essential for the unambiguous characterization of interactions or new particle production. These techniques are also important in many other fields measuring electromagnetic radiation as γ rays, X rays, terahertz radiation, ultraviolett (UV), or infrared (IR) photons up to microwave photons.

This handbook centers on detection techniques in the field of particle physics, medical imaging, and related subjects. It is structured into four parts. The first two parts deal with basic ideas about particle detectors, like interactions of radiation and particles with matter and specific types of detectors. In the third part applications of these devices in high energy physics and related fields are presented. Finally, the last part concerns the ever-growing field of medical imaging using similar detection techniques. The different chapters of the book are written by well-known experts in their field. Clear instructions on the detection techniques and principles in terms of relevant operation parameters for scientists and graduate students are given. Detailed tables, diagrams, and figures will make this a very useful handbook for the application of these techniques in many different fields like physics, medicine, biology, other areas of natural science, and applications in metrology and technology. Also, it is our hope that such a broad presentation of particle detectors and radiation-based imaging can be a source of cross-fertilization between different fields of applications.

Irène Buvat and Claus Grupen
Orsay and Siegen
June 2011

About the Editors



Claus Grupen
Professor of Physics
Department of Physics
Faculty of Natural Science and Engineering
Siegen University
Emmy-Noether-Campus
Walter-Flex-Str. 3
57072 Siegen
Germany

Claus Grupen studied physics and mathematics at Kiel University, Germany. He received his Ph.D. in physics with a thesis on 'Electromagnetic Interactions of Cosmic Ray Muons' (1970). In 1971/1972 he became Visiting Fellow of the Royal Society of England, University of Durham, cooperating with Sir Arnold Wolfendale on cosmic ray physics. In 1974 he changed to Siegen University and obtained his habilitation on his cosmic ray work (1975). As member of the PLUTO experiment he worked on electron–positron interactions at the storage rings DORIS and PETRA at DESY in Hamburg (1974–1985). He became full professor at Siegen University in 1980. In 1995 he received the Special High Energy and Particle Physics Prize of the European Physical Society for the Discovery of the Gluon in 1979, jointly with members of the JADE, Mark J, PLUTO, and TASSO Collaborations. In 1981 and 1985 he spent sabbaticals as visiting professor at the University of Tokyo and as scientific associate at CERN (1990, 1994, and 2000). From 1984 on he is member of the ALEPH experiment (electron–positron interactions at the world largest e^+e^- collider LEP) at CERN.

Professor Grupen is author of a number of textbooks on physics (Particle Detectors 1996; second edition (with Boris Shwartz) 2008; Astroparticle Physics 2005; Introduction to Radiation Protection 2010).

Professor Grupen has organized a number of symposia, is referee for physics journals, and he is member of a number of institutions.

He served as Dean of the physics department at Siegen University 1980/81; 1991–1993; and 2002–2006, and he is still active in research after his retirement in 2006 working on the KASCADE-Grande and LOPES experiments at the Forschungszentrum Karlsruhe on astroparticle physics.



Irène Buvat
IMNC-UMR 8165 CNRS
Paris 7 and Paris 11 Universities
Orsay Campus
Building 440
91406 Orsay Cedex
France

Irène Buvat has received her Ph.D. degree in particle and nuclear physics from Paris Sud University, France, in 1992. During her thesis, she oriented her career towards applications of Nuclear Physics for Medical Imaging. She spent one year as a post-doc at University College London, UK, working on Single Photon Emission Computed Tomography (SPECT) and two years at the National Institutes of Health, Bethesda, USA, specializing in Positron Emission Tomography (PET). In 1995, she entered the French CNRS (Centre National de la Recherche Scientifique) and is currently the head of a "Quantification in Molecular Imaging" team of the "Imaging and Modelling in Neurobiology and Cancerology" CNRS lab in Orsay, France. Her research activities focus on developing correction and tomographic reconstruction methods in PET and SPECT to improve the accuracy and reduce the variability of measurements made from PET and SPECT images. Her methodological approach is based on the use of Monte Carlo simulations to investigate all details of the forward imaging process so as to identify key aspects to be considered when developing quantification methods. She is currently the spokesperson of the worldwide OpenGATE collaboration developing the GATE Monte Carlo simulation tool dedicated to Monte Carlo simulations in emission and transmission tomography and radiotherapy. Irène Buvat is also largely involved into making quantification in SPECT and PET a clinical reality. She contributed to a number of studies demonstrating the clinical values of sophisticated quantification to improve image interpretation, and obtained a large number of research contracts with the major companies in the field. She has authored or co-authored more than 80 peer-reviewed articles.

Irène Buvat teaches medical physics in several French Universities (Nantes, Lyon, Paris Sud). She acts as an associate editor of "IEEE Transactions on Medical Imaging" and of "IEEE Transactions on Nuclear Science" and serves on the Editorial Board of the "Journal of Nuclear Medicine" and on the International Advisory Board of "Physics in Medicine and Biology."

Table of Contents

Preface vii
About the Editors ix
List of Contributors xvii

Volume 1

Part 1 Basic Principles of Detectors and Accelerators 1

1 Interactions of Particles and Radiation with Matter 3
Simon I. Eidelman · Boris A. Shwartz

2 Electronics Part I 25
Helmuth Spieler

3 Electronics Part II 53
Helmuth Spieler

4 Data Analysis 83
Günther Dissertori

5 Statistics 103
Glen Cowan

6 Particle Identification 125
Jürgen Engelfried

7 Accelerators for Particle Physics 139
Helmut Burkhardt

8 Synchrotron Radiation and FEL Instrumentation 159
Shaukat Khan · Klaus Wille

9 Calibration of Radioactive Sources 187
Dirk Arnold · Herbert Janßen

10 Radiation Protection 201
Claus Grupen

Part 2 Specific Types of Detectors **237**

11 Gaseous Detectors **239**
Maxim Titov

12 Tracking Detectors **265**
Manfred Krammer · Winfried Mitaroff

13 Photon Detectors **297**
Peter Križan

14 Neutrino Detectors **313**
Franz von Feilitzsch · Jean-Côme Lanfranchi · Michael Wurm

15 Scintillation Counters **349**
Zane W. Bell

16 Semiconductor Counters **377**
Douglas S. McGregor

17 Gamma-Ray Detectors **411**
William L. Dunn · Douglas S. McGregor

18 Cherenkov Counters **453**
Blair Ratcliff · Jochen Schwiening

19 Muon Spectrometers **473**
Thomas Hebbeker · Kerstin Hoepfner

20 Calorimeters **497**
Richard Wigmans

21 New Solid State Detectors **519**
Christoph J. Ilgner

22 Radiation Damage Effects **535**
R.-Y. Zhu

Volume 2

Part 3 Applications of Detectors in Particle and Astroparticle Physics, Security, Environment and Art 557

23 Astrophysics and Space Instrumentation 559
John W. Mitchell · Thomas Hams

24 Indirect Detection of Cosmic Rays 593
Ralph Engel

25 Technology for Border Security 633
Dudley Creagh

26 Accelerator Mass Spectrometry and its Applications in Archaeology, Geology and Environmental Research 653
Wolfgang Kretschmer

27 Geoscientific Applications of Particle Detection and Imaging Techniques with Special Focus on the Monitoring Clay Mineral Reactions 667
Laurence N. Warr · Georg H. Grathoff

28 Particle Detectors Used in Isotope Ratio Mass Spectrometry, with Applications in Geology, Environmental Science and Nuclear Forensics 685
Nicholas S. Lloyd · Johannes Schwieters · Matthew S. A. Horstwood · Randall R. Parrish

29 Particle Detectors in Materials Science 703
Xin Jiang · Thorsten Staedler

30 Spallation – Neutrons Beyond Nuclear Fission 719
Harald Conrad

31 Neutron Detection 759
Alfred Klett

32 Instrumentation for Nuclear Fusion 791
Rudolf Neu

33 The Use of Neutron Technology in Archaeological and Cultural Heritage Research 813
Dudley Creagh

34 Radiation Detectors and Art 833
Andrea Denker

Part 4 Applications of Particle Detectors in Medicine 855

35 Radiation-Based Medical Imaging Techniques: An Overview 857
John O. Prior · Paul Lecoq

36 CT Imaging: Basics and New Trends 883
Françoise Peyrin · Klaus Engelke

37 SPECT Imaging: Basics and New Trends 917
Brian F. Hutton

38 PET Imaging: Basics and New Trends 935
Magnus Dahlbom

39 Image Reconstruction 973
Claude Comtat

40 Motion Compensation in Emission Tomography 1007
Jörg van den Hoff · Jens Langner

41 Quantitative Image Analysis in Tomography 1043
Irène Buvat

42 Compartmental Modeling in Emission Tomography 1065
Adriaan A. Lammertsma

43 Evaluation and Image Quality in Radiation-Based Medical Imaging 1083
Matthew A. Kupinski

44 Simulation of Medical Imaging Systems: Emission and Transmission Tomography 1095
Robert L. Harrison

45 High-Resolution and Animal Imaging Instrumentation and Techniques 1125
Nicola Belcari · Alberto Del Guerra

46 Imaging Instrumentation and Techniques for Precision Radiotherapy 1153
Katia Parodi · Christian Thieke

47 Tumor Therapy with Ion Beams 1179
Gerhard Kraft · Uli Weber

Index 1207

List of Contributors

Dirk Arnold
Physikalisch-Technische Bundesanstalt
(PTB)
Braunschweig
Germany

Nicola Belcari
University of Pisa
Pisa
Italy

Zane W. Bell
Oak Ridge National Laboratory
Oak Ridge, TN
USA

Helmut Burkhardt
CERN
Geneva
Switzerland

Irène Buvat
UMR 8165 CNRS
Paris 7 and Paris 11 Universities
Orsay CEDEX
France

Claude Comtat
CEA
Orsay
France

Harald Conrad
Forschungszentrum Jülich GmbH
Jülich
Germany

Glen Cowan
Royal Holloway
University of London
Egham, Surrey
UK

Dudley Creagh
University of Canberra
Canberra
Australia

Magnus Dahlbom
David Geffen School of Medicine at UCLA
University of California
Los Angeles, CA
USA

Alberto Del Guerra
University of Pisa
Pisa
Italy

Andrea Denker
Ion Beam Laboratory ISL
Helmholtz-Zentrum Berlin
Berlin
Germany

Günther Dissertori
Institute for Particle Physics
Zurich
Switzerland

William L. Dunn
Kansas State University
Manhattan, KS
USA

Simon I. Eidelman
Budker Institute of Nuclear Physics
Novosibirsk
Russia

Ralph Engel
Karlsruher Institut für Technologie
Karlsruhe
Germany

Jürgen Engelfried
Universidad Autónoma de San Luis Potosí
Manuel Nava #6
San Luis Potosí
Mexico

Klaus Engelke
University of Erlangen
Erlangen
Germany

Georg H. Grathoff
Institut für Geographie und Geologie
Ernst-Moritz-Arndt-Universität
Greifswald
Germany

Claus Grupen
Siegen University
Siegen
Germany

Thomas Hams
NASA/GSFC
Greenbelt, MD
USA
and
CRESST/University of MD Baltimore County
Baltimore, MD
USA

Robert L. Harrison
University of Washington
Seattle, WA
USA

Thomas Hebbeker
RWTH Aachen University
Aachen
Germany

Kerstin Hoepfner
RWTH Aachen University
Aachen
Germany

Matthew S. A. Horstwood
NERC Isotope Geosciences Laboratory
British Geological Survey
Keyworth, Nottingham
UK

Brian F. Hutton
University College London & UCLH
NHS Trust
London
UK

Christoph J. Ilgner
Helmholtz-Zentrum Dresden-Rossendorf
Dresden
Germany

Herbert Janßen
Physikalisch-Technische Bundesanstalt
(PTB)
Braunschweig
Germany

Xin Jiang
University of Siegen
Siegen
Germany

Shaukat Khan
Technische Universität Dortmund
Dortmund
Germany

Alfred Klett
Berthold Technologies GmbH & Co KG
Bad Wildbad
Germany

Gerhard Kraft
GSI Helmholtzzentrum für
Schwerionenforschung GmbH
Darmstadt
Germany

Manfred Krammer
Austrian Academy of Sciences
Vienna
Austria

Wolfgang Kretschmer
Physikalisches Institut
Universität Erlangen
Erlangen
Germany

Peter Križan
University of Ljubljana
Ljubljana
Slovenia

Matthew A. Kupinski
University of Arizona
Tucson, AZ
USA

Adriaan A. Lammertsma
VU University Medical Center
Amsterdam
The Netherlands

Jean-Côme Lanfranchi
Technische Universität München
Garching
Germany

Jens Langner
Helmholtz-Zentrum Dresden-Rossendorf
Dresden
Germany

Paul Lecoq
CERN
Geneva
Switzerland

Nicholas S. Lloyd
Thermo Fisher Scientific
Bremen
Germany

Douglas S. McGregor
Kansas State University
Manhattan, KS
USA

Winfried Mitaroff
Austrian Academy of Sciences
Vienna
Austria

John W. Mitchell
NASA/GSFC
Greenbelt, MD
USA

Rudolf Neu
Max-Planck-Institut für Plasmaphysik
Garching
Germany

Katia Parodi
Heidelberg Ion Beam Therapy Center
Heidelberg
Germany
and
University Clinic Heidelberg
Heidelberg
Germany

Randall R. Parrish
NERC Isotope Geosciences Laboratory
British Geological Survey
Keyworth, Nottingham
UK
and
University of Leicester
Leicester
UK

Françoise Peyrin
Inserm U1044; UMR CNRS 5220; INSA Lyon
Université de Lyon
Villeurbanne
France

John O. Prior
Centre Hospitalier Universitaire Vaudois
and University of Lausanne
Lausanne
Switzerland

Blair Ratcliff
Stanford Linear Accelerator Center
Stanford University
Menlo Park, CA
USA

Jochen Schwiening
Hadronenphysik 1
GSI Helmholtzzentrum für
Schwerionenforschung GmbH
Darmstadt
Germany

Johannes B. Schwieters
Thermo Fisher Scientific
Bremen
Germany

Boris A. Shwartz
Budker Institute of Nuclear Physics
Novosibirsk
Russia

Helmuth Spieler
Lawrence Berkeley National Laboratory
Berkeley, CA
USA

Thorsten Staedler
University of Siegen
Siegen
Germany

Christian Thieke
German Cancer Research Center
Heidelberg
Germany
and
University Clinic Heidelberg
Heidelberg
Germany

Maxim Titov
CEA Saclay Centre d'Études de Saclay
Gif-sur-Yvette Cedex
France

Jörg van den Hoff
Helmholtz-Zentrum Dresden-Rossendorf
Dresden
Germany

Franz von Feilitzsch
Technische Universität München
Garching
Germany

Laurence N. Warr
Institut für Geographie und Geologie
Ernst-Moritz-Arndt-Universität
Greifswald
Germany

Uli Weber
Universitätsklinikum Gießen/Marburg
Germany

Richard Wigmans
Texas Tech University
Lubbock, TX
USA

Klaus Wille
Technische Universität Dortmund
Dortmund
Germany

Michael Wurm
Technische Universität München
Garching
Germany

R.-Y. Zhu
Physics, Mathematics and
Astronomy Division
California Institute of Technology
Pasadena, CA
USA

Part 1

Basic Principles of Detectors and Accelerators

1 Interactions of Particles and Radiation with Matter

Simon I. Eidelman · Boris A. Shwartz
Budker Institute of Nuclear Physics, Novosibirsk, Russia

1 Introduction ***4***

2 Penetration of Charged Particles Through Matter ***4***
2.1 Energy and Angular Spectra of Delta Electrons 4
2.2 Energy Loss by Ionization and Excitation 5
2.3 Fluctuations of Ionization Losses 7
2.4 Multiple Scattering of Charged Particles 8
2.5 Channeling 9
2.6 Radiation Losses – Radiation Length and Critical Energy 11
2.7 Charged-Particle Range Due to Ionization Losses 12

3 Penetration of High-Energy Photons in Matter ***13***
3.1 Photoelectric Effect 13
3.2 Compton Effect 14
3.3 Production of Electron–Positron Pairs 15
3.4 Photon Flux Attenuation by Material 15

4 Electron–Photon Cascades ***16***

5 Nuclear Interactions of Hadrons with Matter ***19***

6 Neutrino Interactions with Matter ***21***

7 Conclusion ***23***

References ***23***

Further Reading ***23***

C. Grupen, I. Buvat (eds.), *Handbook of Particle Detection and Imaging*,
DOI 10.1007/978-3-642-13271-1_1, © Springer-Verlag Berlin Heidelberg 2012

Abstract: Main types of interactions of charged particles and photons are briefly described. For charged particles, we present basic formulas for energy losses by ionization and excitation, fluctuations of ionization losses, δ electrons, channeling, multiple scattering, and bremsstrahlung. We consider photoelectric effect, Compton effect, and electron–positron pair production for photons. Also discussed are nuclear interactions of hadrons with matter as well as neutrino interactions.

1 Introduction

A knowledge of phenomena, which occur when particles and radiation interact with matter, is necessary for development and usage of particle detectors, radiation protection, material studies with the help of ionizing radiation, etc. In this chapter main types of interactions of charged particles, photons, and neutrinos are briefly described.

There is an extensive literature devoted to interactions of particles and radiation with matter. A brief review and further references are given in Amsler et al. (2008), while the detailed consideration of the main relevant issues can be found in the classical book of Rossi (1952).

2 Penetration of Charged Particles Through Matter

The common phenomenon for all charged particles, which causes a change of their energy and direction in matter, is the electromagnetic interactions with electrons and nuclei. Electromagnetic interactions are responsible for particle scattering, ionization and excitation of atoms, bremsstrahlung, Cherenkov and transition radiations. It should be noted that Cherenkov and transition radiations result in a negligible energy loss and do not change a direction of particle motion. The main contribution to the energy loss comes from ionization and bremsstrahlung while the change of the particle trajectory is mostly caused by collisions with nuclei.

2.1 Energy and Angular Spectra of Delta Electrons

When an incident charged particle collides with an electron at rest, a maximum energy transfer is

$$\varepsilon_{\max} = 2m_{\mathrm{e}} \frac{P^2}{M^2 + m_{\mathrm{e}}^2 + 2Em_{\mathrm{e}}/c^2}\,, \tag{1}$$

where M, P, and E are the mass, momentum, and total energy of the incident particle while m_{e} is the electron mass. The following approximations are useful in particular cases:

$$\begin{aligned} \gamma \sim 1 \qquad & \varepsilon_{\max}\,[\mathrm{MeV}] \approx 2(\gamma - 1) \\ 1 \ll \gamma \ll M/(2m_{\mathrm{e}}) \qquad & \varepsilon_{\max}\,[\mathrm{MeV}] \approx \gamma^2 \\ \gamma \gg M/(2m_{\mathrm{e}}) \qquad & \varepsilon_{\max}\,[\mathrm{MeV}] \approx E\,, \end{aligned} \tag{2}$$

where $\gamma = E/Mc^2$.

Recoil electrons are usually referred to as δ electrons. The recoil angle, ϑ_δ, is related to the δ-electron kinetic energy (i.e., equivalent to the energy loss of the incident particle):

$$\cos\vartheta_\delta = \frac{E + m_e c^2}{P}\sqrt{\frac{\varepsilon}{\varepsilon + 2m_e c^2}}\,. \tag{3}$$

The collision of the heavy incident particle with a free electron can be approximately described by the well-known Rutherford formula

$$\frac{\mathrm{d}\sigma}{\mathrm{d}\Omega} = \frac{z^2 r_e^2}{4}\left(\frac{m_e c}{\beta_c p_c}\right)^2 \frac{1}{\sin^4(\theta_c/2)}\,, \tag{4}$$

where β_c, p_c, and θ_c are the electron velocity, momentum, and scattering angle in the center-of-mass system, z is the charge of the incident particle in units of the electron charge, and r_e – classical electron radius. The quantities p_c and θ_c can be easily related to the δ-electron kinetic energy ε:

$$\varepsilon = \frac{p_c^2(1-\cos\theta_c)}{m_e};\quad \mathrm{d}\varepsilon = -\frac{p_c^2}{m_e}\,\mathrm{d}\cos\theta_c;\quad p_c^2 = \frac{m_e\varepsilon_{\max}}{2}\,. \tag{5}$$

Taking into account these relations, we obtain the differential cross section $\mathrm{d}\sigma/\mathrm{d}\varepsilon$:

$$\mathrm{d}\sigma = \frac{2\pi z^2 r_e^2 m_e c^2}{\beta^2\varepsilon^2}\,\mathrm{d}\varepsilon\,. \tag{6}$$

When the incident particle is much heavier than the electron, we can take the electron velocity β_c equal to the velocity of the incident particle β in the laboratory frame. Then the energy distribution of δ electrons is

$$\frac{\mathrm{d}^2 n}{\mathrm{d}\varepsilon\,\mathrm{d}x} = N_A\frac{Z}{A}\frac{\mathrm{d}\sigma}{\mathrm{d}\varepsilon} = 0.15354\frac{Z}{A}\frac{1}{\beta^2\varepsilon^2}, \tag{7}$$

where Z and A are the atomic number and atomic mass, respectively. The thickness of material is measured in units of g/cm^2.

To take into account the electron spin, we have to use the formula for the Mott cross section (Mott 1929) instead of ❯ Eq. 4:

$$\frac{\mathrm{d}\sigma}{\mathrm{d}\Omega} = \frac{z^2 r_e^2}{4}\left(\frac{m_e c}{\beta_c p_c}\right)^2 \frac{1}{\sin^4(\theta_c/2)}\left(1-\beta_c^2\sin^2(\theta_c/2)\right). \tag{8}$$

This modifies the energy distribution to

$$\frac{\mathrm{d}^2 n}{\mathrm{d}\varepsilon\,\mathrm{d}x} = 0.15354\frac{Z}{A}\frac{1}{\beta^2\varepsilon^2}\left(1-\beta^2\frac{\varepsilon}{\varepsilon_{\max}}\right). \tag{9}$$

The formulas for the δ-electron energy distributions in the case of the incident electron, positron, and heavy spin-1/2 particle are given in Rossi (1952).

2.2 Energy Loss by Ionization and Excitation

The total energy transferred by the initial particle to δ electrons with the kinetic energy ε, exceeding a certain $\varepsilon_{\min}$, can be found by integrating ❯ Eq. 9. However, $\varepsilon_{\min}$ cannot approach 0 since electrons are assumed to be free in this expression. The accurate calculations in the Born

approximation result in the Bethe–Bloch equation (Bethe 1930, 1932; Bloch 1933) for the specific energy loss by a heavy spinless particle:

$$-\left(\frac{\mathrm{d}E}{\mathrm{d}x}\right)_{\mathrm{ion}} = Kz^2\frac{Z}{A}\frac{1}{\beta^2}\left(\frac{1}{2}\ln\frac{2m_\mathrm{e}c^2\beta^2\gamma^2\varepsilon_{\max}}{I^2} - \beta^2 - \frac{\delta}{2}\right), \tag{10}$$

where $K = 4\pi N_\mathrm{A} r_\mathrm{e}^2 m_\mathrm{e} c^2 = 0.307075\,\mathrm{MeV}/(\mathrm{g/cm}^2)$, I is a mean excitation energy (average ionization potential), and δ – the density-effect correction. The ionization energy-loss rate in various materials is shown in ❯ *Fig. 1* (Amsler et al. 2008). Common features of the shown dependences are fast growth, as $1/\beta^2$, at low energy, a wide minimum in the range $3 \le \beta\gamma \le 4$, and slow increase at high energy. A particle having $\mathrm{d}E/\mathrm{d}x$ near the minimum is often referred to as a minimum-ionizing particle or mip. The mip's ionization losses for all materials except hydrogen are in the range between 1 and 2 $\mathrm{MeV}/(\mathrm{g/cm}^2)$ slightly increasing from low to large Z.

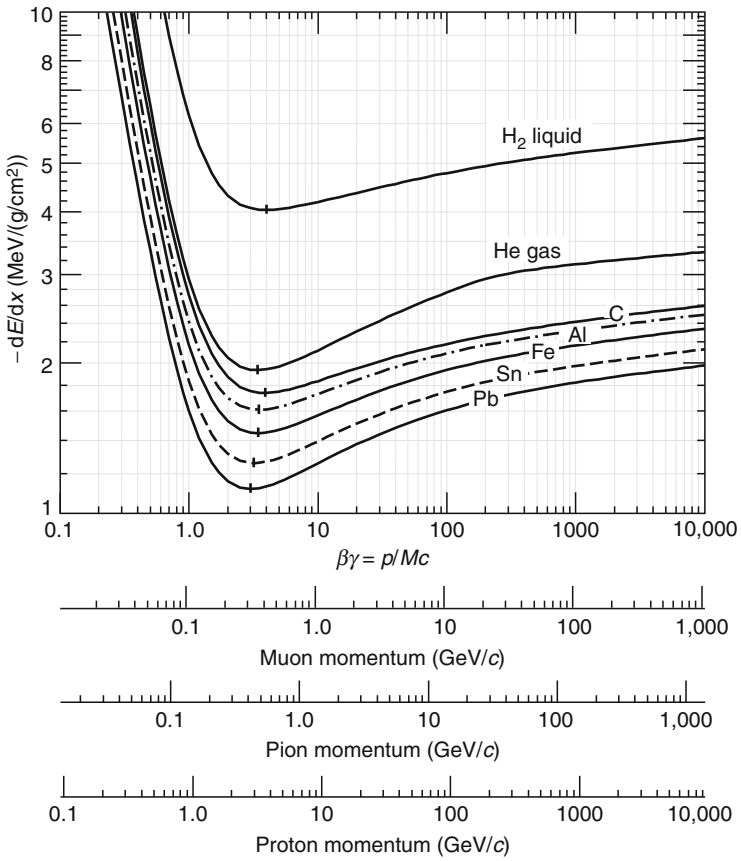


Fig. 1

The ionization energy-loss rate in various materials (Amsler et al. 2008)

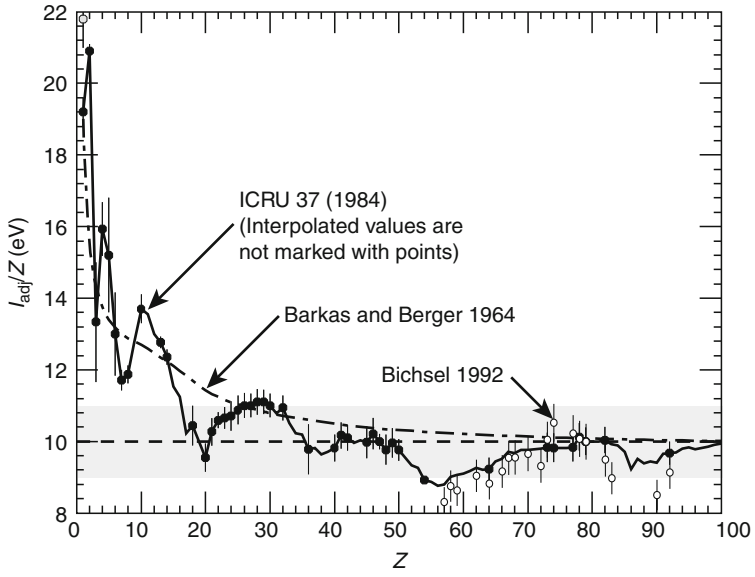


Fig. 2
Mean excitation energies

The I values are obtained from experimental data. A compilation given in Amsler et al. (2008) is presented in Fig. 2. The useful approximation for I is

$$I(\mathrm{eV}) = \begin{cases} 18.7 & \text{for } H_2 \\ Z(12 + 7/Z) & \text{for } Z \le 13 \\ Z(9.76 + 58.8 \cdot Z^{-1.19}) & \text{for } Z > 13 \end{cases} . \tag{11}$$

The density effect limits the increase of ionization losses in liquids and solids. This is a result of the medium polarization that effectively truncates distant collisions. The density effect correction can be calculated according to Sternheimer (1952) and Sternheimer and Peierls (1971).

2.3 Fluctuations of Ionization Losses

When a charged particle passes the layer of material, the energy distribution of the δ electrons (see Eq. 9) and fluctuations of their total number (n_δ) cause fluctuations of the energy losses, ΔE.

One of the typical cases is passage through the relatively thin layer of material by a relativistic particle, when the average energy loss $\langle \Delta E \rangle \ll \varepsilon_{\max}$ while $n_\delta \gg 1$. The probability density function (p.d.f.) for ΔE is strongly asymmetric with the maximum at ΔE_p and a long tail at high losses. The most probable energy loss is (Bichsel 1988)

$$\Delta E = \xi \left(\ln \frac{2 m_e c^2 \beta^2 \gamma^2 \xi}{I^2} + j - \beta^2 - \delta \right) , \tag{12}$$

where $\xi = (K/2)(Z/A)(x/\beta^2)$ and $j = 0.20$.

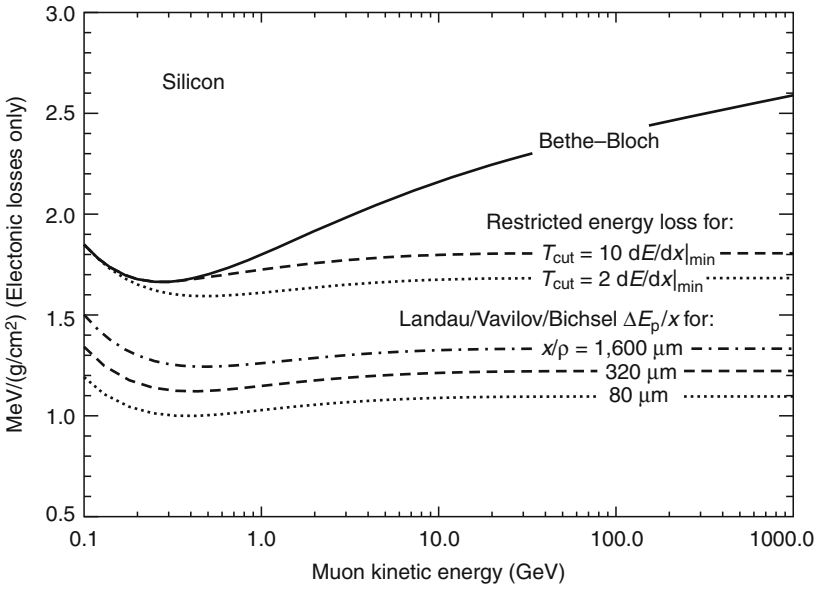


Fig. 3
The energy dependence of $\Delta E_p/x$

The energy dependence of $\Delta E_p/x$ is shown in Fig. 3 in comparison with dE/dx. As can be seen, the value of $\Delta E_p/x$ reaches the Fermi plateau at high energy while dE/dx is slowly growing due to increasing ε_{max}.

The p.d.f. for ultrarelativistic particles, when the parameter $G = \xi \cdot (2m_e c^2/\varepsilon_{max}) \lesssim 0.05$, was given by Landau (1944) and Vavilov (1957). This distribution is often referred to as the Landau distribution.

The position of the peak of this distribution is determined by Eq. 12 and the full width at half maximum is FWHM $\approx 4\xi$. The ionization-loss distributions at $0.05 < G < 10$ are discussed in detail in Rossi (1952).

It should be noted that ionization losses in very thin layers, when $n_\delta \lesssim 1 \div 10$, are not described by the Landau distribution. The most probable loss is still given by Eq. 12, but the distribution is much wider than the Landau function. This case is typical for the gaseous and thin silicon detectors (Onuchin and Telnov 1974; Bichsel 1988, 2006). *Figure 4* shows the energy-loss distributions for 500 MeV pions passing through a silicon layer (Amsler et al. 2008).

2.4 Multiple Scattering of Charged Particles

A particle passing through material undergoes multiple small-angle scattering, mostly due to large-impact-parameter interactions with nuclei. Then an initially parallel particle beam gets the angular spread after traveling through the layer of material. The angular distribution due to the multiple Coulomb scattering is described by the Molière theory (Bethe 1953). For small scattering angles it is normally distributed around the average value $\theta = 0$. Larger scattering

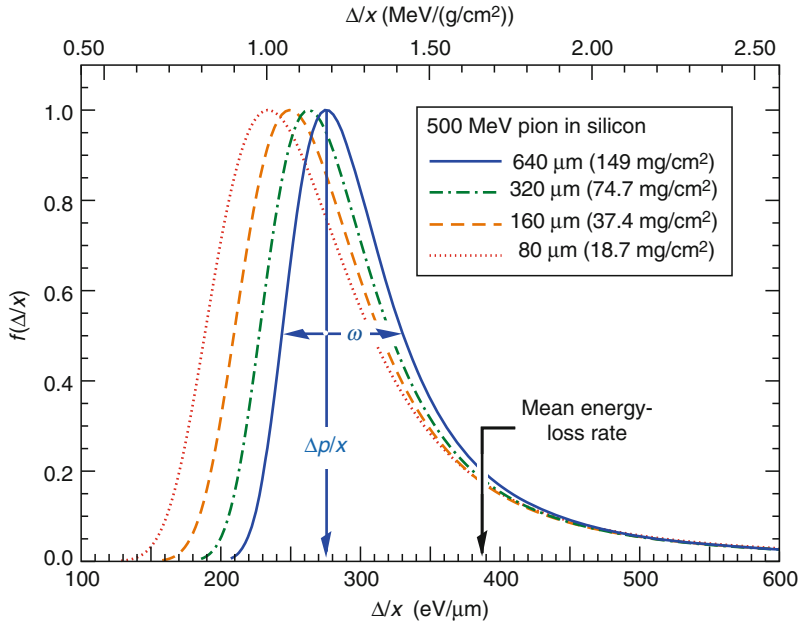


Fig. 4

Energy-loss distribution for 500 MeV pions passing through a silicon layer (Amsler et al. 2008)

angles caused by rare collisions of charged particles with nuclei are, however, more probable than expected from a Gaussian distribution.

The root mean square of the projected scattering-angle distribution is expressed as

$$\theta_{\mathrm{rms}}^{\mathrm{proj}} = \sqrt{\langle\theta^2\rangle} = \frac{13.6\,\mathrm{MeV}}{\beta c p} z \sqrt{\frac{x}{X_0}} \left[1 + 0.038 \ln(x/X_0)\right], \tag{13}$$

where p (in MeV/c) is the momentum, βc the velocity, and z – the charge of the scattered particle. The quantity x/X_0 is the thickness of the scattering medium measured in units of the radiation length (X_0). The meaning of the latter as well as formulas for its calculation are given in Sect. 2.6.

2.5 Channeling

The energy loss of charged particles as described by the Bethe–Bloch formula should be modified for crystals where the collision partners are arranged on a regular lattice. When one looks at a crystal, it becomes immediately clear that the energy loss along certain crystal directions will be quite different from that along a nonaligned direction or in an amorphous substance. Motion along such channeling directions is governed mainly by coherent scattering off strings and planes of atoms rather than by the individual scattering off single atoms. This leads to anomalous energy losses of charged particles in crystalline materials (Møller 1994).

It is obvious from the crystal structure that charged particles can only be channeled along a crystal direction if they are moving more or less parallel to crystal axes. The critical angle necessary for that is small (approximately 0.3° for $\beta \approx 0.1$) and decreases with energy. For the axial direction ($\langle 111 \rangle$, body diagonal) it can be estimated from

$$\psi\,[\text{degrees}] = 0.307 \cdot [z \cdot Z/(E \cdot d)]^{0.5}, \tag{14}$$

where z and Z are the charges of the incident particle and the crystal atom, E is the particle energy in MeV, and d is the interatomic spacing in Å. The quantity ψ is measured in degrees (Gemmel 1974).

For protons ($z = 1$) passing through a silicon crystal ($Z = 14$; $d = 2.35$ Å) the critical angle for channeling along the direction-of-body diagonals becomes

$$\psi = 13\,\mu\text{rad}/\sqrt{E\,[\text{TeV}]}. \tag{15}$$

For planar channeling along the face diagonals ($\langle 110 \rangle$ axis) in silicon one gets (Møller 1994)

$$\psi = 5\,\mu\text{rad}/\sqrt{E\,[\text{TeV}]}. \tag{16}$$

Of course, the channeling process also depends on the charge of the incident particle.

For a field inside a crystal of silicon atoms along the $\langle 110 \rangle$ crystal direction, one obtains $1.3 \cdot 10^{10}$ V/cm. This field extends over macroscopic distances and can be used for the deflection of high-energy charged particles using bent crystals (Gemmel 1974).

Channeled positive particles are kept away from a string of atoms and consequently suffer from a relatively small energy loss. *Figure 5* shows the energy-loss spectra for 15 GeV/c

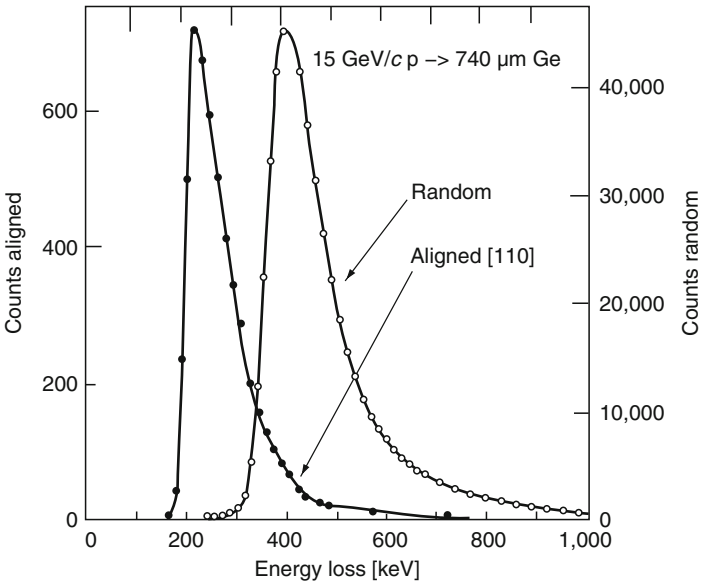


Fig. 5

The energy-loss spectra for 15 GeV/*c* protons passing through a 740 µm thick germanium crystal (Gemmel 1974)

protons passing through a 740 µm thick germanium crystal (Gemmel 1974). The energy loss of channeled protons is lower by about a factor of 2 compared to random directions in the crystal.

2.6 Radiation Losses – Radiation Length and Critical Energy

Fast charged particles may lose energy not only due to the ionization, but also by bremsstrahlung when these particles are decelerated in the Coulomb field of the nucleus. For electrons, the radiation losses become dominant from the energy of a few tens of MeV. Radiation losses are characterized by the radiation length, X_0. After passing a layer of material with a thickness X_0 the average electron energy decreases to 1/e from the initial energy. Then the radiation energy loss can be expressed as

$$-\left(\frac{\mathrm{d}E}{\mathrm{d}x}\right)_{\mathrm{rad}} = \frac{E}{X_0}. \tag{17}$$

The values of X_0 for various materials were calculated and tabulated by Tsai (1974):

$$\begin{aligned}\frac{1}{X_0} &= 4\alpha r_\mathrm{e}^2 \frac{N_\mathrm{A}}{A}\left\{Z^2\left[L_\mathrm{rad} - f(Z)\right] + ZL'_\mathrm{rad}\right\} \\ &= \frac{1}{A \cdot 716.408\,\mathrm{g/cm^2}}\left\{Z^2\left[L_\mathrm{rad} - f(Z)\right] + ZL'_\mathrm{rad}\right\}.\end{aligned} \tag{18}$$

The function $f(Z)$ can be approximated (Davies et al. 1954):

$$f(Z) = a^2\left[(1+a^2)^{-1} + 0.20206 - 0.0369a^2 + 0.0083a^4 - 0.002a^6\right], \tag{19}$$

where $a = \alpha Z$. For $Z > 4$, $L_\mathrm{rad} = \ln(184.15Z^{-1/3})$ and $L'_\mathrm{rad} = \ln(1194Z^{-2/3})$. For the lightest elements, the values of L_rad, L'_rad deviate from those given by the quoted formulas by ~ 10–15% and can be found in Tsai (1974).

The radiation length of a mixture of elements or a compound can be approximated by

$$X_0 = \frac{1}{\sum \rho_i / X_0^i}, \tag{20}$$

where ρ_i are the mass fractions of the components with the radiation lengths X_0^i.

Contributions of various processes of the energy loss for electrons and positrons are presented in ❯ *Fig. 6* (Amsler et al. 2008). The energy at which specific radiation losses, $(\mathrm{d}E/\mathrm{d}x)_\mathrm{rad}$, reach the ionization losses, $(\mathrm{d}E/\mathrm{d}x)_\mathrm{ion}$, is called the critical energy, E_c. The critical energy for electrons can be approximated by the formulas

$$E_\mathrm{c} = \frac{610\,\mathrm{MeV}}{Z + 1.24}\ (\text{for solids and liquids}), \quad E_\mathrm{c} = \frac{710\,\mathrm{MeV}}{Z + 0.92}\ (\text{for gases}). \tag{21}$$

In a wide energy range from about 100 MeV up to 1 TeV the differential probability of the emission of a photon with energy ε_ph can be approximated by (Rossi 1952):

$$\frac{\mathrm{d}n}{\mathrm{d}\varepsilon_\mathrm{ph}\,\mathrm{d}x} = \frac{1}{X_0 \cdot \varepsilon_\mathrm{ph}}\left(\frac{4}{3} - \frac{4}{3}y + y^2\right), \tag{22}$$

where $y = \varepsilon_\mathrm{ph}/E_\mathrm{e}$ is a portion of the initial electron energy, E_e, transferred to the photon. The path in a material as well as X_0 are measured in g/cm^2. This formula is valid in the so-called complete-screening approximation, which is determined by the condition

$$\frac{50\,\mathrm{MeV}}{E_\mathrm{e}} \frac{y}{1-y} Z^{-1/3} \ll 1. \tag{23}$$

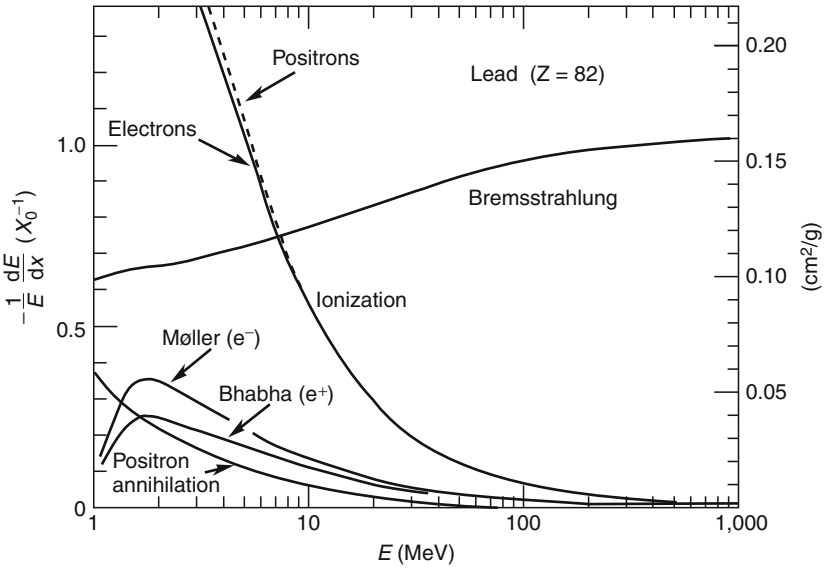


Fig. 6
Contributions of the different processes to the energy loss for electrons and positrons (Amsler et al. 2008)

Obviously, this is not fulfilled at $y \approx 1$, however, Eq. 22 is not accurate at $y \approx 0$, especially at very high energies ($E_e > 100$ GeV) due to bulk-medium effects (see Amsler et al. (2008) for discussion and further references).

The angular distribution for emitted photons is quite narrow at high electron energy. The root mean square of the emission angle is (Rossi 1952)

$$\sqrt{\langle\theta^2\rangle} \sim \frac{m_e c^2}{E_e} \ln \frac{E_e}{m_e c^2} . \tag{24}$$

At very high energies, the radiation processes become important for all charged particles. *Figure 7* shows the specific energy loss by muons in copper including excitation, ionization, and radiation processes (Amsler et al. 2008).

2.7 Charged-Particle Range Due to Ionization Losses

The range of charged particles can be calculated from their energy losses:

$$R = \int_0^{E_0} \frac{dE}{(dE/dx)} . \tag{25}$$

In practice, only a range for heavy particles is well defined when the main mechanism of the energy loss is medium ionization. In this case dE/dx should be taken from Eq. 10. The range of heavy particles for several substances as a function of particle momentum is presented in *Fig. 8* (Amsler et al. 2008). The range for electrons is not well defined because

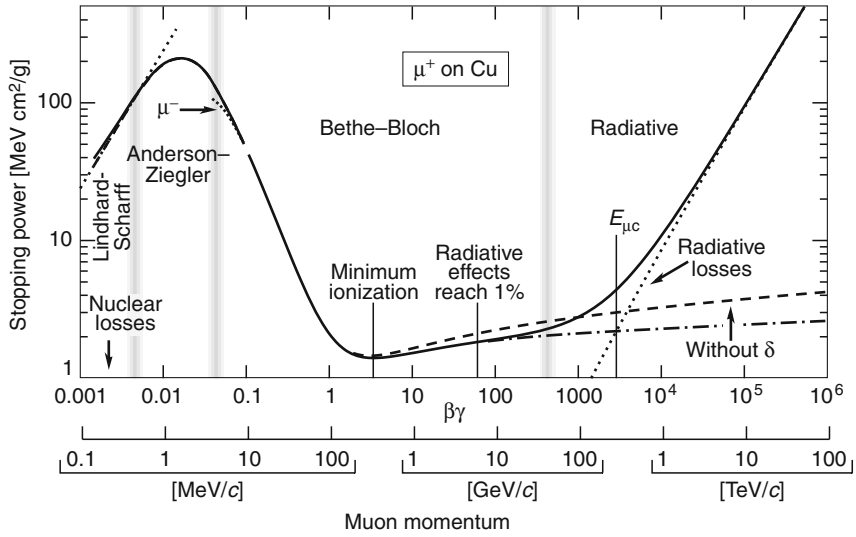


Fig. 7

The total energy losses (–d*E*/d*x*) for muons in copper (*solid line*) (Amsler et al. 2008). *Vertical bands* indicate different approximations

of large energy fluctuations in the bremsstrahlung losses and large path-length variations due to multiple scattering.

3 Penetration of High-Energy Photons in Matter

In contrast to the charged particles, photons in each interaction with electrons or nuclei disappear or change dramatically their energy and direction. Thus, the photon beam subsides according to an exponential law:

$$I = I_0 \exp(-\mu x)\,, \tag{26}$$

where I is a photon flux and μ – mass attenuation coefficient. The processes of photon interactions in material are discussed in the next subsections.

3.1 Photoelectric Effect

Photoelectric effect implies an absorption of a photon by the electron bound in atom and transfer of the photon energy to this electron. The photoeffect cross section for the photon of energy $E_\gamma > E_K$, where E_K is the K-shell energy, is particularly large for the K-shell electrons. The total cross section is

$$\sigma_{ph}^{K} = \sqrt{\frac{32}{\zeta^7}}\alpha^4 Z^5 \sigma_{Th}\ [\mathrm{cm}^2/\mathrm{atom}]\,, \tag{27}$$

where $\zeta = E_\gamma / m_e c^2$ and $\sigma_{Th} = \frac{8}{3}\pi r_e^2 = 665$ mb is the cross section of Thompson scattering.

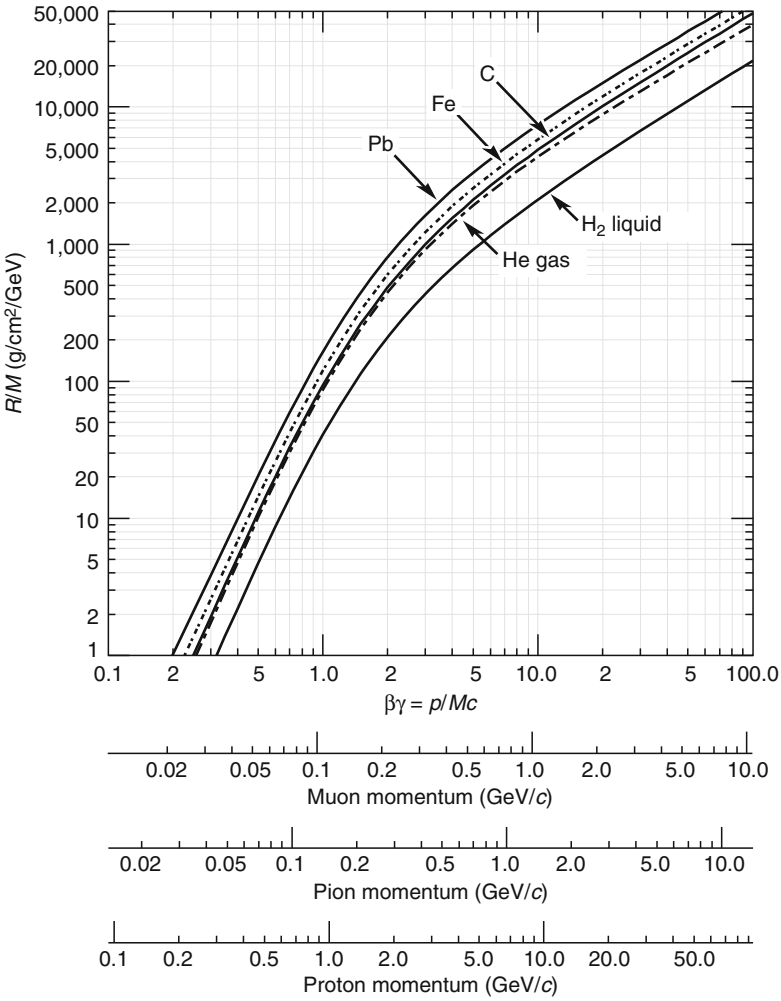


Fig. 8
Range of heavy charged particles in various substances (Amsler et al. 2008)

The photoelectric cross section has sharp discontinuities when E_γ is equal to the binding energy of the atomic shells. After a photoelectric effect in the K shell, the atomic electrons are rearranged and characteristic X rays (like K_α) or Auger electrons are emitted.

3.2 Compton Effect

The Compton effect is inelastic scattering of photons by quasi-free atomic electrons. After this scattering the photon energy, E'_γ, and the scattering angle, θ_γ, are related by the formula

$$\frac{E'_\gamma}{E_\gamma} = \frac{1}{1+\zeta(1-\cos\theta_\gamma)}\,, \tag{28}$$

where $\zeta = E_\gamma / m_e c^2$. The total cross section of Compton scattering derived by the integration of the Klein–Nishina formula (Klein and Nishina 1929) is

$$\sigma_C = \frac{\pi r_e^2}{\zeta}\left[\left(1 - \frac{2}{\zeta} - \frac{2}{\zeta^2}\right)\ln(1+2\zeta) + \frac{1}{2} + \frac{4}{\zeta} - \frac{1}{2(1+2\zeta)^2}\right]. \tag{29}$$

This formula provides the cross section per one electron. In the ultrarelativistic case, when $\zeta \gg 1$, the formula for the Compton cross section reduces to

$$\sigma_C = \frac{\pi r_e^2}{\zeta}\left(\ln 2\zeta + \frac{1}{2}\right). \tag{30}$$

3.3 Production of Electron–Positron Pairs

The production of an electron–positron pair by the photon becomes possible when the photon energy, E_γ, exceeds the threshold

$$E_\gamma \geq 2m_e c^2 + \frac{2m_e^2 c^2}{M_{\text{nucleus}}} \approx 2m_e c^2. \tag{31}$$

As for the bremsstrahlung, the screening parameter is defined,

$$\gamma = 100\frac{m_e c^2}{E_\gamma}\frac{1}{\nu(1-\nu)}Z^{-1/3}, \quad \nu = \frac{E_e + m_e c^2}{E_\gamma}, \tag{32}$$

where E_e is an electron (or positron) kinetic energy.

In the case of complete screening ($\gamma \ll 1$), the pair-production cross section is given by

$$\sigma_{\text{pair}} = 4\alpha r_e^2 Z^2\left(\frac{7}{9}\ln\frac{183}{Z^{1/3}} - \frac{1}{54}\right)\ [\text{cm}^2/\text{atom}]. \tag{33}$$

Then the probability dw of photon conversion at the small path length dx is approximately equal to

$$dw = \frac{7}{9}\frac{dx}{X_0}. \tag{34}$$

The energy distribution of the produced electrons and positrons is given by the differential probability

$$\frac{dw}{dE_+\,dx} = \frac{\alpha r_e^2}{E_\gamma}\frac{N_A}{A}Z^2 f(Z, E_\gamma, \nu) \tag{35}$$

where E_+ is the energy of the positron and $f(Z, E_\gamma, \nu)$ is a dimensionless function shown in ❯ *Fig. 9*. As can be seen from the figure, the distributions are close to being flat except areas near the edges of the spectra.

3.4 Photon Flux Attenuation by Material

An overview of the processes contributing to the attenuation of the photon flux is presented in ❯ *Fig. 10* (Amsler et al. 2008) for carbon and lead. The flux of the photons penetrating in matter

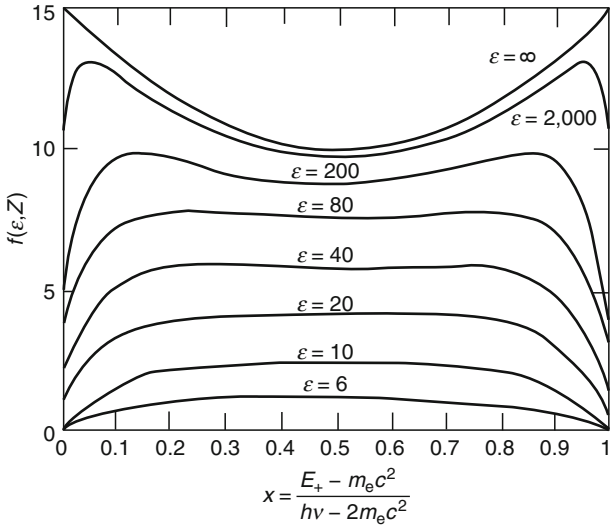


Fig. 9
Energy-partition function $f(Z, E_\gamma, x)$ with $\varepsilon = E_\gamma/m_e c^2$ as a parameter

is described by Eq. 26, where μ is called the total mass attenuation coefficient that is related to the sum of cross sections of all contributing processes:

$$\mu = \frac{N_A}{A} \sum_i \sigma_i, \tag{36}$$

where σ_i is the atomic cross section for the process i. The value $1/\mu$, called the photon attenuation length, is shown in *Fig. 11* for several elements as a function of the photon energy (Amsler et al. 2008). For a chemical compound, the effective value of $\mu_{\text{eff}} = 1/\lambda_{\text{eff}}$ can be found as $1/\lambda_{\text{eff}} = \sum w_Z/\lambda_Z$, where w_Z and λ_Z are the weight content and attenuation length, respectively, for the element with a nucleus charge Z.

4 Electron–Photon Cascades

At high energies (higher than 100 MeV), electrons lose their energy almost exclusively by bremsstrahlung while the main interaction process for photons is electron–positron pair production. Thus, these processes lead to the development of an electromagnetic cascade in matter where the number of particles, i.e., electrons, positrons, and photons, increases until the energy of the particles decreases to the critical energy, E_c (see Sect. 2.6, Eq. 21). Since both processes, bremsstrahlung and pair production, are characterized by the radiation length, X_0, this unit is a natural measure for the electron–photon shower development.

Shower development is characterized by the number of charged particles and photons as well as the energy-deposition rate at the depth $t = x/X_0$ in an absorber. The longitudinal

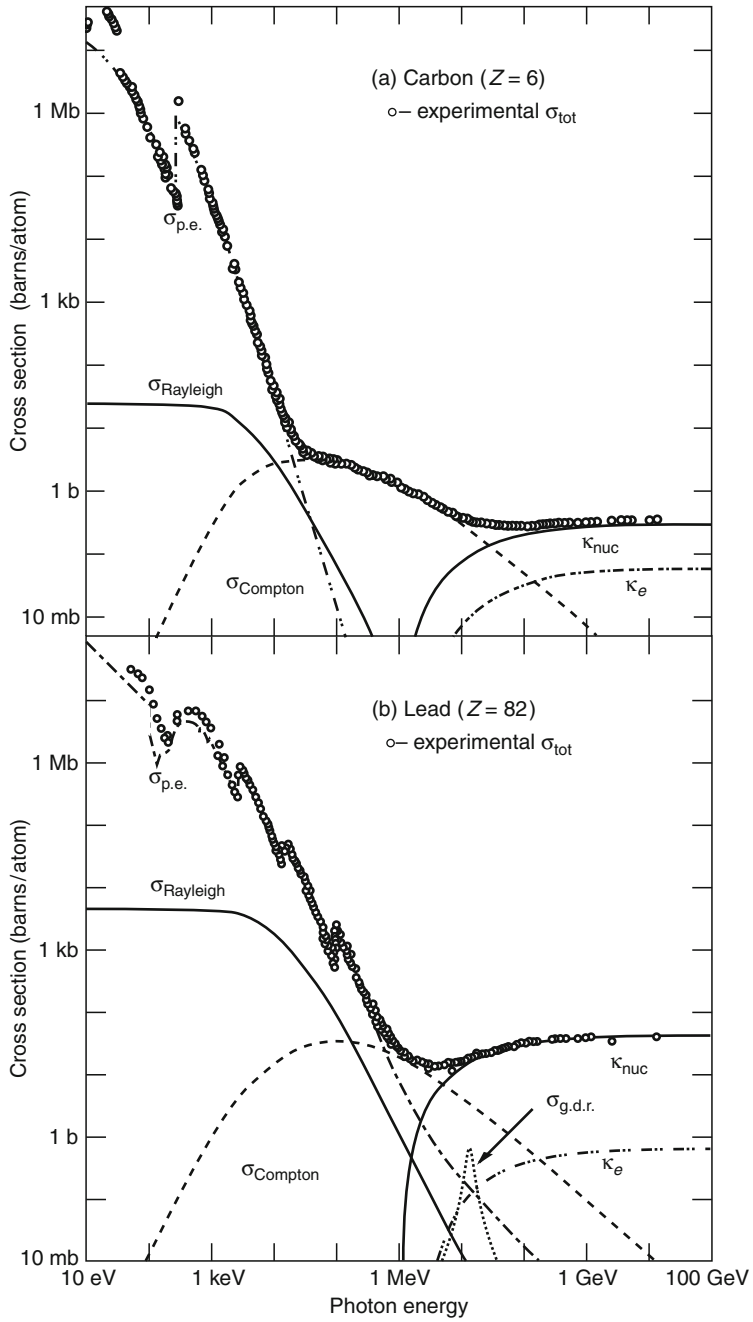


Fig. 10

Photon total and partial cross sections in carbon and lead (Amsler et al. 2008). Here $\sigma_{\text{p.e.}}$ is for photoelectric effect, σ_{Rayleigh} Rayleigh (coherent) scattering, σ_{Compton} Compton (incoherent) scattering by atomic electrons, κ_{nuc} pair production at nuclei; κ_{e} pair production at electrons, $\sigma_{\text{g.d.r.}}$ photonuclear interactions, most notably the Giant Dipole Resonance

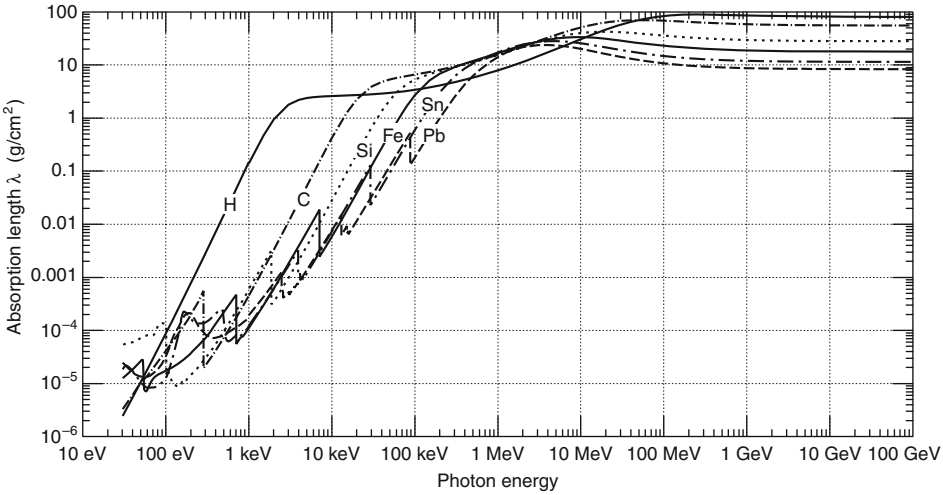


Fig. 11
The photon mass attenuation length (Amsler et al. 2008)

distribution of the energy deposition in electromagnetic cascades can be approximated by the following expression evaluated from the Monte Carlo simulation (Longo and Sestili 1975),

$$\frac{\mathrm{d}E}{\mathrm{d}t} = E_0 b \frac{(bt)^{a-1}\,\mathrm{e}^{-bt}}{\Gamma(a)}\,, \tag{37}$$

where $\Gamma(g)$ is the Euler's Γ function, defined by

$$\Gamma(g) = \int_0^\infty \mathrm{e}^{-x} x^{g-1}\,\mathrm{d}x\,. \tag{38}$$

The gamma function has the property

$$\Gamma(g+1) = g\,\Gamma(g). \tag{39}$$

Here a and b are model parameters and E_0 is the energy of the incident particle. In this approximation, the maximum of shower development is reached at

$$t_{\max} = \frac{a-1}{b} = \ln\left(\frac{E_0}{E_\mathrm{c}}\right) + C_{ye}\,, \tag{40}$$

where $C_{ye} = 0.5$ for a gamma-induced shower and $C_{ye} = -0.5$ for an incident electron. The parameter b as obtained from simulation results is $b \approx 0.5$ for heavy absorbers from iron to lead. Then the energy-dependent parameter a can be derived from Eq. 40.

The experimentally measured distributions (Akchurin et al. 2001; Baumgart et al. 1987) are well described by Monte Carlo simulation using the code EGS4 (Nelson et al. 1985; Amsler et al. 2008). Equation 37 provides a reasonable approximation for electrons and photons with energies larger than 1 GeV and a shower depth of more than $2\,X_0$, while for other conditions it gives a rough estimate only. The longitudinal development of the electromagnetic shower initiated by an electron in matter is shown in *Fig. 12*.

The angular distribution of the particles produced by bremsstrahlung and pair production is very narrow, the characteristic angles are of the order of $m_\mathrm{e}c^2/E_\gamma$. That is why the lateral

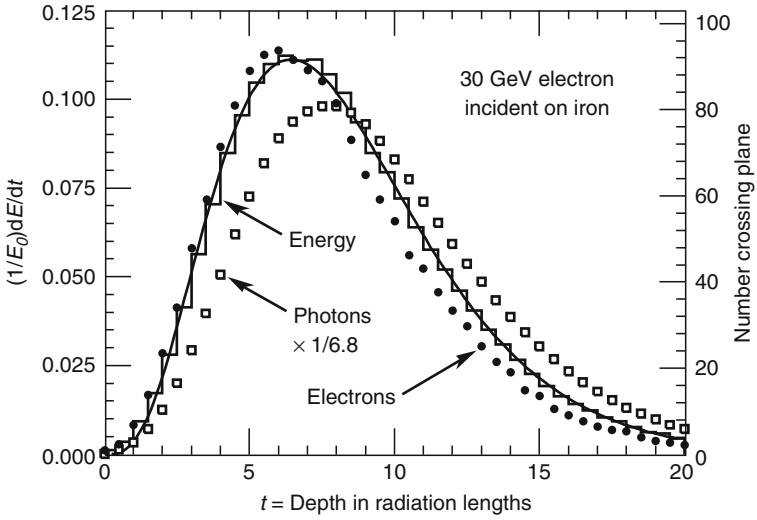


Fig. 12

Longitudinal shower development of a 30 GeV electron-induced cascade obtained by the EGS4 simulation in iron (Nelson et al. 1985; Amsler et al. 2008). The solid histogram shows the energy deposition; *black circles* and *open squares* represent the number of electrons and photons, respectively, with the energy larger than 1.5 MeV; the *solid line* is the approximation given by Eq. 37

width of an electromagnetic cascade is mainly determined by multiple scattering and can be described in units of the Molière radius

$$R_{\mathrm{M}} = \frac{21\,\mathrm{MeV}}{E_{\mathrm{c}}} X_0\,[\mathrm{g/cm^2}]. \tag{41}$$

The largest fraction of energy is deposited in a relatively narrow shower core. About 95% of the shower energy is contained in a cylinder around the shower axis whose radius is $R(95\%) = 2R_{\mathrm{M}}$ almost independently of the energy of the incident particle.

5 Nuclear Interactions of Hadrons with Matter

As an example of the hadronic cross section, the pion–deuteron and pion–proton total and inelastic cross sections are presented in *Fig. 13*. The total pion–nucleus cross section has similar momentum dependence scaled as $A^{1/3}$; however, the peaks in the range of the isobar production are less pronounced. Similar cross-section dependence is typical for other hadrons as well. As can be seen, the momentum dependence in the range higher than 1 GeV is rather flat. For momenta exceeding a few hundred MeV, the interaction is mostly inelastic, which implies a production of secondaries and knock-out nucleons from the nuclei. These processes induce a development of the hadron cascades when the energy of the incident hadron is high enough ($\gtrsim$10 GeV).

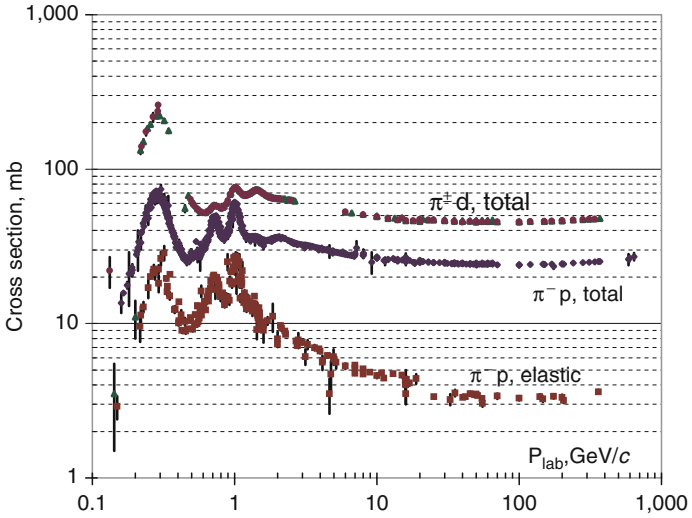


Fig. 13
Pion–proton and pion–deuteron cross sections (Amsler et al. 2008)

The longitudinal development of the hadron shower is determined by the average nuclear interaction length, λ_I, which can be roughly estimated as (Amsler et al. 2008)

$$\lambda_I \approx 35\,\mathrm{g/cm^2} \cdot A^{1/3}. \tag{42}$$

In most detector materials, this quantity is much larger than the radiation length X_0, which describes the behavior of electron–photon cascades. In the inelastic hadronic processes mainly charged and neutral pions, but with lower multiplicities also kaons, nucleons, and other hadrons are produced. The average particle multiplicity per interaction varies only weakly with energy ($\propto \ln E$). The average transverse momentum of secondary particles can be characterized by

$$\langle p_T \rangle \approx 0.35\,\mathrm{GeV}/c. \tag{43}$$

The average inelasticity, that is, the fraction of energy, which is transferred to secondary particles in the interaction, is around 50%.

A large component of the secondary particles in hadron cascades is neutral pions, which represent approximately one third of the pions produced in each inelastic collision. Therefore, a considerable fraction of the energy of the hadron shower, f_{em}, is deposited in the form of an electromagnetic shower which can be approximated by (Gabriel et al. 1994)

$$f_{em} = 1 - \left(\frac{E_0}{E_{sc}}\right)^{k-1}, \tag{44}$$

where E_0 is the energy of the incident hadron, E_{sc} is a parameter varying from 0.7 GeV (for iron) to 1.3 GeV (for lead), and k is between 0.8 and 0.85. Details can be found in Wigmans (2000).

In contrast to electrons and photons, whose electromagnetic energy is almost completely recorded in the detector, a substantial fraction of the energy in hadron cascades remains "invisible" (f_{inv}). This is related to the fact that some part of the hadron energy is used to break up nuclear bonds. This nuclear binding energy is provided by the primary and secondary hadrons and does not contribute to the energy deposition within a hadronic shower. In addition, long-lived or stable neutral particles like neutrons, K_L^0, or neutrinos can escape from the calorimeter,

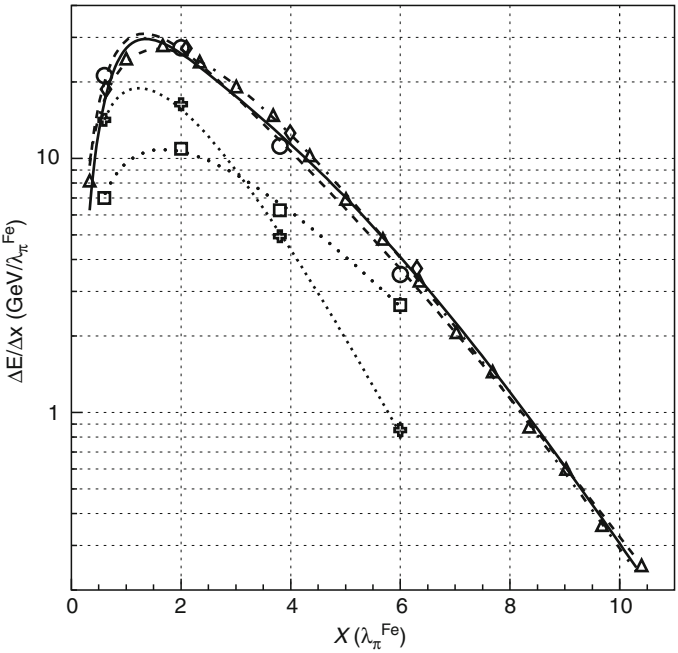


Fig. 14

The longitudinal energy distribution in a hadronic shower in iron induced by 100 GeV pions. The depth *X* is measured in units of the interaction length λ_I. *Open circles* and *triangles* are experimental data; *diamonds* are predictions of simulation. The *dash-dotted* line is a simple fit by Eq. 37 with optimal *a* and *b*, the other lines are more sophisticated approximations. *Crosses* and *squares* are contributions of electromagnetic showers and the non-electromagnetic part, respectively (Amaral et al. 2000)

thereby reducing the visible energy. The total invisible energy fraction of a hadronic cascade can be estimated as $f_{\text{inv}} \approx 30$–$40\%$ (Wigmans 2000). *Figure 14* shows the measured longitudinal shower development of 100 GeV pions in iron (Amaral et al. 2000) in comparison to Monte Carlo calculations and empirical approximations.

Apart from the longer longitudinal development of hadron cascades, their lateral width is also sizably increased compared to electron cascades. While the lateral structure of electron showers is mainly determined by multiple scattering, in hadron cascades it is caused by large transverse momentum transfers in nuclear interactions.

6 Neutrino Interactions with Matter

A neutrino interacts with matter due to the charged-current and neutral-current weak interactions. The typical charged-current interactions are

$$\nu_l + n \to X + l^- \,, \tag{45}$$

$$\overline{\nu}_l + p \to X + l^+ \,, \tag{46}$$

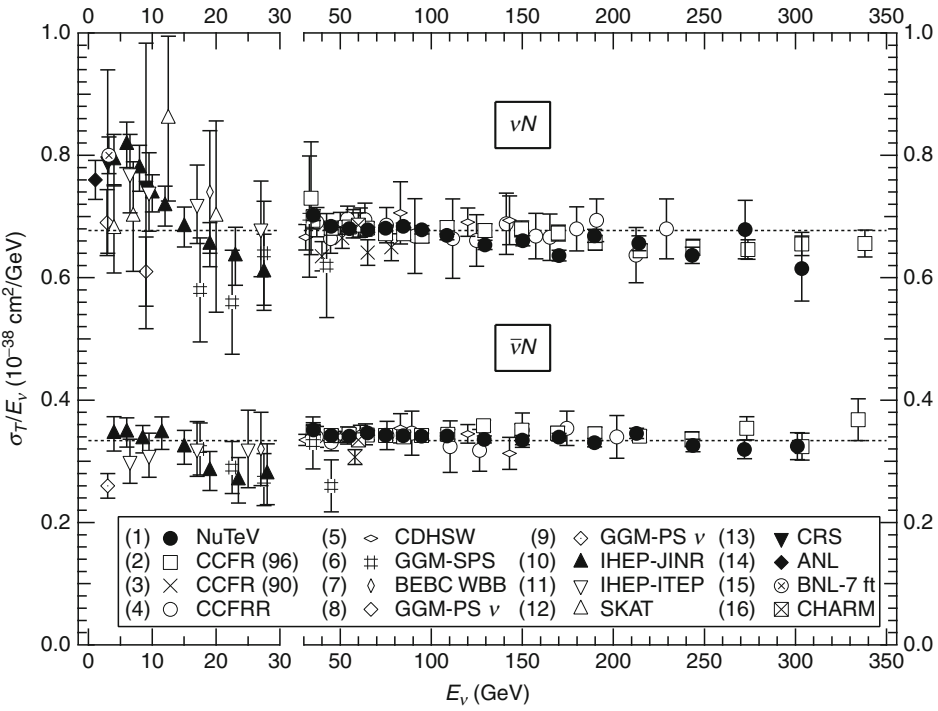


Fig. 15

Total cross sections for muon-neutrino interactions with nucleons at high energies (Amsler et al. 2008). The *straight lines* are the isoscalar-corrected values averaged over 30–200 GeV

where l stands for an electron, muon, or τ lepton. X means just p (reaction 45) or n (reaction 46) for the low-energy neutrino and a multiparticle final state for the high-energy neutrino. The total neutrino–nucleon cross sections measured in fixed-target experiments are shown in *Fig. 15*. These cross sections at high neutrino energies can be approximated as

$$\sigma_{\nu N}\,[\mathrm{cm}^2] = (0.677 \pm 0.014)\cdot 10^{-38} E\,[\mathrm{GeV}]\,,$$
$$\sigma_{\bar{\nu}N}\,[\mathrm{cm}^2] = (0.334 \pm 0.008)\cdot 10^{-38} E\,[\mathrm{GeV}]\,. \tag{47}$$

The charged-current interaction also contributes to the elastic νe scattering.

The neutral-current interactions induce the processes of the elastic and quasielastic scattering:

$$\begin{aligned} \nu_l + N &\to \nu_l + X\,,\\ \bar{\nu}_l + N &\to \bar{\nu}_l + X\,,\\ \nu_l + e &\to \nu_l + e\,,\\ \bar{\nu}_l + e &\to \bar{\nu}_l + e\,. \end{aligned} \tag{48}$$

Cross sections of these processes are much smaller than those induced by the charged currents.

In general, all neutrino cross sections are too small to provide a noticeable attenuation of the neutrino flux for any material thickness available on the Earth.

7 Conclusion

It is obviously impossible to give a detailed review of a variety of phenomena related to interactions of particles and radiation with matter in a brief chapter. Moreover, since many different physical processes contribute to these interactions, the subject itself becomes so multidisciplinary that its complete description is hardly possible in one, even large review. For more details and further references we can recommend the books mentioned in "Further Reading."

References

Akchurin N et al (2001) Electromagnetic shower profile measurements in iron with 500 MeV electrons. Nucl Instrum Meth A 471:303

Amaral P et al (2000) Hadronic shower development in iron-scintillator tile calorimetry. Nucl Instrum Meth A 443:51

Amsler C et al (2008) Passage of particles through matter. Phys Lett B 667:267

Baumgart R et al (1987) Performance characteristics of an electromagnetic streamer tube calorimeter. Nucl Instrum Meth A 256:254

Bethe HA (1930) Theory of the passage of fast corpuscular rays through matter. Ann Phys 5:325

Bethe HA (1932) Bremsformel für Elektronen mit relativistischen Geschwindigkeiten. Z Phys 76:293

Bethe HA (1953) Moliere's theory of multiple scattering. Phys Rev 89:1256

Bichsel H (1988) Straggling in thin silicon detectors. Rev Mod Phys 60:663

Bichsel H (2006) A method to improve tracking and particle identification in TPCs and silicon detectors. Nucl Instrum Meth A 562:154

Bloch F (1933) On the stopping of fast-moving particles in passage through matter. Z Phys 81:369

Davies H, Bethe HA, Maximon LC (1954) Theory of bremsstrahlung and pair production. 2. Integral cross section for pair production. Phys Rev 93:788

Gabriel TA et al (1994) Energy dependence of hadronic activity. Nucl Instrum Meth A 338:336

Gemmell DS (1974) Channeling and related effects in the motion of charged particles through crystals. Rev Mod Phys 46:129

Klein O, Nishina Y (1929) Über die Streuung von Strahlung durch Freie Elektronen nach der neuen relativistischen Quantenmechanik von Dirac. Z Phys 52:853

Landau LD (1944) On the energy loss of fast particles by ionization. J Exp Phys (USSR) 8:201

Longo E, Sestili I (1975) Monte Carlo calculation of photon-initiated electromagnetic showers in lead glass. Nucl Instrum Meth A 128:283

Møller SP (1994) Crystal channeling or how to build a '1000 Tesla Magnet'. Preprint CERN-94-05

Mott NF (1929) The scattering of fast electrons by atomic nuclei. Proc R Soc Lond A 124:425

Nelson WR et al (1985) The EGS4 code system. Preprint SLAC-R-265

Onuchin AP, Telnov VI (1974) Fluctuations of ionization losses in proportional chambers. Nucl Instrum Meth A 120:365

Rossi B (1952) High energy particle. Prentice-Hall, Englewood Cliffs

Sternheimer RM (1952) The density effect for ionization loss in materials. Phys Rev 88:851

Sternheimer RM, Peierls RF (1971) General expression for the density effect for the ionization loss of charged particles. Phys Rev B 3:3681

Tamm IE (1930) Über die Wechselwirkung der Freien Elektronen mit der Strahlung nach der Diracschen Theorie des Elektrons und nach der Quantenelektrodynamik. Z Phys 62:545

Tsai YS (1974) Pair production and bremsstrahlung of charged leptons. Rev Mod Phys 46:815

Vavilov PV (1957) Ionization losses of high energy heavy particles. Sov Phys JETP 5:749

Wigmans R (2000) Calorimetry: energy measurement in particle physics. Clarendon, Oxford

Further Reading

Grupen C, Shwartz B (2008) Particle detectors. Cambridge University Press, Cambridge

Kleinknecht K (1986) Detectors for particle radiation, Cambridge University Press, Cambridge

Knoll GF (2000) Radiation detection and measurement, 3rd edn. Wiley, New York

2 Electronics Part I

Helmuth Spieler
Lawrence Berkeley National Laboratory, Berkeley, CA, USA

1 Why Understand Electronics? ***26***

2 Detector Types ***28***

3 Signal Fluctuations ***29***

4 Signal Formation ***31***

5 Electronic Noise ***37***
5.1 Electronic Noise Levels 38
5.2 Noise in Amplifiers 40
5.3 Noise Versus Dynamic Range 42

6 Signal Charge Measurements ***43***
6.1 Charge-Sensitive Amplifiers 44
6.2 Noise in a Charge-Sensitive Amplifier System 45
6.3 Realistic Charge-Sensitive Amplifiers 46

7 Detector Equivalent Circuits ***48***
7.1 Thermistor Detecting IR Radiation 49
7.2 Piezoelectric Transducer 49
7.3 Ionization Chamber 49
7.4 Position-Sensitive Detector with Resistive Charge Division 49

8 Summary ***52***

References ***52***

C. Grupen, I. Buvat (eds.), *Handbook of Particle Detection and Imaging*,
DOI 10.1007/978-3-642-13271-1_2, © Springer-Verlag Berlin Heidelberg 2012

Abstract: Detectors come in many different forms and apply a wide range of technologies, but their principles can be understood by applying basic physics. In analyzing the signal acquisition, relatively simple models provide sufficient information to assess the effect of different readout schemes. This chapter discusses signal formation in various types of detectors and fluctuations in signal magnitude. It then moves on to baseline fluctuations, i.e. electronic noise, and the properties of amplifiers used for signal acquisition.

1 Why Understand Electronics?

Detector electronics is a highly developed field, and most users simply take it for granted. Many standard items are readily available or if custom systems are required, they can often be developed by applying standard recipes. However, real scientists understand their tools and if they are capable of applying basic physics to practical uses, understanding the electronic functions is quite practical without detailed knowledge of electronics engineering. It does require a real understanding of basic classical physics and the ability to recognize which aspects of physics apply in practical situations. For scientists and engineers to work together efficiently, it is necessary that scientists understand basic principles, so they do not request things that cannot work. Conversely, engineers should also understand the relevant aspects of the applications, so scientists have to be capable of explaining the important rerquirements. In detector systems that push the envelope, practical solutions are a balance between functions allowed by physics and the constraints of technology. However, revolutionary detector techniques often are not based on new inventions, but on combining existing technologies in novel ways. Understanding electronics can also help in recognizing subtle malfunctions that can fake physics results.

Radiation detectors are used for three basic functions: detecting the presence of radiation, measuring the energy spectrum, and recording the relative timing between events. These functions also provide position sensing. Energy resolution is one of the most useful properties, as shown in ❯ *Fig. 1*.

Energy resolution is determined by processes within the detector, where fluctuations in signal magnitude are inherent, and the readout system, where baseline fluctuations are superimposed on the signal. ❯ *Figure 2* shows the effect of noise added to a signal taken at four different times. The resulting peak amplitude can be both higher and lower. The effect of noise on timing measurements can also be seen. If the timing signal is derived from a threshold discriminator, where the output fires when the signal crosses a fixed threshold, amplitude fluctuations in the leading edge translate into time shifts. If one derives the time of arrival from a centroid analysis, the timing signal also shifts (compare the top and bottom right figures). Random fluctuations in the detector and in the baseline add in quadrature. ❯ *Figure 3* shows how either detector or readout baseline fluctuations can dominate to yield the same overall resolution. The baseline fluctuations can have many origins, external interference, artifacts due to imperfect electronics, etc., but the fundamental limit is electronic noise. Whether one or the other fluctuation dominates depends on the type of detector.

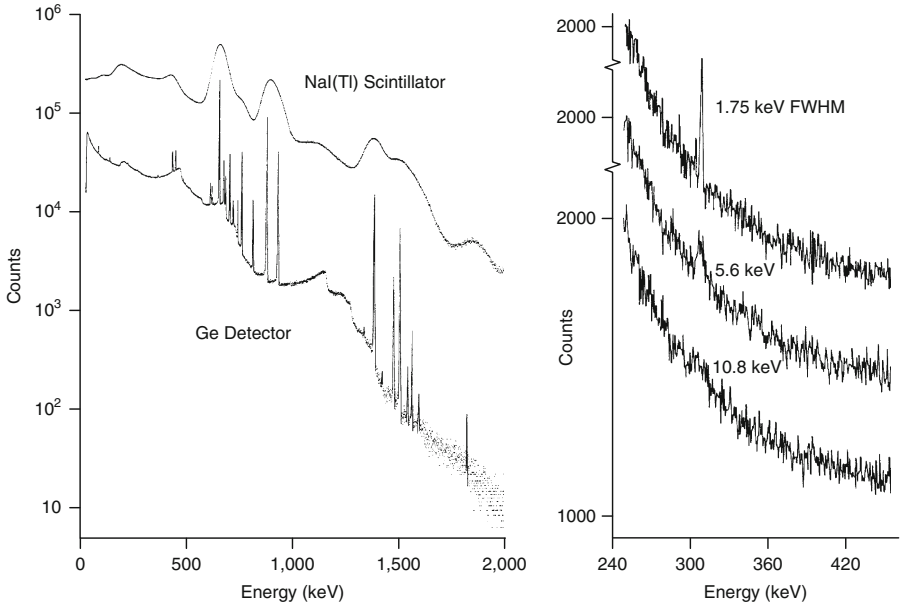


Fig. 1

The comparison of gamma-ray spectra taken with a scintillator and a semiconductor detector (*left*) clearly shows how improved resolution reveals detailed structure (adapted from Philippot 1970). Higher resolution also improves the signal-to-noise ratio (*right*). (Adapted from Armantrout et al. 1972. Figures ©IEEE, reprinted with permission)

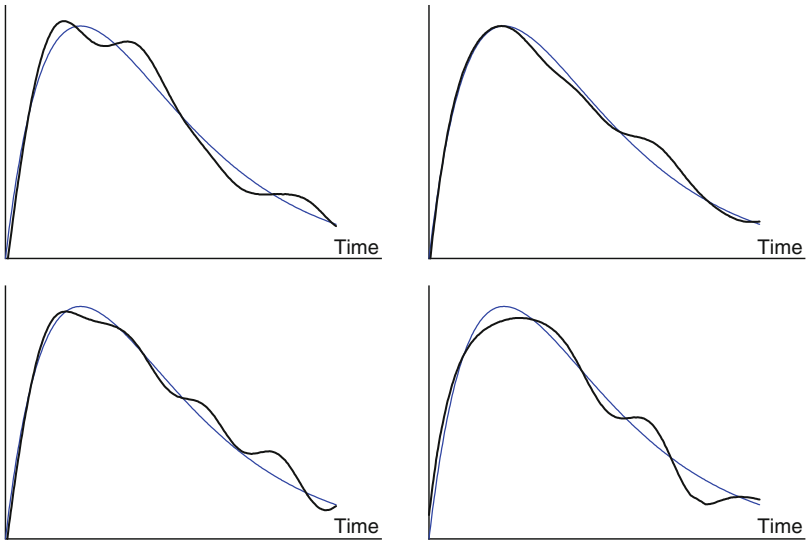


Fig. 2

Signal plus noise at four different times, shown for a signal-to-noise ratio of about 20. The noiseless signal is superimposed for comparison

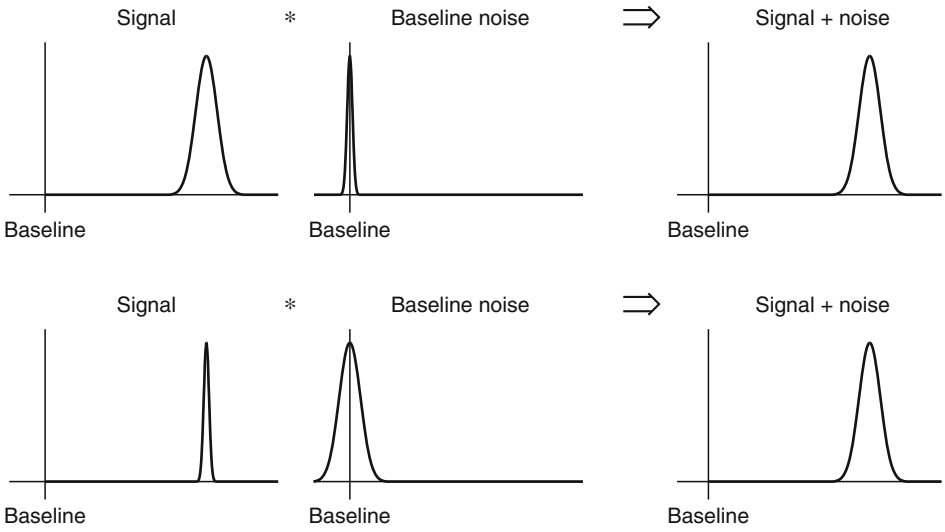


Fig. 3

Signal and baseline fluctuations add in quadrature. For large signal variance (*top*), as in scintillation detectors or proportional chambers, the baseline noise is usually negligible, whereas for small signal variance as in semiconductor detectors or liquid-Ar ionization chambers, baseline noise is critical

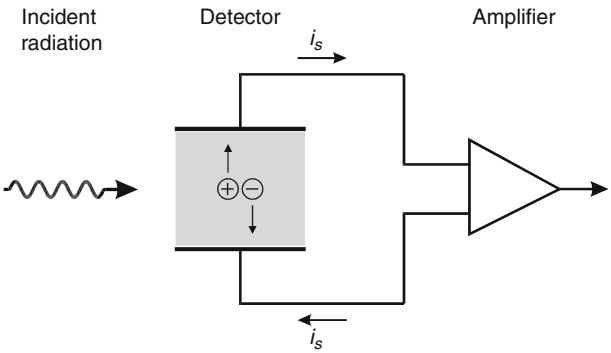


Fig. 4

In an ionization chamber, the absorbed energy is converted directly into signal charges

2 Detector Types

Detectors can transform absorbed energy directly or indirectly into signals. ❯ *Figure 4* shows the principle of an ionization chamber where absorbed radiation is converted directly into a charge signal. Here the statistical fluctuations in the number of signal charges place a fundamental limit on the energy resolution. A scintillation detector as shown in ❯ *Fig. 5* is an example of an indirect detector. In this case, another function comes into play that affects the energy resolution.

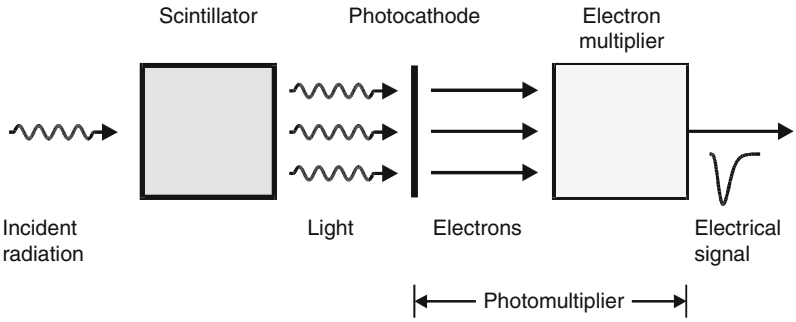


Fig. 5
In a scintillation detector, charged-particle or photon energy absorbed in the scintillator is converted into visible or near-visible light. When the scintillation photons impinge on the photocathode, they produce photoelectrons, which can be detected directly, or sent through a gain mechanism as in a photomultiplier

3 Signal Fluctuations

In a photomultiplier, the photocathode converts only a fraction of the impinging scintillation photons into photoelectrons. This quantum efficiency is commonly in the 10–30% range. Furthermore not all of the scintillation light ends up at the photocathode. Assume a 511 keV gamma ray absorbed in a NaI(Tl) scintillator:

- 25,000 photons are created in the scintillator
- 15,000 photons impinge on the photocathode
- 3,000 photoelectrons arrive at the first dynode
- $3 \cdot 10^9$ electrons appear at the anode

Defects and incomplete translation of absorbed energy into signal quanta can degrade detector resolution, but the inherent resolution of the detector cannot be better than the statistical variance in the number of signal quanta N produced by the absorbed energy E,

$$\frac{\Delta E}{E} = \frac{\Delta N}{N} = \frac{\sqrt{N}}{N} = \frac{1}{\sqrt{N}} . \tag{1}$$

The smallest quantity in the above sequence is the number of photoelectrons impinging on the first dynode, so in this example (see also Eq. 5)

$$\frac{\Delta E}{E} = \frac{1}{\sqrt{3{,}000}} = 2\% \text{ rms} \approx 5\% \text{ FWHM} . \tag{2}$$

The gain in a photomultiplier does not add much additional variance, so the relative resolution is the same at the anode, although the number of electrons is 10^6 times larger. Other nonstatistical mechanisms can also contribute to fluctuations. In many scintillators the ratio of scintillation photons to absorbed energy changes slightly with energy. In events where the absorbed energy is distributed over several separate interactions that deposit different energies, e.g. Compton scattering, these differences in the individual interactions lead to variations in the total scintillation light.

Equation 1 illustrates a basic mechanism. In general, however, additional considerations come into play when assessing statistical fluctuations.

$$\sigma_N = \sqrt{FN} = \sqrt{F\frac{E}{E_{SQ}}}, \tag{3}$$

where E_{SQ} is the energy required to form a signal quantum, e.g. electron–ion pair, and F is the Fano factor (Fano 1947), which in some detectors is < 1. Variance is limited relative to Eq. 2 because the absorbed energy sets an upper limit, but the largest reduction occurs in materials where energy is transferred by multiple paths with different excitation energies. For example, in silicon the bandgap is 1.2 eV, which is the energy required to excite an electron from the valence into the conduction band, i.e. the ionization energy. However, absorbed energy can also excite phonons, i.e. lattice vibrations linked to a temperature increase. In silicon the average phonon energy is 37 meV, i.e. orders of magnitudes smaller than the ionization energy. Since the total energy must be conserved, any fluctuation in the number of signal charges must be balanced by the fluctuation in the number of phonons. As the number of phonons is much greater, its relative variance is small and this reduces the overall fluctuations. In silicon and germanium $F = 0.1$. In liquid Xe $F \approx 20$, whereas in gaseous Xe $F \approx 0.15$. Since energy deposition also requires momentum conservation, the number of excited phonons is significant at energies well beyond the bandgap. Above several keV, the energy required to form a signal charge pair in silicon is 3.6 eV, i.e. about 2.8 times the bandgap. This ratio is the same for practically all other semiconductors and also holds for photons and different particle types. The Fano factor can be derived by formal mathematical procedures (e.g. Grupen and Shwartz 2008) or in a simplified form by physical processes (Spieler 2005, pp 52–55).

Expressed in terms of energy the signal fluctuation

$$\sigma_E = E_{SQ}\sqrt{F\frac{E}{E_{SQ}}} \tag{4}$$

and the relative energy resolution expressed as full width at half maximum (FWHM)

$$\frac{\Delta E}{E} = 2.35 \cdot \frac{E_{SQ}}{E}\sqrt{F\frac{E}{E_{SQ}}} = 2.35 \cdot \sqrt{F\frac{E_{SQ}}{E}}, \tag{5}$$

so the relative energy resolution improves with increasing energy.

Other detector mechanisms are the excitation of optical states in scintillators, lattice vibrations (phonons), the breakup of Cooper pairs in superconductors, and the formation of superheated droplets in superfluid He, for example. Typical energies required to form a single signal quantum are

- Ionization in gases: order 30 eV
- Ionization in semiconductors: 1–5 eV
- Scintillation: order 10 eV
- Phonons: meV
- Breakup of Cooper pairs: meV

Figure 6 shows the intrinsic energy resolution of Si and Ge detectors. When measuring the energy of a single deposition as in calorimeters, expressing the relative resolution $\Delta E/E$ is appropriate. However, when multiple additional energy peaks are near the desired energy

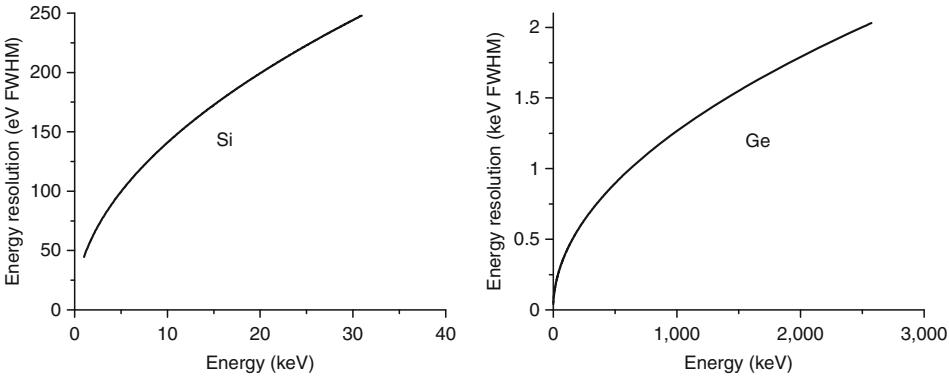


Fig. 6

Intrinsic energy resolution of silicon and germanium detectors versus energy. The silicon resolution is shown in eV and the Ge resolution in keV

measurement, expressing the energy resolution as an absolute number (e.g. eV or keV) is more useful, as it indicates how closely spaced other peaks may be to separate them.

4 Signal Formation

In a scintillation detector, the pulse shape is determined by the decay times of optical states. Multiple decay paths with non-radiative energy transfer are often involved before the final optical states are populated and decay. As a result, the overall pulse shape has a rise and a decay time as shown in *Fig. 7*, with the approximate form

$$I(t) \approx I_0 \left(e^{-t/\tau_d} - e^{-t/\tau_r} \right), \tag{6}$$

where τ_r and τ_d are the rise and decay time constants. Rise times are commonly in the ns range, whereas the decay times range from nanoseconds to microseconds. The rise time is often increased substantially by subsequent components in the system and variations in the path length in large scintillators, since the signal from a given event is often the combination of multiple interactions (e.g. Compton scattering). Depending on the decay sequences, some scintillators also spread the signal over numerous distinct pulses, so appropriate signal integration is essential.

The signal formation in ionization chambers, whether gaseous or solid state, is determined by the motion of the charges in the active region. Particles deposit energy in the detection volume, forming positive and negative charge carriers. Under an applied electric field, the charge carriers move and cause a change in induced charge on the electrodes. The duration of the induced signal depends on the carriers' velocity, which depends on the electric field. A high field in the detection volume is desirable for fast response, but also for improved charge collection efficiency.

As illustrated in *Fig. 8*, charge moving in the sensitive volume of the sensor gives rise to a signal current, as indicated in the accompanying equivalent circuit. At this point we need to determine $i_s(t)$. When does the signal current begin? When the charge reaches the electrode

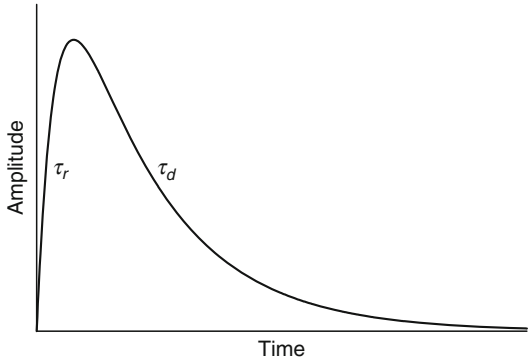


Fig. 7

Since the scintillation energy states are fed through non-emitting states, the scintillation pulse is characterized by a rise and fall time. In many scintillators the fall time is much longer than the rise time

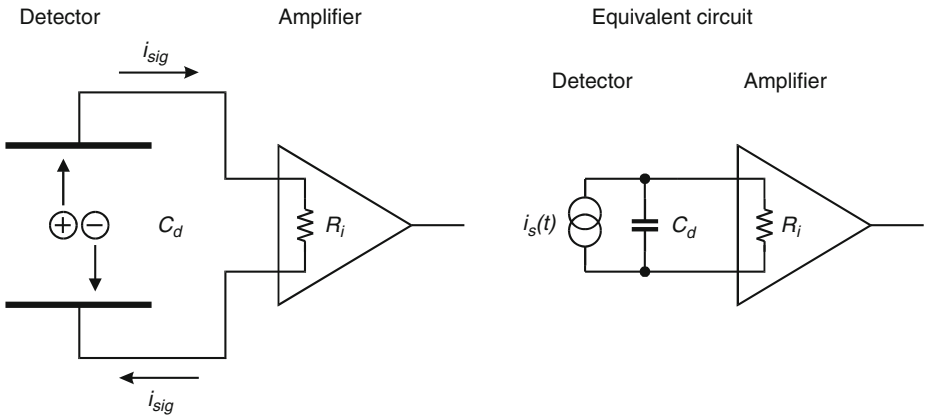


Fig. 8

Charge moving in the detector volume induces a signal current in the external circuit (*left*). The detector's equivalent circuit is shown at the *right*

or when the charge begins to move? Although the first answer is quite popular (encouraged by the phrase "charge collection"), the second is correct; current flow begins instantaneously.

To understand the physics of induced charge, first consider a charge q near a single, infinitely large electrode. All electric field lines from the charge terminate on the electrode. Integrating the field on a Gaussian surface S surrounding the charge yields

$$\oint_S \mathbf{E} \cdot \mathrm{d}\mathbf{a} = q\,.$$

Correspondingly, integrating over a Gaussian surface enclosing only the electrode yields the charge $-q$. Since the direction of the field lines is opposite relative to the first integral, this charge – the "induced charge" – has the opposite sign. Next, add a second electrode, as shown in *Fig. 9*. If the charge is positioned midway between the two electrodes, half of the field

lines will terminate on the upper and the other half on the lower electrode. Integrating over a Gaussian surface S_1 enclosing the upper electrode yields $-q/2$, as does integration around the lower electrode. If the charge is moved very close to the lower electrode, as in the right panel of ❯ *Fig. 9*, most of the field strength will terminate there, and the induced charge will be correspondingly higher. Thus, a charge moving from the upper to the lower electrode will initially induce most of its charge on the upper electrode, with an increasing proportion shifting to the lower electrode as the charge moves toward it.

We cannot observe the induced charge directly, but we can measure its change. If the two electrodes are connected to form a closed circuit, the change in induced charge manifests itself as a current. Integrating the induced current as the charge traverses the distance from the top to the bottom electrode yields the difference in induced charge q.

If the lower electrode is segmented, as illustrated in ❯ *Fig. 10*, charge is initially induced on all subelectrodes, but as the charge moves closer to the electrode on which it terminates, the charge distribution becomes more localized, and when the charge ends on the electrode, the total charge will be there.

The magnitude of the signal as a function of the charge's position relative to the readout electrode is determined by a "weighting function" that can be calculated easily by applying Ramo's theorem (Ramo 1939). The instantaneous current at a given electrode

$$i_k = -q\mathbf{v} \cdot \mathbf{F}_Q. \tag{7}$$

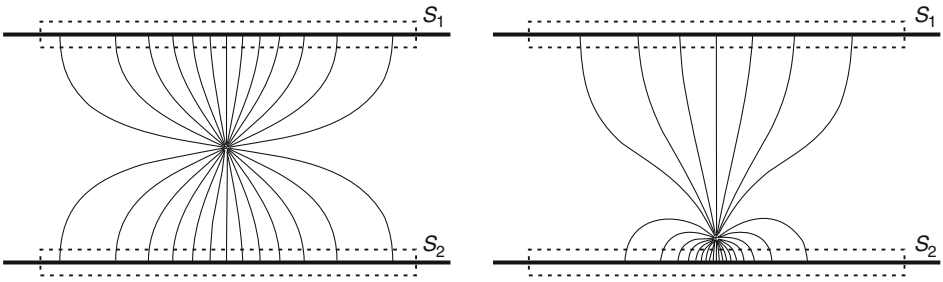


Fig. 9

A charge q positioned midway between two parallel plates induces equal charge on each plate (*left*). Integrating over the Gaussian surface S_1 or S_2 yields the induced charge $-q/2$. When positioned close to the bottom plate (*right panel*), more field lines terminate on the lower than on the upper plate, so the charge enclosed by S_2 is larger than the charge enclosed by S_1, i.e. the induced charge on the lower plate has increased

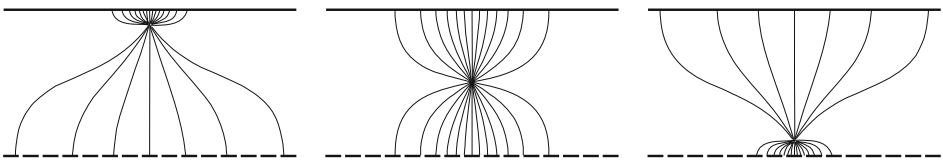


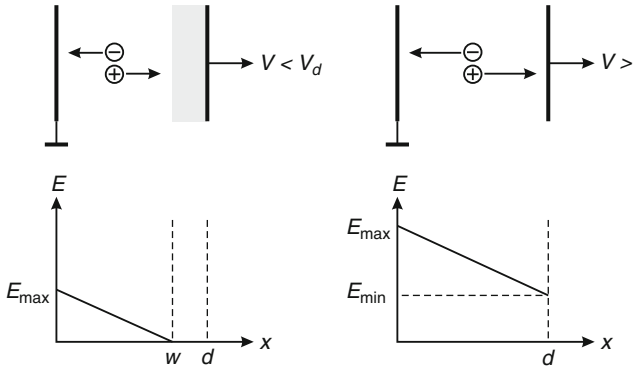
Fig. 10

In a segmented electrode, initially charge is induced on many strips (*left*). As the charge approaches the strips, it becomes more localized (*middle*), and when it is close to the strips the signal is concentrated on a few strips. The same principle applies to pixels

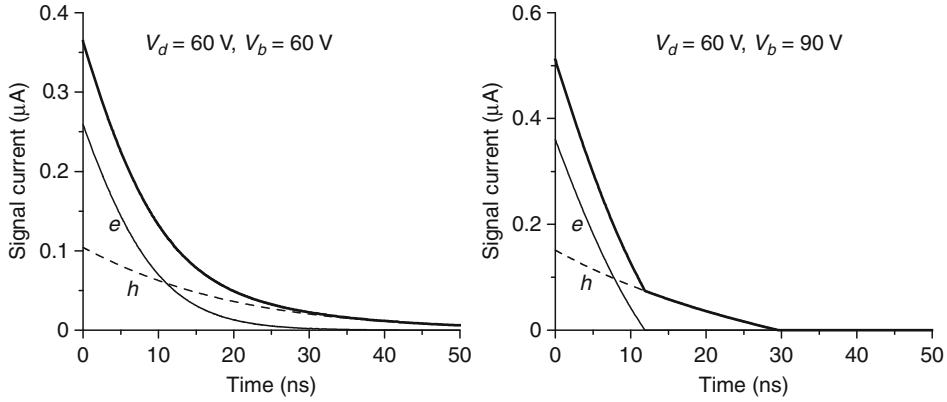
The weighting field F_Q is calculated by applying unit potential to the signal electrode and zero to all others. Note that for charge movement this field is irrelevant. It is not an electric field with the unit V/cm, but a field that is not derived from a physical quantity, so its unit is simply 1/cm. For a more detailed discussion, see Spieler (2005 pp 73–82).

In a simple parallel-electrode detector the weighting field is uniform, so for a constant charge velocity the signal current is constant. In a semiconductor detector, the field in the region depleted of mobile charge depends on the space charge of the atomic cores of the dopants, so it has a linear slope, which for overbias is superimposed on a constant field, as shown in ❯ *Fig. 11*.

Operating in partial depletion results in a zero field at the side opposite from the p–n junction, so to reduce the collection time operation with overbias is desirable. ❯ *Figure 12* shows



❑ Fig. 11
Electric field in a reverse-biased semiconductor diode in partial depletion (*left*) and with overbias (*right*)



❑ Fig. 12
Current signals for tracks traversing a semiconductor detector with parallel-plate electrodes and applied bias voltages of 60 and 90 V. The *left-hand* plot is at the depletion voltage, where the field at the ohmic contact is zero and the *right-hand* plot shows the effect of overbias

the resulting signal currents at bias levels of 60 and 90 V in a diode with a depletion voltage of 60 V. ❯ *Figure 13* shows the signal currents on both sides of a double-sided strip detector operating at the same voltages. Because of the different weighting field in the segmented detector, the pulse shapes are very different, although the overall collection time is the same. In all four signal examples, the integrated current yields the same charge.

Although induced signal currents are usually associated with detectors, the same physics applies to charge motion in any electrode configuration. Ramo derived the equations to analyze the signals in multigrid vacuum tubes. Another example is the photomultiplier tube (PMT).

❯ *Figure 14* shows the PMT configuration together with a typical resistive voltage divider that applies the dynode voltages. Electrons moving from the last dynode to the anode form the anode signal, as illustrated in ❯ *Fig. 15*. Note that the current path is closed through the bypass capacitor of the last dynode, so the practical implementation should be configured to

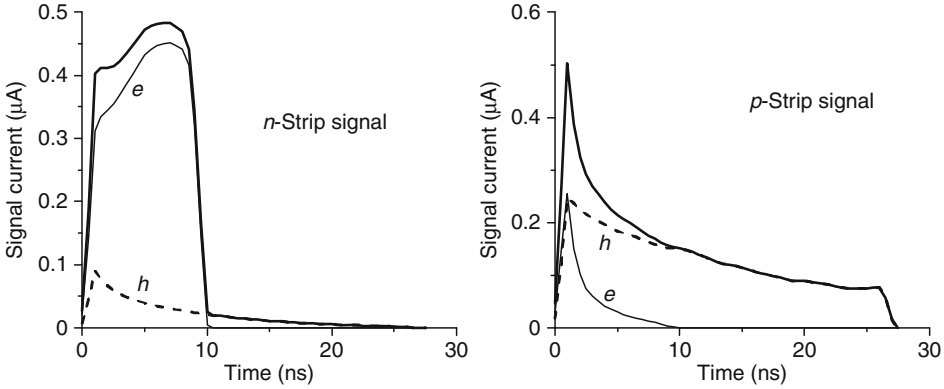


Fig. 13

Strip detector signals for an *n*-bulk device with 60 V depletion voltage operated at a bias voltage of 90 V. The electron (*e*) and hole (*h*) components are shown together with the total signal (*bold*). Despite marked differences in shape, the total charge is the same on both sides

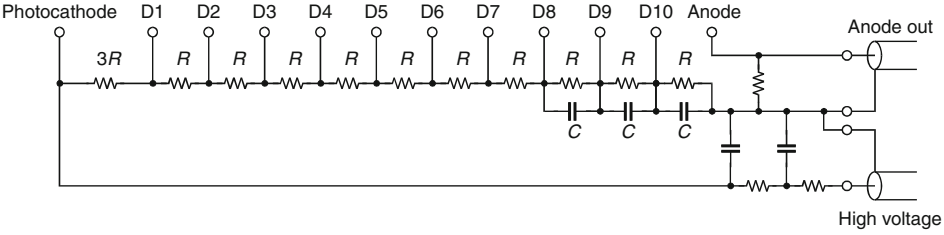


Fig. 14

Voltage divider circuit for a photomultiplier tube. The voltage ratios between the various electrodes are set by the resistance value *R*. D1 through D10 are the dynode connections. The resistor connected to the anode is included to avoid the anode charging up to a high voltage when the output is disconnected. When made much larger than the output load resistance, the signal fed to the readout will not be significantly reduced

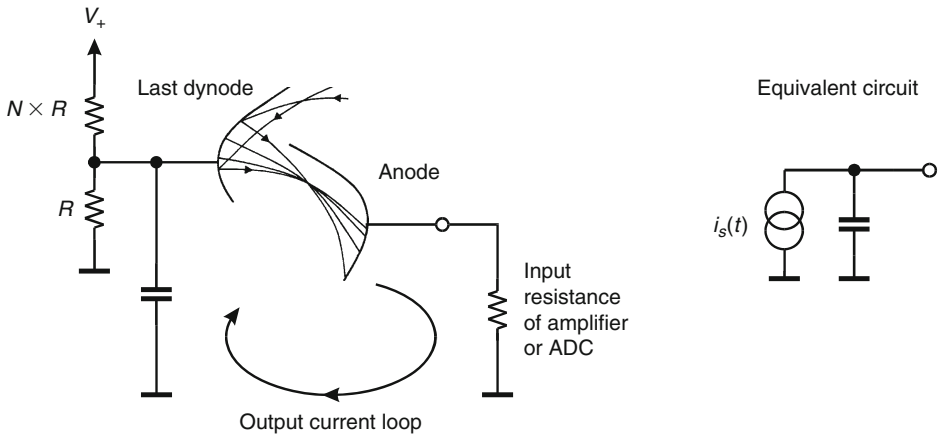


Fig. 15
The anode signal in a photomultiplier tube is formed by electrons moving from the last dynode to the anode. The equivalent circuit is shown at the *right*

keep it short, especially if fast pulse response is required. The circuit in *Fig. 14* is drawn accordingly.

The gain of a photomultiplier tube is very sensitive to the operating voltage. Given an individual dynode gain of G in the regime where the gain is proportional to the voltage difference between dynodes, the gain of a tube with N dynodes is G^N. Then the overall gain will be proportional to V^N, so fluctuations in the supply voltage V can degrade the energy resolution. Another contribution to changes in gain is the event rate. The charge required for the additional electrons emitted from each dynode is supplied by the voltage divider. At the final dynodes, this charge can be sufficient to cause a significant voltage drop. Just to estimate the current drawn by the dynode, assume a triangular output pulse with 1 V peak amplitude and a base width of 10 ns. Into a 50 Ω load the peak current I_{pk} = 20 mA. Also assume that the PMT is operating with a supply voltage of 1 kV and a divider current of 1 mA. Then the peak signal current of 20 mA is much larger than the current that can be provided by the voltage divider, so the gain will drop. This effect is reduced by the capacitors connected between the final dynodes. For a dynode voltage $V_d = V_n - V_{n-1}$ (e.g. the voltage difference between dynodes 10 and 9), the stored charge $Q = C \cdot V_d$. If this charge is much larger than the signal charge, then the voltage drop will be small. In this example, the signal charge at the last dynode is about 100 fC. If V_d = 100 V and C = 1 nF, the stored charge Q = 100 nC, so the relative voltage drop will be 10^{-3}, small with respect to a typical scintillator resolution. However, the capacitor has to charge up before the next pulse if the gain is to be maintained. The voltage will recover in time as $V(t) = V_d(1 - e^{t/\tau})$, where $\tau = RC$, i.e. about 100 μs (10^{-4} s), so the time between pulses should be at least 10^{-3} to 10^{-2} s. A more accurate calculation takes into account that since the total supply voltage remains constant, a reduced voltage at one dynode will change the voltage distribution and increase the voltages at the preceding dynodes, but the simple estimate performed above indicates where attention is required to maintain energy resolution. In principle the rate capability can be increased by raising the divider current, but power dissipation is a major constraint (1 W in the above example).

The voltages at the last dynodes are most susceptible. They can be stabilized by incorporating transistor drivers at the critical stages. Another scheme splits the voltage divider and feeds the final dynodes from a separate supply with a higher divider current. However, the simplest solution is to reduce the gain of the PMT and make up for it by feeding a low-noise amplifier. This has been a practical solution for decades, but not widely applied, since the combination of amplifier speed and electronic noise must be considered, and many PMT users believe that they don't have to understand low-noise electronics.

5 Electronic Noise

Although the mechanisms are different from signal formation in detectors, electronic noise is also caused by statistical fluctuations. Consider a current flowing through a sample bounded by two electrodes, i.e. n electrons with a total charge charge ne, moving with velocity v. The induced current depends on the spacing s between the electrodes (see "Ramo's theorem" discussed above), so

$$i = \frac{nev}{s}. \tag{8}$$

The fluctuation of this current is given by the total differential

$$\langle \mathrm{d}i \rangle^2 = \left(\frac{ne}{s} \langle \mathrm{d}v \rangle \right)^2 + \left(\frac{ev}{s} \langle \mathrm{d}n \rangle \right)^2, \tag{9}$$

where the two terms add in quadrature, as they are statistically uncorrelated. From this, one sees that two mechanisms contribute to the total noise: velocity, and number fluctuations.

Velocity fluctuations originate from thermal motion. Superimposed on the average drift velocity are random velocity fluctuations due to thermal excitation. The induced charge fluctuations can be translated from the Maxwellian velocity distributions and by applying the Wiener–Khinchine theorem in the Fourier transform to the frequency domain. However, a simpler procedure is to derive the maximum power that can be extracted from a thermal component in equilibrium by applying Planck's theory of blackbody radiation. The energy per mode at an absolute temperature T as a function of frequency f is

$$\overline{E} = \frac{hf}{\mathrm{e}^{hf/kT} - 1}, \tag{10}$$

where k is the Boltzmann constant. Thus, since $E = \int P \, \mathrm{d}t$, the spectral density of the emitted power

$$\frac{\mathrm{d}P}{\mathrm{d}f} = \frac{hf}{\mathrm{e}^{hf/kT} - 1}. \tag{11}$$

This is the power that can be extracted in equilibrium. At low frequencies $hf \ll kT$, the spectral power density

$$\frac{\mathrm{d}P}{\mathrm{d}f} \approx \frac{hf}{\left(1 + \frac{hf}{kT}\right) - 1} = kT, \tag{12}$$

so the spectral density is independent of frequency ("white noise"), and for a total bandwidth Δf, the noise power that can be transferred to an external device is $P_n = kT \cdot \Delta f$. The important result in this context is that the noise increases with bandwidth, so fast electronics where short rise times contain high-frequency Fourier components will have higher noise.

Number fluctuations occur in many circumstances. One source is carrier flow that is limited by emission over a potential barrier. Examples are thermionic emission or current flow in a semiconductor diode. The probability of a carrier crossing the barrier is independent of any other carrier being emitted, so the individual emissions are random and not correlated. Another mechanism applies in a reverse-biased diode, where the current is created by statistically independent generation and recombination processes. These number fluctuations are called "shot noise," which also has a "white" spectrum.

Another source of number fluctuations is carrier trapping. Impurities or imperfections in a crystal lattice can trap charge carriers and release them after a characteristic lifetime. Typically, multiple defects with a range of lifetimes are present. This leads to a frequency-dependent power spectrum $\mathrm{d}P_n/\mathrm{d}f = 1/f^{\alpha}$, where α is typically in the range of 0.5–2. Given a random distribution of trapping levels with a wide range of lifetimes, the noise power spectrum has a $1/f$ distribution (see Spieler 2005, pp 113–114).

5.1 Electronic Noise Levels

As noted above, both thermal and shot noise have "white" power spectra, i.e. if the noise power is measured with a fixed bandwidth, the magnitude is independent of frequency. A typical source of thermal noise is a resistor. For a resistance R the spectral noise power density

$$\frac{\mathrm{d}P_n}{\mathrm{d}f} = 4kT, \tag{13}$$

where k is the Boltzmann constant and T the absolute temperature. Since the power in a resistance R is

$$P = \frac{v^2}{R} = i^2 R, \tag{14}$$

the spectral voltage noise density

$$\frac{\mathrm{d}v_n^2}{\mathrm{d}f} \equiv e_n^2 = 4kTR \tag{15}$$

and the spectral current noise density

$$\frac{\mathrm{d}i_n^2}{\mathrm{d}f} \equiv i_n^2 = \frac{4kT}{R}. \tag{16}$$

The spectral noise density of shot noise

$$i_n^2 = 2eI, \tag{17}$$

where I is the average current and e the electronic charge. Note that the criterion for shot noise is that carriers are injected independently of one another, as in thermionic or semiconductor diodes. Current flow determined by an ohmic conductor ($I = V/R$) does not carry shot noise. Any local fluctuation in electron density relative to the stationary positive charge of the host atoms will set up an electric field that can easily draw in additional carriers to equalize the disturbance.

The above expressions for voltage and current noise can be derived in various ways (see Spieler 2005, pp 109–115). Other aspects that have to be considered when assessing the effect of current noise on the noise charge will be explained in Chap. 3, "Electronics Part II" on signal processing.

The spectral distribution of noise is a power density $\mathrm{d}P_n/\mathrm{d}f$, or in other words, the power in a narrow slice of frequency space. However, in analyzing electronic noise, we need to describe the noise in terms of voltage and current spectral densities $\mathrm{d}v_n/\sqrt{\mathrm{d}f}$ and $\mathrm{d}i_n/\sqrt{\mathrm{d}f}$. In circuit design literature and data sheets, these are commonly abbreviated as $\mathrm{d}v_n/\sqrt{\mathrm{d}f} \equiv e_n$ and $\mathrm{d}i_n/\sqrt{\mathrm{d}f} \equiv i_n$, so we'll follow that convention. This does lead to inconsistencies; i_s might represent a signal current with the unit A, whereas i_n represents a spectral density $\mathrm{A}/\sqrt{\mathrm{Hz}}$. In the following, just bear in mind that e_n and i_n have this special connotation.

The total noise is obtained by integrating the noise power over the relevant frequency range of the system, the bandwidth. Since power is proportional to either the voltage or current squared, the output noise of an amplifier with a frequency-dependent gain $A(f)$ is

$$v_{no}^2 = \int_0^\infty e_n^2 A^2(f)\,\mathrm{d}f \quad \text{or} \quad i_{no}^2 = \int_0^\infty i_n^2 A^2(f)\,\mathrm{d}f\,. \tag{18}$$

The total noise v_{no} or i_{no} increases with the square root of bandwidth. Since large bandwidths correspond to fast rise times, increasing the speed of a pulse measurement system will increase the noise.

In practice, amplifier stages limit the bandwith with both low- and high-frequency cutoff frequencies, beyond which the gain drops off. The distribution of bandwidths must also be considered. If a low-noise amplifier with a small bandwidth is feeding a wide-bandwidth stage prior to measurement, e.g. a wide-bandwidth digitizer, the noise of the second stage can dominate.

The frequency and time response of an amplifier are related as shown in *Fig. 16*. Beginning at low frequencies, basic amplifiers have a constant gain, which begins to drop off at a cutoff frequency f_u, beyond which the gain drops linearly with frequency. In the time domain, a fast-rise-time input pulse is slowed down to a rise time constant $\tau = 1/(2\pi f_u)$. At higher frequencies, additional cutoff frequencies may come into play, but well-designed amplifiers have a frequency response that is characterized by a single cutoff frequency ("pole") until the gain has dropped off significantly.

The basic electronic noise sources are expressed in terms of voltage and current, but the overall detector signal is a charge. The noise level can be expressed accordingly by measuring

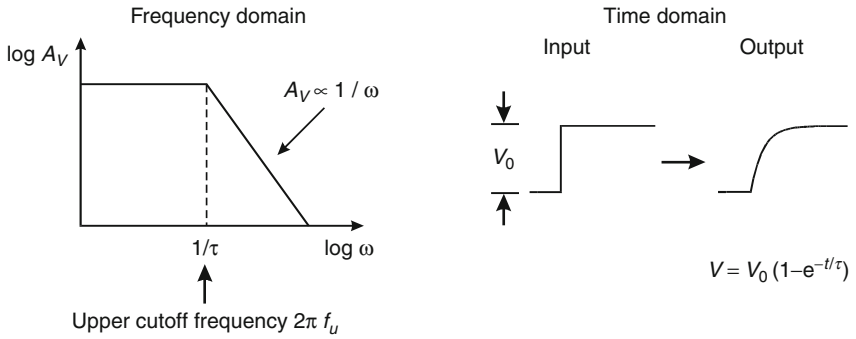


Fig. 16

The time constants of an amplifier affect both the frequency and the time response. Both are fully equivalent representations

the noise level at the output of the amplifier and the amplitude of the signal for a known input charge. From this measured signal-to-noise ratio the signal charge yielding a signal-to-noise ratio of one can be determined. This is the equivalent noise charge (ENC).

5.2 Noise in Amplifiers

The noise of an amplifier can be fully characterized in terms of a voltage and current noise source at the input as shown in Fig. 17. Typical magnitudes are $\mathrm{nV}/\sqrt{\mathrm{Hz}}$ and fA to $\mathrm{pA}/\sqrt{\mathrm{Hz}}$. Rather than specifying the total noise over the full bandwidth, the magnitude of each noise source is characterized by its spectral density. This is convenient because the effects of frequency-dependent impedances in the input circuit and of the amplifier bandwidth can be assessed separately.

The noise sources do not have to be physically present at the input. Noise also originates within the amplifier. Assume that at the output the combined contribution of all internal noise sources has the spectral density e_{no}. If the amplifier has a voltage gain A_v, this is equivalent to a voltage noise source at the input $e_n = e_{no}/A_v$.

First, let's assess the effect of external noise sources. Assume a signal fed to the amplifier input through a series resistance R_S, e.g. the resistance of a detector strip electrode or the series resistor used as part of an overvoltage protection circuit. Furthermore, a source of noise current is shunting the amplifier input. The equivalent circuit is shown in Fig. 18.

The signal at the input of the amplifier

$$v_{Si} = v_S \frac{R_i}{R_S + R_i} \,. \tag{19}$$

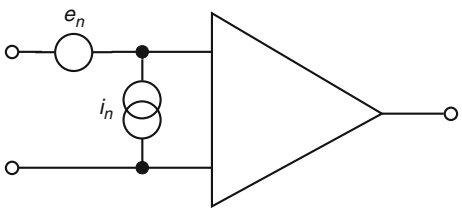


Fig. 17
An amplifier's noise can be fully characterized by assessing the effects of equivalent voltage and current sources at the input

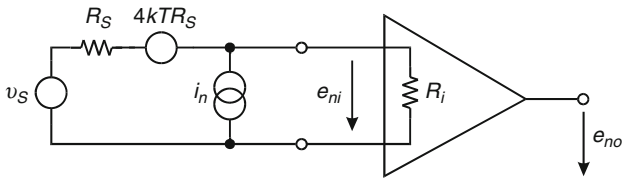


Fig. 18
The amplifier's noise equivalent circuit including a resistance in the signal path

The noise current flows through both the series resistance R_S and the amplifier input resistance R_i. Since the two are effectively in parallel, the resulting voltage at the input is the noise current flowing through the resistance of R_S and R_i in parallel. Since the contributions from independent, i.e. uncorrelated, noise sources add in quadrature, the noise voltage at the input of the amplifier

$$e_{ni}^2 = (4kTR_S)\left(\frac{R_i}{R_i + R_S}\right)^2 + i_n^2\left(\frac{R_i R_S}{R_i + R_S}\right)^2, \tag{20}$$

where the bracket in the i_n^2 term is the parallel combination of R_i and R_S. The signal-to-noise ratio at the output of the amplifier,

$$\left(\frac{S}{N}\right)^2 = \frac{A_v^2 v_{Si}^2}{A_v^2 e_{ni}^2} = \frac{v_S^2\left(\frac{R_i}{R_i + R_S}\right)^2}{(4kTR_S)\left(\frac{R_i}{R_i + R_S}\right)^2 + i_n^2\left(\frac{R_i R_S}{R_i + R_S}\right)^2},$$

$$\left(\frac{S}{N}\right)^2 = \frac{v_S^2}{(4kTR_S) + i_n^2 R_S^2}, \tag{21}$$

is the same as for an infinite input resistance, since the effect of the amplifier input resistance on the external noise arriving at the amplifier input is the same as for the signal. This result also holds for a complex input impedance, i.e. a combination of resistive and capacitive or inductive components. Since S/N is independent of amplifier input impedance, we can perform valid noise analyses using "idealized" voltage amplifiers with infinite input impedance.

When taking the amplifier noise into account, the origins of the noise shown in *Fig. 17* must be considered. In amplifiers the major contributions to the noise at the output originate inside the amplifier. The equivalent input noise is then the output noise divided by the amplifier gain. The contribution of this noise is independent of the input or source resistances. If an amplifier noise source is directly present at the input, e.g. the base current of a bipolar transistor, the noise current will flow through the source resistance and generate an input voltage, but not through the amplifier input resistance, as this is an integral part of the noise source. In general, the amplifier noise sources shown in *Fig. 17* only feed the impedances external to the amplifier. When analyzing a voltage-sensitive amplifier with a very high input impedance, the amplifier equivalent input noise voltage appears in series with the input signal and adds in quadrature with any other noise voltages in the input loop. The noise current originating at the input will flow through the source impedances resulting in a voltage that adds in quadrature.

With a current-sensitive amplifier, the output noise divided by the amplifier gain results in an equivalent input noise current, whose contribution in this case is independent of the source impedances. Any noise voltage sources in series with the input, e.g. resistors, will result in a noise current flowing through the input loop. This current will depend on the total source impedance.

Consider a chain of two amplifiers (or amplifying devices), with gains A_1 and A_2, and input noise levels N_1 and N_2, as shown in *Fig. 19*. When a signal S is applied to the first amplifier, the input signal-to-noise ratio is S/N_1. At the output of the first amplifier, the signal is $A_1 S$ and the noise $A_1 N$.

Both the signal and the noise are amplified by the second amplifier, but in addition the second amplifier contributes its noise, which adds in quadrature to the noise from the first stage. Then the signal-to-noise ratio at the output of the second amplifier

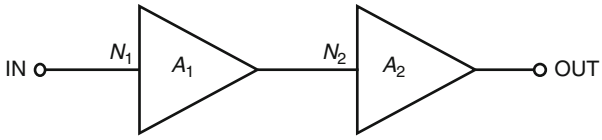


Fig. 19
In cascaded amplifiers the equivalent input noise of the first amplifier amplified by the first stage's gain can override the noise of the second stage

$$\left(\frac{S}{N}\right)^2 = \frac{(SA_1A_2)^2}{(N_1A_1A_2)^2 + (N_2A_2)^2} = \frac{S^2}{N_1^2 + \left(\frac{N_2}{A_1}\right)^2},$$

$$\left(\frac{S}{N}\right)^2 = \left(\frac{S}{N_1}\right)^2 \frac{1}{1 + \left(\frac{N_2}{A_1N_1}\right)^2}. \tag{22}$$

The second stage adds to the overall noise, but if the gain of the first stage is sufficiently high, it can be negligible. In a well-designed system the noise is dominated by the first gain stage.

5.3 Noise Versus Dynamic Range

Photomultiplier tubes often provide sufficient gain that they can directly feed an analog-to-digital converter (ADC) without an additional amplifier. However, ADCs also have internal noise sources, i.e. their input stage is an amplifier. Often the equivalent input noise is quite high, e.g. an order of magnitude higher than a good amplifier, and the ADC's resolution is worse than implied by the number of bits. The dynamic range is determined by the maximum signal level of the digitizer divided by the equivalent input noise. As noted above, the rate capability of a PMT can be increased by operating the PMT at a reduced gain and using a low-noise amplifier to bring the signal to the level required by the ADC. Adding an amplifier can reduce the overall noise, but the required gain can reduce the overall dynamic range.

Given the preamplifier's input noise of v_{n1} and gain G together with ADC's equivalent input noise v_{n2}, the resulting noise referred to the preamplifier input

$$v_n = \frac{\sqrt{(v_{n1}G)^2 + v_{n2}^2}}{G} = \sqrt{v_{n1}^2 + \left(\frac{v_{n2}}{G}\right)^2}, \tag{23}$$

so the ratio of the combined noise to the ADC noise

$$\frac{v_n}{v_{n2}} = \sqrt{\left(\frac{v_{n1}}{v_{n2}}\right)^2 + \left(\frac{1}{G}\right)^2}. \tag{24}$$

For example, for a gain of 10 and $v_{n1} = v_{n2}/10$, the resulting noise is reduced by a factor $\sqrt{2}/10$, so the minimum signal level is about seven times lower.

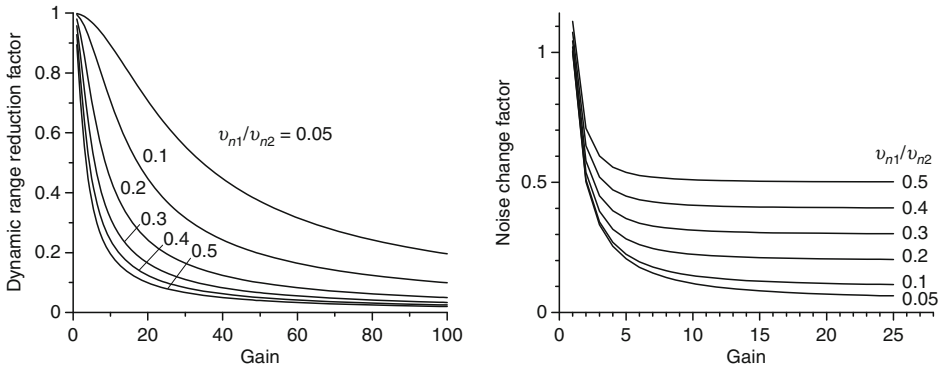


Fig. 20

Changes in dynamic range (*left*) and noise (*right*) versus gain for various noise ratios of the preamplifier to the ADC's equivalent input noise v_{n1}/v_{n2}

However the maximum signal level is reduced by the preamplifier gain G, so the dynamic range

$$\frac{V_i^{\max}}{v_n} = \frac{V_{i0}^{\max}/G}{\sqrt{v_{n1}^2 + \left(\frac{v_{n2}}{G}\right)^2}} = \frac{V_{i0}^{\max}}{\sqrt{(v_{n1}G)^2 + v_{n2}^2}} = \frac{V_{i0}^{\max}}{v_{n2}} \cdot \frac{1}{\sqrt{\left(\frac{v_{n1}}{v_{n2}}G\right)^2 + 1}}, \tag{25}$$

where the second factor is the reduction in dynamic range. For the above example with a gain of 10 and $v_{n1} = v_{n2}/10$, the overall dynamic range is $1/\sqrt{2}$ smaller. *Figure 20* shows the dynamic range and noise versus gain for various noise ratios v_{n1}/v_{n2}. For $v_{n1}/v_{n2} = 0.05$ and a gain of 5, the dynamic range is reduced by only 3% and the noise is reduced to 20.6% of the ADC noise. Doubling the preamp noise to $v_{n1}/v_{n2} = 0.1$ reduces the dynamic range by 11% and the noise is reduced to 22.4%.

6 Signal Charge Measurements

As shown in *Fig. 8*, the change in induced charge produces the signal in an ionization chamber. The form of the measured signal depends on the input time constant formed by the detector capacitance and the input resistance of the amplifier: $\tau = R_i C_d$. If the time constant is much smaller than the detector charge collection time, the detector capacitance will discharge quickly and the current flowing into the amplifier's input resistance will match the induced detector current. On the other hand, if the input time constant is much larger than the collection time, the signal charge Q_S will be stored on the capacitance C_d and produce a peak voltage $V_S = Q_S/C_d$. This illustrates that a given amplifier can be either current or voltage sensitive, depending on the detector capacitance, not its label.

If the amplifier is operating in the voltage mode with an equivalent input noise voltage v_n, then the signal-to-noise ratio

$$\frac{S}{N} = \frac{V_S}{v_n} = \frac{Q_S}{C_d} \cdot \frac{1}{v_n}, \tag{26}$$

which is inversely proportional to the capacitance. This is a general result for many systems, but not for the same reason, as illustrated in Sect. 6.2. In an ideal current-sensitive amplifier the noise can be independent of capacitance, but in reality the effect of parasitic noise sources can depend on detector capacitance, so it should always be considered.

A drawback of using a voltage amplifier to measure the signal charge is that the calibration depends on the capacitance. With a partially depleted detector the capacitance depends on the applied bias voltage, so the same energy deposition can yield different signal levels depending on the detector bias. Tracking detectors using semiconductor detectors often use varying strip lengths, so within a given system the calibration will vary. A solution to this problem is to use an amplifier whose ouput depends only on the signal charge, but not on the detector capacitance.

6.1 Charge-Sensitive Amplifiers

Figure 21 shows the principle of a charge-sensitive amplifier. The basic building block is an inverting voltage amplifier with a high input resistance. For simplicity assume an infinite input resistance, so that no signal current can flow into the amplifier. Since the amplifier inverts, the voltage gain $\mathrm{d}v_o/\mathrm{d}v_i = -A$, so $v_o = -Av_i$. A feedback capacitor C_f is connected from the output to the input. If an input signal produces a voltage v_i at the amplifier input, the voltage at the amplifier output is $-Av_i$. Thus, the voltage difference across the feedback capacitor $v_f = (A+1)v_i$ and the charge deposited on C_f is $Q_f = C_f v_f = C_f(A+1)v_i$. Since no current can flow into the amplifier's infinite input resistance, all of the signal current must charge up the feedback capacitance, so $Q_f = Q_i$. The amplifier input appears as a "dynamic" input capacitance

$$C_i = \frac{Q_i}{v_i} = C_f(A+1)\,. \tag{27}$$

The enhanced input capacitance corresponds to a reduction in input impedance $1/(\omega C_i)$, as expected for a shunt feedback amplifier.

The voltage output per unit input charge

$$A_Q = \frac{v_o}{Q_i} = \frac{Av_i}{C_i v_i} = \frac{A}{C_i} = \frac{A}{A+1}\cdot\frac{1}{C_f} \approx \frac{1}{C_f}\quad (A \gg 1)\,, \tag{28}$$

so the charge gain is determined by a well-controlled component, the feedback capacitor.

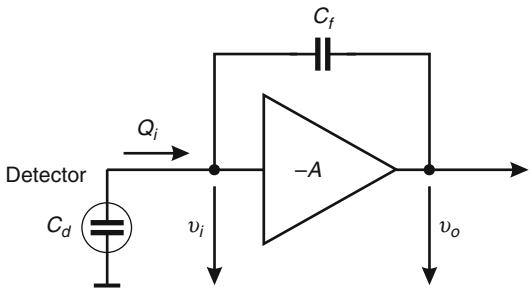


Fig. 21
Principle of a charge-sensitive amplifier

The signal charge Q_S will be distributed between the sensor capacitance C_d and the dynamic input capacitance C_i. The ratio of measured charge to signal charge

$$\frac{Q_i}{Q_s} = \frac{Q_i}{Q_d + Q_{s,\mathrm{amp}}} = \frac{C_i}{C_d + C_i} = \frac{1}{1 + \dfrac{C_d}{C_i}}, \tag{29}$$

so the dynamic input capacitance must be large compared to the sensor capacitance.

Another very useful feature of the integrating amplifier is the ease of charge calibration. By adding a test capacitor as shown in ❯ *Fig. 22*, a voltage step injects a well-defined charge into the input node. If the dynamic input capacitance C_i is much larger than the test capacitance C_T, the voltage step ΔV at the test input will be applied nearly completely across the test capacitance C_T, thus injecting a charge $C_T \Delta V$ into the input. More precisely, the injected charge

$$Q_T = \frac{C_T}{1 + \dfrac{C_T}{C_i + C_d}} \cdot \Delta V \approx C_T \left(1 - \frac{C_T}{C_i + C_d}\right) \Delta V, \tag{30}$$

so for the best accuracy the system should be calibrated with the detector connected.

6.2 Noise in a Charge-Sensitive Amplifier System

Unlike an ideal voltage-sensitive amplifier, where the noise is independent of the detector capacitance, but the signal voltage is inversely proportional, $V_S = Q_S/C$, in a charge-sensitive amplifier the dependence of the signal output voltage on the detector capacitance is small. However, in this case the noise increases with detector capacitance since the detector is part of the feedback loop as shown in ❯ *Fig. 23*. When no detector is connected to the input, all of the output noise is fed back to the input, resulting in an additional correlated noise of opposite sign at the output, which reduces the output noise. With a detector connected, the feedback capacitor

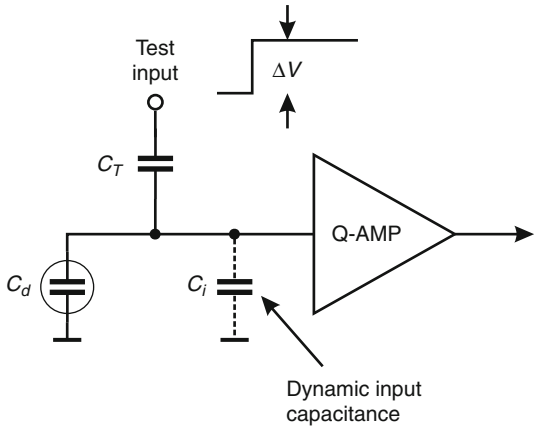


Fig. 22

Adding a test input to a charge-sensitive amplifier provides a simple means of absolute charge calibration

and the detector capacitance form a voltage divider, so only a fraction of the output noise is fed to the input and less negative feedback results in a noise increase. In circuits of this type, where the signal source is part of the feedback loop, the noise gain differs from the signal gain.

For a more detailed description of the calculation, see Spieler (2005, pp 123–125). Given a sufficiently high open-loop amplifier gain A to make the charge gain $A_Q = 1/C_f$, the signal-to-noise ratio

$$\frac{Q_s}{Q_{ni}} = \frac{1}{F_S}\frac{Q_s}{e_{nf}(C_d + C_f)} \approx \frac{1}{F_S}\frac{1}{C_d}\frac{Q_s}{e_n} \quad (C_d \gg C_f), \tag{31}$$

where the factor $F_S = A_{VS}\sqrt{\Delta f_n}$ combines the noise bandwidth and gain that characterize the system. This is the same result as for a voltage-sensitive amplifier, *but here the signal is constant and the noise grows with increasing* C_d. However, note that the additional feedback capacitor adds to the detector capacitance in determining the noise.

6.3 Realistic Charge-Sensitive Amplifiers

The preceding discussion assumed that the amplifiers are infinitely fast, so they respond instantaneously to the applied signal. In reality this is not the case; charge-sensitive amplifiers often respond much more slowly than the time duration of the current pulse from the sensor. However, as shown in ❯ *Fig. 24*, this does not obviate the basic principle. Initially, signal charge is integrated on the sensor capacitance, as indicated by the left-hand current loop. Subsequently, as the amplifier responds the signal charge is transferred to the amplifier. Thus, the signal charge

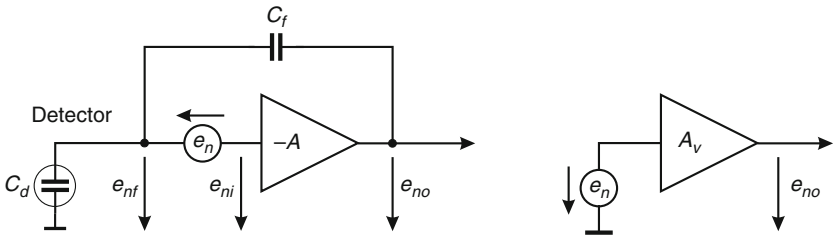


Fig. 23
Circuit for the noise analysis of a charge-sensitive amplifier. The convention for setting the phase of e_n is shown at the *right*

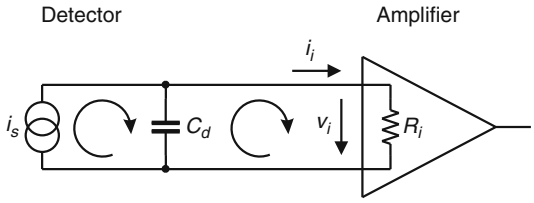


Fig. 24
Charge integration in a realistic charge-sensitive amplifier. First, charge is integrated on the sensor capacitance and subsequently transferred to the charge-sensitive loop, as it becomes active

is preserved and the full signal appears at the amplifier output, even if the amplifier is much slower than the collection time.

At low frequencies, where the amplifier's open-loop phase shift is 180°, the input impedance of a charge-sensitive amplifier utilizing capacitive shunt feedback is capacitive. However, the open-loop corner frequency where the gain begins to roll off and introduces a 90° phase shift is typically well below the passband set by the shaper, approximately centered at the inverse peaking time. There the input impedance is resistive. The relevant frequency range is determined by the pulse shaper. Some common examples are shown in ❯ Chap. 3, "Electronics Part II." ❯ *Figure 25* shows the input impedance versus frequency for an amplifier with an open-loop corner frequency of 100 kHz. At low frequencies, where the amplifier phase shift is 180°, the input impedance is capacitive and decreases with frequency. Beyond the amplifier cutoff frequency where its phase shift is 90°, the input impedance levels off and is resistive (phase shift 0°). Its value

$$Z_i = \frac{1}{\omega_0 C_f} \tag{32}$$

depends on the amplifier's unity gain frequency, extrapolated from the frequency regime where the gain falls off linearly with frequency, as shown in in the left panel of ❯ *Fig. 25*. Note that the amplifier also has a second pole (cutoff frequency) at 100 MHz, which reduces the gain more rapidly at higher frequencies. The low-frequency gain of the amplifier is 1,000, which improves the baseline stability, but at a peaking time of 20 ns, corresponding to a mid-band frequency of about 10 MHz, the resistive input impedance is about 1.6 kΩ. This is also the impedance presented by a capacitance of 10 pF, so in a 6 cm-long silicon strip detector, about half of the signal current will go to the neighbors. The mid-band frequency scales inversely with the peaking time. It depends on the type of shaper and, as noted above, examples are shown in ❯ Chap. 3, "Electronics Part II."

The frequency response shown in ❯ *Fig. 25* was chosen in a circuit simulation to demonstrate the effect of the open-loop phase shift. Today low-power amplifiers with a much higher

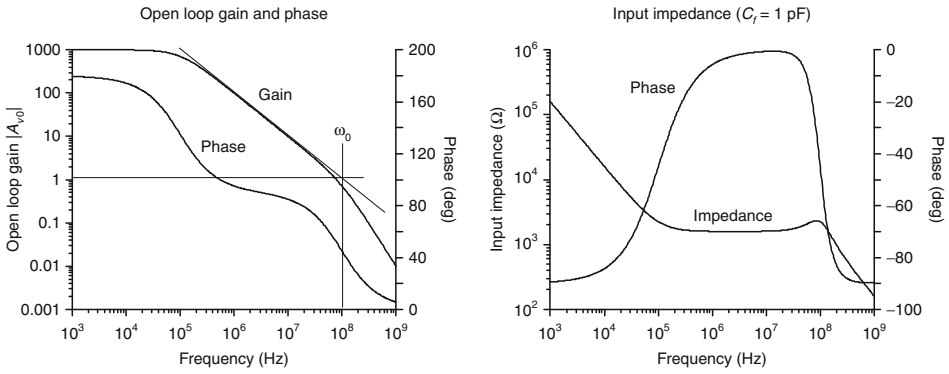


Fig. 25

At low frequencies where the amplifier's open loop phase shift is 180°, the input impedance is capacitive (phase shift 90°). Beyond the amplifier cutoff frequency where an additional phase shift of –90° is introduced, the input impedance levels off and is resistive (phase shift 0°)

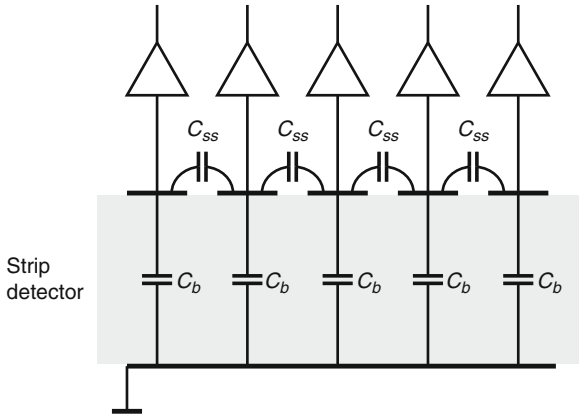


Fig. 26
The amplifier input impedance determines cross-coupling between channels. It must be small compared to the impedance presented by the inter-electrode capacitance C_{ss}

unity gain frequency ω_0 are quite feasible, so sufficiently low input impedances can be achieved. However, this should not be taken for granted.

Besides optimizing charge transfer to the front-end amplifier, the input impedance is also important in strip and pixel detectors because it determines cross-talk between adjacent electrodes. As shown in *Fig. 26*, if the amplifier had an infinite input impedance, the signal current from a given electrode would simply flow to the neighbors and the charge deposited on the neighbor electrodes would depend on the inter-electrode capacitance C_{ss} and the backplane capacitance C_b. To capture the bulk of the signal charge on the target electrode, the amplifier's input impedance must be small compared to the impedance presented by the inter-electrode capacitance C_{ss}. This impedance $1/(\omega C_{ss})$ depends on the range of frequencies at which the signal amplitude is measured, i.e. the shaping time.

The above example illustrates the relevant parameters. The effect of the input impedance is illustrated for a 20 ns peaking time, but in this example the resistive impedance will be effective at shaping times up to 1 μs. Whether the effective input impedance meets the specific application's requirements depends on the preamplifier and the signal shaper, so the complete system must be analyzed both in the time and frequency domain.

7 Detector Equivalent Circuits

In calculating the signal-to-noise ratio, the magnitude of the signal applied to the amplifier input must also be assessed. As indicated in *Fig. 18* and discussed in the preceding section, the equivalent circuit of the detector affects both the signal and the noise. *Figure 15* illustrates how the signal from a PMT can be described by an equivalent circuit, where the anode is a current source with a shunt capacitance formed by the capacitance of the last dynode to the anode plus any additional stray capacitance. The same equivalent circuit can be applied to a microchannel plate. Simple circuits can often represent rather complex detectors.

7.1 Thermistor Detecting IR Radiation

A thermistor is a resistor whose value depends on temperature. When biased with a constant current, the voltage across the thermistor is proportional to the thermistor resistance. Constant current bias can be simply implemented by feeding the thermistor from a voltage source through a resistor whose value is much larger than the thermistor resistance R_T, so a change in R_T will not significantly affect the bias current. ❯ *Figure 27* shows the actual circuit and the model.

7.2 Piezoelectric Transducer

In a piezoelectric transducer, the molecular dipole moments of the sensitive material create a change in voltage across the sensitive volume when the material is stressed. ❯ *Figure 27* shows the equivalent circuit.

7.3 Ionization Chamber

Gas-filled ionization or proportional chambers directly convert absorber radiation into charge, whose motion through the detector volume induces a signal current. The same principle applies to semiconductor detectors, ranging from simple two-electrode configurations to strips or pixels. ❯ *Figure 28* shows the basic detector configuartion and the equivalent circuit.

7.4 Position-Sensitive Detector with Resistive Charge Division

The input impedance is also important in other applications, for example using charge division to measure the z coordinate in long-strip detectors. This technique is less susceptible to "ghosting" than crossed-strip configurations, as in crossed strips the probability of "ghosting"

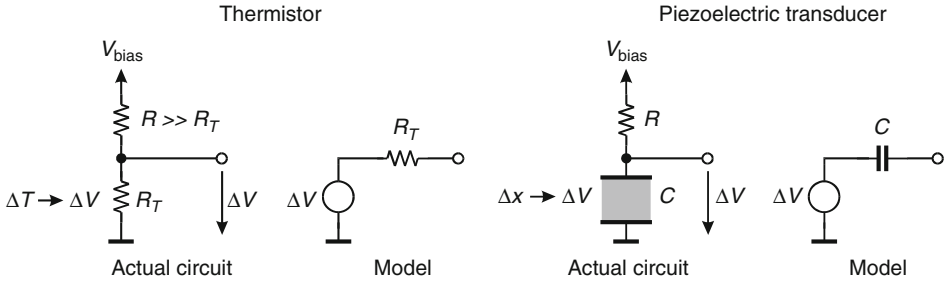


Fig. 27
***Left panel*: Equivalent circuit of a thermistor biased at a constant current. *Right panel*: Equivalent circuit of a piezoelectric transducer. The bias voltage is applied through a high-value resistor**

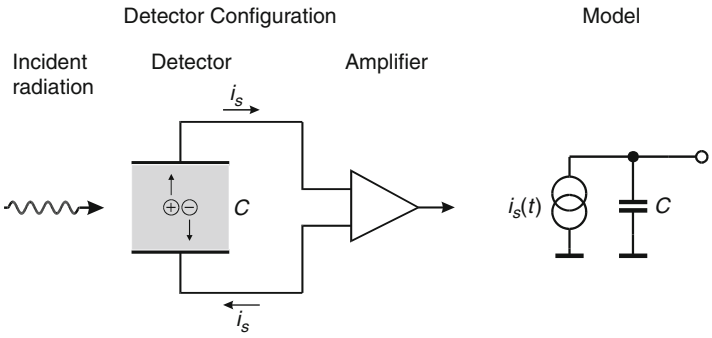


Fig. 28

In an ionization chamber the charge formed by the absorbed radiation moves under the influence of the applied electric field and the induced current appears at the detector electrodes. The equivalent current source is shunted by the capacitance formed by the detector electrodes

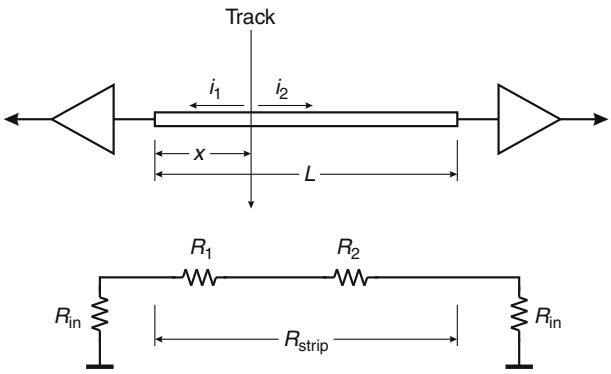


Fig. 29

The track coordinate can be determined by measuring the ratio of currents at both ends of the strip. Maximum position sensitivity requires a low amplifier input impedance

is proportional to the area subtended by the crossed strips, whereas here multiple-hit probability depends only on the area covered by the strip and its neighbors. It also reduces the material and mechanical complexity of crossed-strip configurations. However, several interlinked contributions must be considered in optimizing the design, and it is easy to obtain inferior results and discredit the technique. This technique is not necessarily an adequate replacement for crossed strips and can easily require a higher signal-to-noise ratio, so it must be evaluated in detail.

The principle is illustrated in *Fig. 29*. The position sensitivity is given by

$$\frac{\mathrm{d}x}{\mathrm{d}(i_1/i_2)} = -L\frac{\left(\frac{x}{L}+\frac{R_{\mathrm{in}}}{R_{\mathrm{strip}}}\right)^2}{1+2\frac{R_{\mathrm{in}}}{R_{\mathrm{strip}}}}, \tag{33}$$

so the input impedance must be small with respect to the strip resistance. *Figure 30* shows the simplest equivalent circuit together with a more complex version that takes the distributed resistance and capacitance into account in applications where a fast preamplifier is used.

For a given input impedance one could choose a higher strip resistance. However, noise is also an issue. With increasing strip resistance the thermal noise of the strip resistance increases. The noise from the strip is anticorrelated in the two readout channels, so for a given signal-to-noise ratio this puts an upper limit on the strip resistance. On the other hand, in a shunt-feedback amplifier, the load, whether capacitive or resistive, that is presented to the input determines the noise gain, so the strip resistance in series with the input impedance of the amplifier at the other end forms a load to the preamplifier, in addition to the strip capacitance. This increases the amplifier noise and constrains the minimum strip resistance. This effect can be reduced by lowering the noise of the preamplifier. If this is done by increasing the current in the input transistor, i.e. its transconductance, the gain–bandwidth product will also increase, which will lower the input impedance.

Reducing the amplifier noise is also useful for another reason, as the strip resistance cross-couples noise from the two amplifiers, as shown in *Fig. 31*. This will typically increase noise by about $\sqrt{2}$, somewhat reduced by the strip-to-backplane capacitance. Cross-coupling of noise from neighboring strips through the strip-to-strip capacitance is another contribution. For a more detailed discussion of noise cross-coupling, see Spieler (2005 pp 129–132).

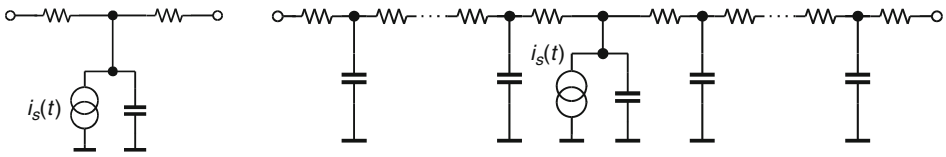


Fig. 30

A simple equivalent circuit is shown at the *left*. When using a fast readout, it may be necessary to include the distributed resistance and capacitance as shown to the *right*

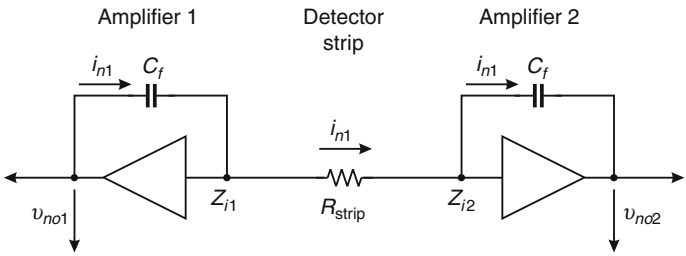


Fig. 31

Noise from one amplifier is cross-coupled to the other. A noise voltage at the output of one amplifier will inject a current through the feedback capacitor and strip resistance into the other amplifier. Note that for a signal originating from its output, the left-hand amplifier presents a high impedance at its input, so the current i_{n1} flows to the right-hand amplifier, which presents a low impedance to an external signal (see Spieler 2005)

As noted above, charge division can offer advantages in certain applications. However, details depend on the application requirements, and a more sophisticated analysis than in standard strip detectors is needed to achieve optimum results.

8 Summary

Optimizing a detector system requires an understanding of both the detector physics and the characteristics of the readout electronics. In the preceding discussions the basic sources of fluctuations were discussed. However, when electronic noise is a significant contribution, signal processing is a key element. This is discussed in Chap. 3, "Electronics Part II."

References

Fano U (1947) Ionization yield of radiations. II. The fluctuations of the number of ions. Phys Rev 72:26–29

Grupen C, Shwartz BA (2008) Particle detectors, 2nd edn. Cambridge University Press, Cambridge

Ramo S (1939) Currents induced by electron motion. Proc IRE 27:584–585

Spieler H (2005) Semiconductor detector systems. Oxford University Press, Oxford

3 Electronics Part II

Helmuth Spieler
Lawrence Berkeley National Laboratory, Berkeley, CA, USA

1 Basic Principles of Signal Processing ***54***

2 Signal Processing ***56***

3 Noise Analysis of a Detector–Preamplifier–Shaper System ***60***

4 Timing Measurements ***64***

5 Digital Electronics ***66***
5.1 Logic Elements 66
5.2 Propagation Delays and Power Dissipation 69
5.3 Logic Arrays 70

6 Analog-to-Digital Conversion ***71***

7 Time-to-Digital Converters (TDCs) ***74***

8 Digital Signal Processing ***76***

9 Summary ***80***

References ***81***

C. Grupen, I. Buvat (eds.), *Handbook of Particle Detection and Imaging*,
DOI 10.1007/978-3-642-13271-1_3, © Springer-Verlag Berlin Heidelberg 2012

Abstract: Signal processing is needed to optimize energy and time resolution. This chapter discusses the basic principles of signal processing and the resulting electronic noise, beginning with analog systems and then moving on to digital electronics and digital signal processing.

1 Basic Principles of Signal Processing

Radiation impinges on a sensor and creates an electrical signal, either directly or indirectly, as discussed in Chap. 2, "Electronics Part I." Especially in high-resolution systems, the signal level is often low and must be amplified to allow digitization and storage. Both the sensor and the electronics introduce signal fluctuations, i.e. fluctuations in the signal magnitude introduced by the sensor and noise from the electronics that is superimposed on the signal. The detection limit and measurement precision are determined by the signal-to-noise ratio.

Electronic noise affects all types of measurements. To merely detect the presence of a hit, the noise level determines the minimum threshold. If the threshold is set too low, the output will be dominated by noise hits. In energy measurements the noise "smears" the signal amplitude, and in time measurements noise alters the time dependence of the signal pulse, as shown in Chap. 2, "Electronics Part I."

To increase the signal-to-noise ratio one can increase the signal, reduce the noise, or both. Problems are often best solved by viewing components from different perspectives. Since pulses are viewed in the time domain, whereas noise is characterized in the frequency domain, optimizing the signal-to-noise ratio is best addressed in both the time and frequency domain. *Figure 1* shows a unipolar and bipolar pulse in the time and frequency domains. The upper extent of the frequency distribution depends on the maximum slope of the signal pulse, as a fast change in time requires Fourier components at high frequencies. The unipolar pulse has a net charge, so the frequency spectrum extends all the way down to zero. The bipolar pulse has zero net area, so in the frequency domain the distribution goes down to zero at zero frequency. The increased slope at the zero crossing extends the spectrum to higher frequencies.

Figure 2 shows a typical noise spectrum in the frequency domain. A constant noise level extends over the mid-range, but noise increases at low frequencies due to "$1/f$" noise and also increases at high frequencies because the amplifier gain is decreasing.

As noted in Chap. 2, "Electronics Part I," operating with excessive bandwidth will increase the noise level without improving the signal level. Since the signal and noise spectra differ, minimizing noise will also reduce the signal amplitude, so the choice of the optimum frequency response requires balancing between reduction in noise and loss of signal. This optimum depends on the measurement goal. For example, a short detector pulse would imply a fast, i.e. high bandwidth, electronics response. This would be appropriate when signal timing is important. However, if only the magnitude of the signal charge is to be measured, the fast pulse can be integrated to form a long-duration pulse whose amplitude is proportional to the signal charge, as obtained from a charge-sensitive amplifier. This allows a subsequent amplifier chain with a smaller bandwidth, which will yield a lower noise level.

However, most radiation measurements are not restricted to single pulses, but to a random sequence in time. This imposes an additional constraint on signal processing as the duration of a signal pulse determines the probability of pulse pileup, as illustrated in *Fig. 3*.

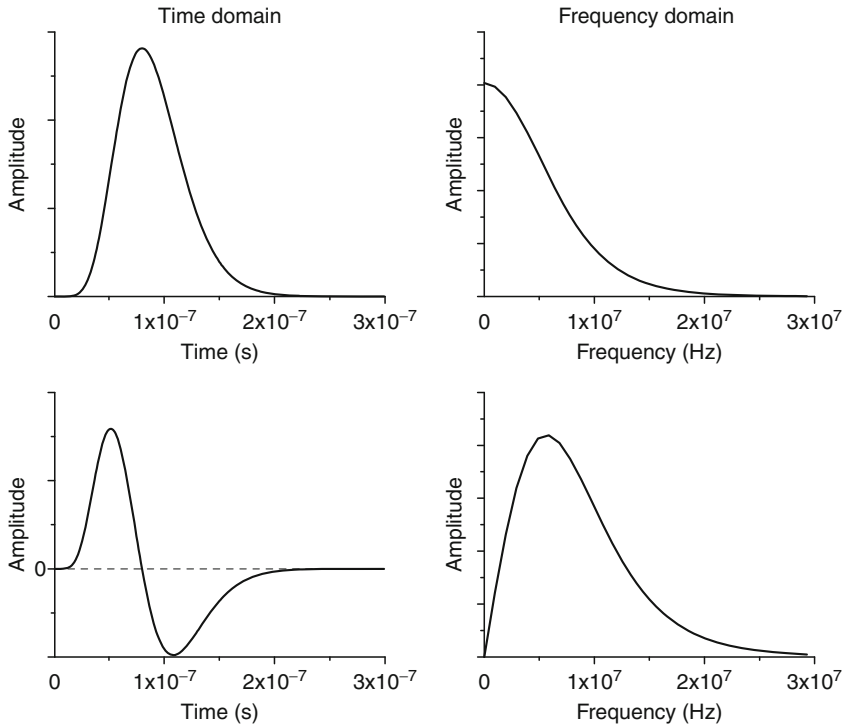


Fig. 1

Signal distribution in both the time and frequency domains, a unipolar pulse in the *upper row* and a bipolar pulse *below*. The bipolar pulse has a zero net area, so the amplitude in the frequency domain goes to zero at zero frequency. The bipolar pulses slope is greater at the zero crossing, so the frequency spectrum extends to higher frequencies

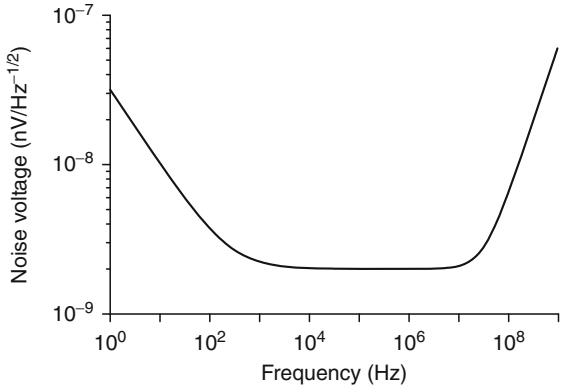


Fig. 2

A typical noise spectrum in the frequency domain. Distinguishing the individual components in the time domain is more difficult

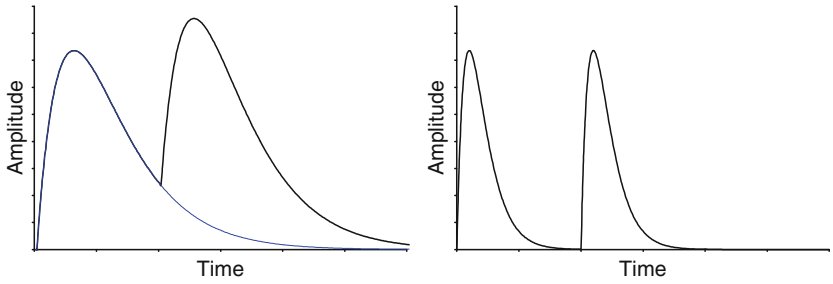


Fig. 3

Longer pulses will increase the probability of pileup, which will affect the peak amplitude of the second pulse. A smaller pulse width reduces pileup

2 Signal Processing

The upper half of *Fig. 4* shows the basic components of a detector readout. A preamplifier integrates the detector current to derive the signal charge and also transforms a short detector pulse into a long step output, which allows the following stages to have a small bandwidth to reduce the noise. The pulse shaper forms the signal into a pseudo-Gaussian shape whose peak amplitude is proportional to the signal charge. This peak amplitude is recorded by an analog-to-digital converter, and the digitized signal is stored and used for further data analysis. The bottom section of *Fig. 4* digitizes the detector signal and then sends the digitized signal to a digital signal processor, which performs pulse shaping, but can also perform other functions, such as correcting the signal amplitude when pileup occurs. Digital signal processing is discussed in Sect. 8.

Other components of signal processing systems handle the data readout. In single-channel detector systems the digitized output is sent directly to the memory. This becomes more complicated in large-scale detector systems where the data from thousands or millions of detector channels must be recorded. Monolithic integrated circuits (ICs) commonly include 128 readout channels and a detector module includes multiple ICs. *Figure 5* shows the barrel strip-detector module in the ATLAS SemiConductor Tracker (SCT) (ATLAS Collaboration 2008). The ICs are read out sequentially as shown in *Fig. 6*. For more details on large-scale semiconductor detector systems, see Spieler (2005).

To explain the basic functions of a signal processor, the simplest form is shown in *Fig. 7*. The first shaper function sets the pulse duration by sending the step input through a C–R "differentiator," which in the frequency domain acts as a high-pass filter. This reduces the pulse width, which increases the signal rate capability, but has just little effect on the noise bandwidth. To reduce the noise bandwidth, the signal is sent through an R–C "integrator," which is also a low-pass filter. This results in a pulse with a rounded peak, whose amplitude can be accurately measured by the subsequent digitizer. Pulse shapers can also produce outputs with a sharp peak, but since digitizers also have a limited bandwidth and need sufficient time to reach the peak amplitude, the accuracy of recording the peak amplitude can be uncertain.

The simple CR–RC shaper shows the basic functions of a pulse shaper, limiting the noise bandwidth and constraining the pulse duration. Although this simple shaper is 37% worse than the optimum shaper, it is quite useful, especially in circuits where size and power dissipation

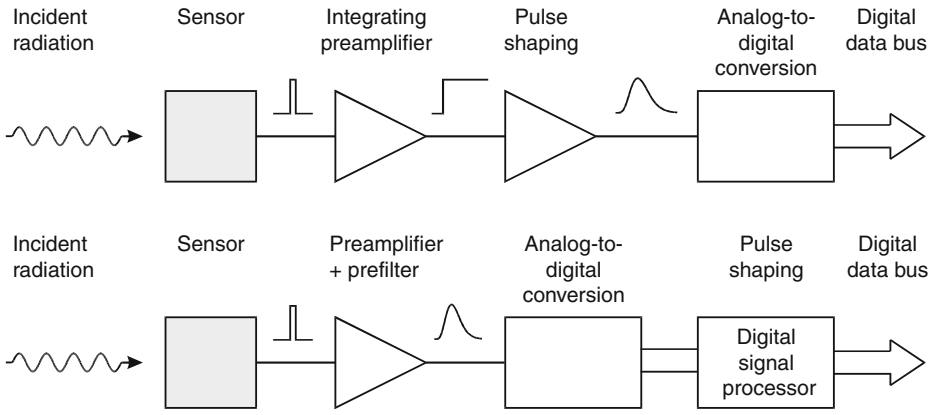


Fig. 4

A typical signal processing system consists of a preamplifier, here shown as a charge-sensitive amplifier, a pulse shaper, and a digitizer that converts the signal for subsequent signal storage. Pulse processing can be performed both with analog and digital circuitry. In the *bottom* system, the sensor signal is prefiltered in the preamplifier and then digitized. Pulse processing to optimize the signal-to-noise ratio and perform other functions is then performed in a digital signal processor

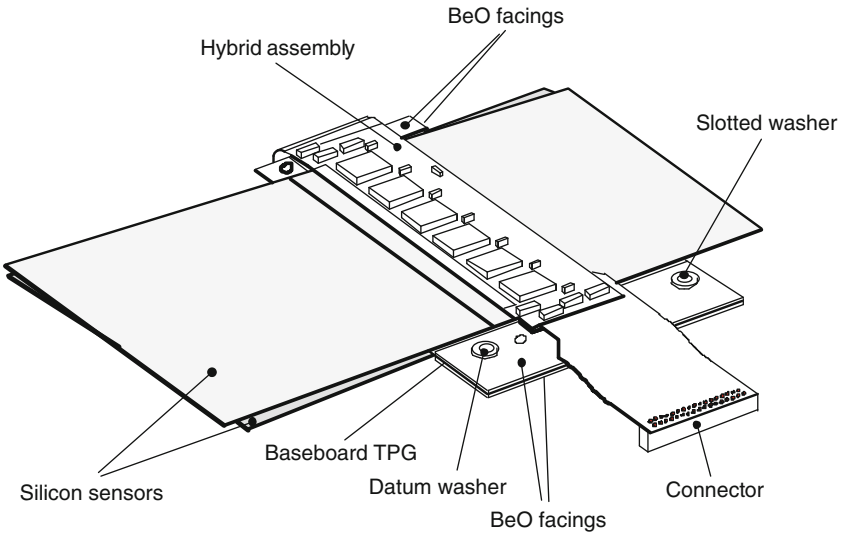


Fig. 5

ATLAS SCT barrel detector module. Two single-sided sensors are glued back-to-back with an intermediate thermal pyrolytic graphite (TPG) heat spreader. The "ear" extending from the module attaches to the support/cooling stave of the SCT barrel structure (Unno 2003) (figure courtesy of T. Kondo.)

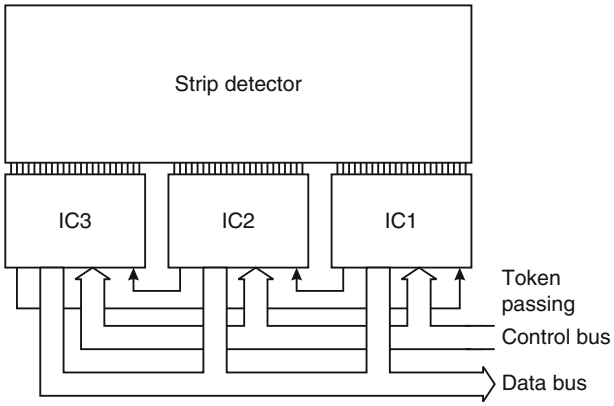


Fig. 6

Multiple ICs are ganged to read out a strip detector. The right-most chip IC1 is the master. A command on the control bus initiates the readout. When IC1 has written all of its data it passes the token to IC2. When IC2 has finished it passes the token to IC3, which in turn returns the token to the master IC1

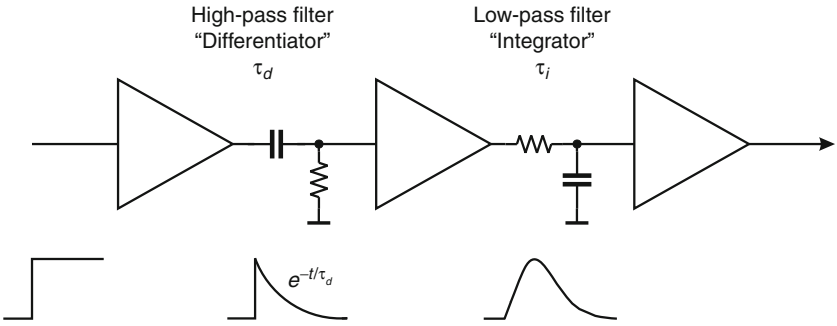


Fig. 7

A simple pulse shaper using a *CR* "differentiator" as a high-pass and an *RC* "integrator" as a low-pass filter. For a given decay time constant τ_d, optimum noise results when $\tau_i = \tau_d$

are important. However, by adding additional integrator stages, the output pulses can be made more symmetrical, as shown in *Fig. 8*, so for a given peaking time the pulse will return to the baseline more quickly and allow higher pulse rates. Typically several gain stages are needed to bring the signal level up to the requirements of the digitizer, so by tailoring the bandwidths of the amplifiers the equivalent of a two- to four-stage integrator can be achieved without additional circuit complexity. For higher integration levels more complex circuits are commonly used.

Figure 9 shows the frequency response of a *CR*–*RC* and *CR*–4*RC* shaper, both with a peaking time of 100 ns. The frequency spectrum simply scales inversely with peaking time. For the 100 ns peaking time shown in the figure, the peaking frequency of the *CR*–*RC* shaper is 1.6 MHz, and for a 10 ns peaking time it is 16 MHz. The bandwidth scales accordingly. For the *CR*–*RC* shaper, the peaking frequency is easily derived from the peaking time, which is equal

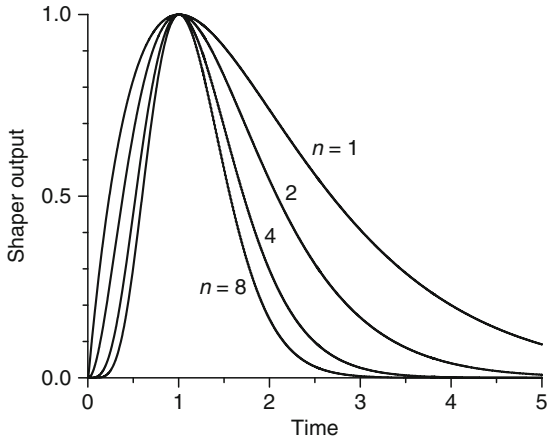


Fig. 8

Pulse shape *vs.* number of integrators in a *CR–nRC* shaper. The integration and differentiation time constants are scaled with the number of integrators ($\tau = \tau_{n=1}/n$) to maintain the peaking time

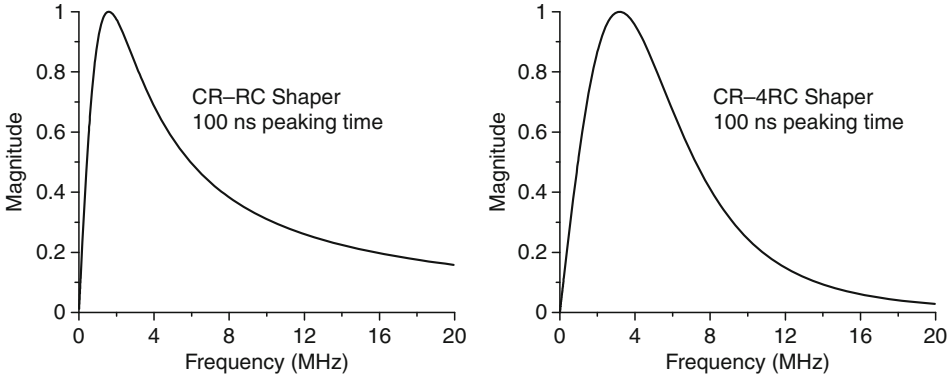


Fig. 9

The frequency response of *CR–RC* and *CR–4RC* pulse shapers, both with a peaking time of 100 ns. The peaking frequencies are 1.6 MHz for the *CR–RC* shaper and 3.2 MHz for the *CR–4RC*

to the integration and differentiation time constants τ. Then the peaking frequency is $\omega_P = 1/\tau$ or $f_P = (2\pi\tau)^{-1}$. However for the CR–$4RC$ shaper with 100 ns peaking time, the peaking frequency is 3.2 MHz, i.e. about twice as high as for the CR–RC shaper. The bandwidth, i.e. the difference between the upper and lower half-power frequencies is 3.2 MHz for the CR–RC shaper and 4.3 MHz for the CR–$4RC$ shaper.

For a given input signal, the signal level at the shaper output depends on the shaper type, assuming that the total gain of the amplifiers in the signal chain is the same. The shaper type also affects the noise bandwidth, so both the signal attenuation and the noise bandwidth must be assessed to optimize the signal-to-noise ratio.

3 Noise Analysis of a Detector–Preamplifier–Shaper System

To determine how the pulse shaper affects the signal-to-noise ratio, consider the detector front end in *Fig. 10*. The detector is represented by the capacitance C_d, a relevant model for many radiation sensors. Sensor bias voltage is applied through the resistor R_b. The bypass capacitor C_b shunts any external interference coming through the bias supply line to ground. For high-frequency signals, this capacitor appears as a low impedance, so for sensor signals the "far end" of the bias resistor is connected to ground. The coupling capacitor C_c blocks the sensor bias voltage from the amplifier input, which is why a capacitor serving this role is also called a "blocking capacitor." The series resistor R_s represents any resistance present in the connection from the sensor to the amplifier input. This includes the resistance of the sensor electrodes, the resistance of the connecting wires or traces, any resistance used to protect the amplifier against large voltage transients ("input protection"), and parasitic resistances in the input transistor.

The following implicitly includes a constraint on the bias resistance, whose role is often misunderstood. It is often thought that the signal current generated in the sensor flows through R_b and the resulting voltage drop is measured. If the time constant $R_b C_d$ is small compared to the peaking time of the shaper T_P, the sensor will have discharged through R_b and much of the signal will be lost. Thus, we have the condition $R_b C_d \gg T_P$, or $R_b \gg T_P/C_d$. The bias resistor must be sufficiently large to block the flow of signal charge, so that all of the signal is available for the amplifier.

To analyze this circuit, a voltage amplifier will be assumed, so all noise contributions will be calculated as a noise voltage appearing at the amplifier input. Steps in the analysis are: (1) Determine the frequency distributions of all noise voltages presented to the amplifier input from all individual noise sources. (2) Integrate over the frequency response of the shaper (for simplicity a CR–RC shaper) and determine the total noise voltage at the shaper output. (3) Determine the peak amplitude of the output pulse for a known input signal charge. The equivalent noise charge (ENC) is the signal charge for which $S/N = 1$.

The equivalent circuit for the noise analysis (second panel of *Fig. 10*) includes both current and voltage noise sources. The "shot noise" i_{nd} of the sensor leakage current is represented by a current noise generator in parallel with the sensor capacitance. As noted in Chap. 2, "Electronics Part I," resistors can be modeled either as a voltage or current generator. Generally, resistors shunting the input act as noise current sources, and resistors in series with the input act as noise voltage sources (which is why some in the detector community refer to current

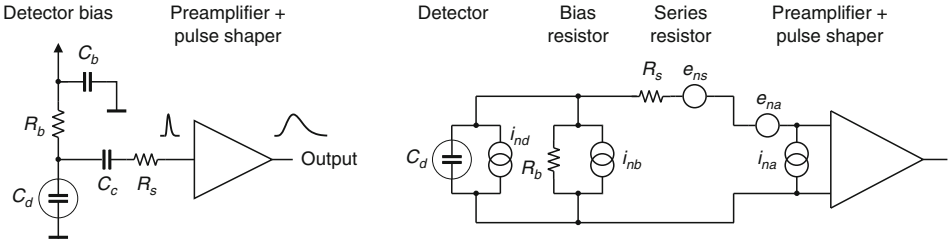


Fig. 10
A detector front-end circuit and its equivalent circuit for noise calculations

and voltage noise as "parallel" and "series" noise). Since the bias resistor effectively shunts the input, as the capacitor C_b passes current fluctuations to ground, it acts as a current generator i_{nb}, and its noise current has the same effect as the shot-noise current from the detector. The shunt resistor can also be modeled as a noise voltage source, yielding the result that it acts as a current source. Choosing the appropriate model merely simplifies the calculation. Any other shunt resistances can be incorporated in the same way. Conversely, the series resistor R_s acts as a voltage generator. The electronic noise of the amplifier is described fully by a combination of voltage and current sources at its input, shown as e_{na} and i_{na}.

Thus, the noise sources are

$$\begin{aligned} \text{Sensor bias current: } i_{nd}^2 &= 2eI_d, \\ \text{Shunt resistance: } i_{nb}^2 &= \frac{4kT}{R_b}, \\ \text{Series resistance: } e_{ns}^2 &= 4kTR_s, \\ \text{Amplifier: } & \quad e_{na},\ i_{na}, \end{aligned}$$

where e is the electronic charge, I_d the sensor bias current, k the Boltzmann constant, and T the temperature. Typical amplifier noise parameters e_{na} and i_{na} are of order $\mathrm{nV}/\sqrt{\mathrm{Hz}}$ and $\mathrm{fA}/\sqrt{\mathrm{Hz}}$ (FETs) to $\mathrm{pA}/\sqrt{\mathrm{Hz}}$ (bipolar transistors). Amplifiers tend to exhibit a "white" noise spectrum at high frequencies (greater than order kHz), but at low frequencies show excess noise components with the spectral density

$$e_{nf}^2 = \frac{A_f}{f}, \tag{1}$$

where the noise coefficient A_f is device specific and of order 10^{-10}–$10^{-12}\ \mathrm{V}^2$.

The noise voltage generators are in series and simply add in quadrature. White noise distributions remain white. However, a portion of the noise currents flows through the detector capacitance, resulting in a frequency-dependent noise voltage $i_n/(\omega C_d)$, so the originally white spectrum of the sensor shot noise and the bias resistor now acquires a $1/f$ dependence and their contribution increases as the shaper is shifted to lower frequencies, i.e. longer shaping times. The frequency distribution of all noise sources is further altered by the combined frequency response of the amplifier–shaper chain $A(f)$. Integrating over the cumulative noise spectrum at the amplifier–shaper output and comparing to the output voltage for a known input signal yields the signal-to-noise ratio. In this example, the shaper is a simple CR–RC shaper, where for a given differentiation time constant the signal-to-noise ratio is maximized when the integration time constant equals the differentiation time constant, $\tau_i = \tau_d \equiv \tau$. Then the output pulse assumes its maximum amplitude at the time $T_P = \tau$.

Although the basic noise sources are currents or voltages, since radiation detectors are typically used to measure charge, the system's noise level is conveniently expressed as an equivalent noise charge Q_n. As noted previously, this is equal to the detector signal that yields a signal-to-noise ratio of one. The equivalent noise charge is commonly expressed in coulombs, the corresponding number of electrons, or the equivalent deposited energy (eV). For the above circuit the equivalent noise charge is

$$Q_n^2 = \left(\frac{\mathrm{e}^2}{8}\right)\left[\left(2eI_d + \frac{4kT}{R_b} + i_{na}^2\right)\cdot\tau + \left(4kTR_s + e_{na}^2\right)\cdot\frac{C_d^2}{\tau} + 4A_fC_d^2\right]. \tag{2}$$

The prefactor $\mathrm{e}^2/8 = \exp(2)/8 = 0.924$ normalizes the noise to the signal gain. The first term combines all noise current sources and increases with shaping time. The second term combines

all noise voltage sources and decreases with shaping time, but increases with sensor capacitance. The third term is the contribution of amplifier $1/f$ noise and, as a voltage source, also increases with sensor capacitance. The $1/f$ term is independent of shaping time, since for a $1/f$ spectrum the total noise depends on the ratio of upper to lower cutoff frequency, which depends only on shaper topology, but not on the shaping time.

The equivalent noise charge can be expressed in a more general form that applies to all types of pulse shapers:

$$Q_n^2 = i_n^2 F_i T_S + e_n^2 F_v \frac{C^2}{T_S} + F_{vf} A_f C^2, \tag{3}$$

where F_i, F_v, and F_{vf} depend on the shape of the pulse determined by the shaper and T_S is a characteristic time, e.g. the peaking time of a CR–nRC-shaped pulse or the prefilter time constant in a correlated double sampler. C is the total parallel capacitance at the input, including the amplifier input capacitance. The shape factors F_i, F_v are easily calculated,

$$F_i = \frac{1}{2T_S} \int_{-\infty}^{\infty} [W(t)]^2 \, \mathrm{d}t \quad \text{and} \quad F_v = \frac{T_S}{2} \int_{-\infty}^{\infty} \left[\frac{\mathrm{d}W(t)}{\mathrm{d}t} \right]^2 \mathrm{d}t. \tag{4}$$

For time-invariant pulse shaping, $W(t)$ is simply the system's impulse response (the output signal seen on an oscilloscope) with the peak output signal normalized to unity. For a time-variant shaper, the same equations apply, but $W(t)$ is determined differently. See references Goulding (1972, 1982) and Radeka (1972, 1974) for more details.

A CR–RC shaper with equal time constants $\tau_i = \tau_d$ has $F_i = F_v = 0.9$ and $F_{vf} = 4$, independent of the shaping time constant, so for the circuit in ❯ *Fig. 10* ❯ Eq. 3 becomes

$$Q_n^2 = \left(2q_e I_d + \frac{4kT}{R_b} + i_{na}^2 \right) F_i T_S + \left(4kTR_s + e_{na}^2 \right) F_v \frac{C^2}{T_S} + F_{vf} A_f C^2. \tag{5}$$

Pulse shapers can be designed to reduce the effect of current noise, to mitigate radiation damage, for example. Increasing pulse symmetry tends to decrease F_i and increase F_v, e.g. to $F_i = 0.45$ and $F_v = 1.0$ for a shaper with one CR differentiator and four cascaded RC integrators.

❯ *Figure 11* shows how equivalent noise charge is affected by shaping time. At short shaping times the voltage noise dominates, whereas at long shaping times the current noise takes over. Minimum noise is obtained where the current and voltage contributions are equal. The noise minimum is flattened by the presence of $1/f$ noise. Also shown is that increasing the detector capacitance will increase the voltage noise contribution and shift the noise minimum to longer shaping times, albeit with an increase in minimum noise.

For quick estimates one can use the following equation, which assumes a field effect transistor (FET) amplifier (negligible i_{na}) and a simple CR–RC shaper with peaking time τ. The noise is expressed in units of the electronic charge e, and C is the total parallel capacitance at the input, including C_d, all stray capacitances, and the amplifier's input capacitance,

$$Q_n^2 = 12 \left[\frac{e^2}{\mathrm{nA\,ns}} \right] I_d \tau + 6 \cdot 10^5 \left[\frac{e^2\,\mathrm{k\Omega}}{\mathrm{ns}} \right] \frac{\tau}{R_b} + 3.6 \cdot 10^4 \left[\frac{e^2\,\mathrm{ns}}{(\mathrm{pF})^2 (\mathrm{nV})^2/\mathrm{Hz}} \right] e_n^2 \frac{C^2}{\tau}. \tag{6}$$

The noise charge is improved by reducing the detector capacitance and leakage current, judiciously selecting all resistances in the input circuit, and choosing the optimum shaping time constant. The noise parameters of a well-designed amplifier depend primarily on the input device. Fast, high-gain transistors are generally best.

In field effect transistors, both junction field effect transistors (JFETs) or metal oxide semiconductor field effect transistors (MOSFETs), the noise current contribution is very small,

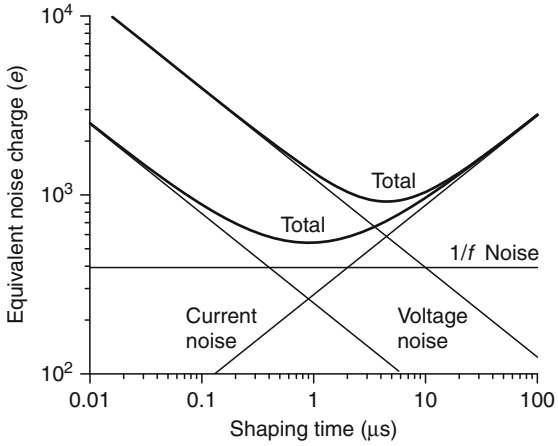


Fig. 11

Equivalent noise charge vs. shaping time. At small shaping times (large bandwidth), the equivalent noise charge is dominated by voltage noise, whereas at long shaping times (large integration times), the current noise contributions dominate. The total noise assumes a minimum where the current and voltage contributions are equal. The "1/*f*" noise contribution is independent of shaping time and flattens the noise minimum. Increasing the voltage or current noise contribution shifts the noise minimum. Increased voltage noise is shown as an example

so reducing the detector leakage current and increasing the bias resistance will allow long shaping times with correspondingly lower noise. The equivalent input noise voltage is $e_n^2 \approx 4kT/g_m$, where g_m is the transconductance, which increases with operating current. For a given current, the transconductance increases when the channel length is reduced, so reductions in feature size with new process technologies are beneficial. At a given channel length, minimum noise is obtained when a device is operated at maximum transconductance. If lower noise is required, the width of the device can be increased (equivalent to connecting multiple devices in parallel). This increases the transconductance (and required current) with a corresponding decrease in noise voltage, but also increases the input capacitance. At some point the reduction in noise voltage is outweighed by the increase in total input capacitance. The optimum is obtained when the FET's input capacitance equals the external capacitance (sensor + stray capacitance). Note that this capacitive matching criterion only applies when the input-current noise contribution of the amplifying device is negligible.

Capacitive matching comes at the expense of power dissipation. Since the minimum is shallow, one can operate at significantly lower currents and reduced input capacitance with just a minor increase in noise. In large detector arrays power dissipation is critical, so FETs are hardly ever operated at their minimum noise. Instead, one seeks an acceptable compromise between noise and power dissipation (see Spieler 2005 for a detailed discussion). Similarly, the choice of input devices is frequently driven by available fabrication processes. High-density integrated circuits tend to include only MOSFETs, so this determines the input device, even where a bipolar transistor would provide better performance.

In bipolar transistors the shot noise associated with the base current I_B is significant, $i_{nB}^2 = 2eI_B$. Since $I_B = I_C/\beta_{DC}$, where I_C is the collector current and β_{DC} the direct current gain, this contribution increases with device current. On the other hand, the equivalent input noise voltage

$$e_n^2 = \frac{2(kT)^2}{eI_C} \tag{7}$$

decreases with collector current, so the noise assumes a minimum at a specific collector current,

$$Q_{n,\min}^2 = 4kT\frac{C}{\sqrt{\beta_{DC}}}\sqrt{F_i F_v} \quad \text{at} \quad I_C = \frac{kT}{e}C\sqrt{\beta_{DC}}\sqrt{\frac{F_v}{F_i}}\frac{1}{T_S}. \tag{8}$$

For a CR–RC shaper and $\beta_{DC} = 100$,

$$Q_{n,\min} \approx 250\left[\frac{e}{\sqrt{\text{pF}}}\right]\cdot\sqrt{C} \quad \text{at} \quad I_C = 260\left[\frac{\mu\text{A ns}}{\text{pF}}\right]\cdot\frac{C}{T_S}. \tag{9}$$

The minimum obtainable noise is independent of shaping time (unlike FETs), but only at the optimum collector current I_C, which does depend on shaping time.

In bipolar transistors, the input capacitance is usually much smaller than the sensor capacitance (of order 1 pF for $e_n \approx 1\,\text{nV}/\sqrt{\text{Hz}}$) and substantially smaller than in FETs with comparable noise. Since the transistor input capacitance enters into the total input capacitance, this is an advantage. Note that capacitive matching does not apply to bipolar transistors because their noise current contribution is significant. Due to the base current noise, bipolar transistors are best at short shaping times, where they also require lower power than FETs for a given noise level.

When the input noise current is negligible, the noise increases linearly with the total capacitance at the input. If the detector capacitance dominates, the noise slope

$$\frac{dQ_n}{dC_d} \approx 2e_n \cdot \sqrt{\frac{F_v}{T_S}} \tag{10}$$

depends both on the preamplifier (e_n) and the shaper (F_v, T_S). The zero intercept can be used to determine the amplifier input capacitance plus any additional capacitance at the input node. Note that the noise slope also depends on the pulse shaping, so this should be included in the specification (and often isn't).

Practical noise levels range from $<1\,e$ for charge-coupled devices (CCDs) at long shaping times to $\sim 10^4\,e$ in high-capacitance liquid-Ar calorimeters. Silicon strip detectors typically operate at $\sim 10^3$ electrons, whereas pixel detectors with fast readout provide noise of 100–200 electrons. Transistor noise is discussed in more detail in Spieler (2005).

4 Timing Measurements

Pulse-height measurements discussed up to now emphasize measurement of signal charge. Timing measurements seek to optimize the determination of the time of occurrence. Although, as in amplitude measurements, signal-to-noise ratio is important, the determining parameter is not signal-to-noise, but slope-to-noise ratio. This is illustrated in *Fig. 12*, which shows the leading edge of a pulse fed into a threshold discriminator (comparator), a "leading-edge trigger."

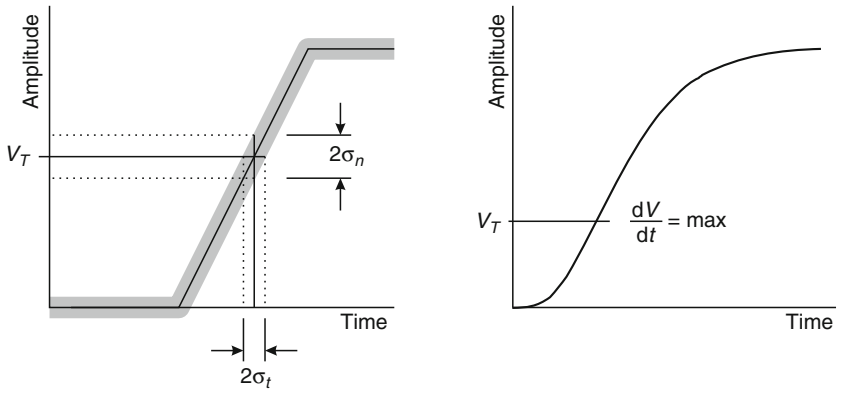


Fig. 12

Fluctuations in signal amplitude crossing a threshold translate into timing fluctuations (*left*). With realistic pulses the slope changes with amplitude, so minimum timing jitter occurs with the trigger level at the maximum slope

The instantaneous signal level is modulated by noise, where the variations are indicated by the shaded band. Because of these fluctuations, the time of threshold crossing fluctuates. By simple geometrical projection, the timing variance or "jitter" is

$$\sigma_t = \frac{\sigma_n}{(\mathrm{d}V/\mathrm{d}t)_{V_T}} \approx \frac{t_r}{S/N}, \tag{11}$$

where σ_n is the rms noise and the derivative of the signal $\mathrm{d}V/\mathrm{d}t$ is evaluated at the trigger level V_T. Although this example shows a voltage pulse, it applies to any other signal. To increase $\mathrm{d}V/\mathrm{d}t$ without incurring excessive noise, the amplifier bandwidth should match the rise time of the detector signal. The 10–90% rise time of an amplifier with the upper cutoff frequency f_u is

$$t_r = 2.2\,\tau = \frac{2.2}{2\pi f_u} = \frac{0.35}{f_u}. \tag{12}$$

For example, an oscilloscope with 350 MHz bandwidth has a 1 ns rise time. When amplifiers are cascaded, which is invariably necessary, the individual rise times add in quadrature:

$$t_r \approx \sqrt{t_{r1}^2 + t_{r2}^2 + \ldots + t_{rn}^2}. \tag{13}$$

Increasing signal-to-noise ratio improves time resolution, so minimising the total capacitance at the input is also important. At high signal-to-noise ratios, the time jitter can be much smaller than the rise time.

The second contribution to time resolution is time walk, where the timing signal shifts with amplitude as shown in *Fig. 13*. This can be corrected by various means, either in hardware or software. For more detailed tutorials on timing measurements, see Spieler (1982, 2005).

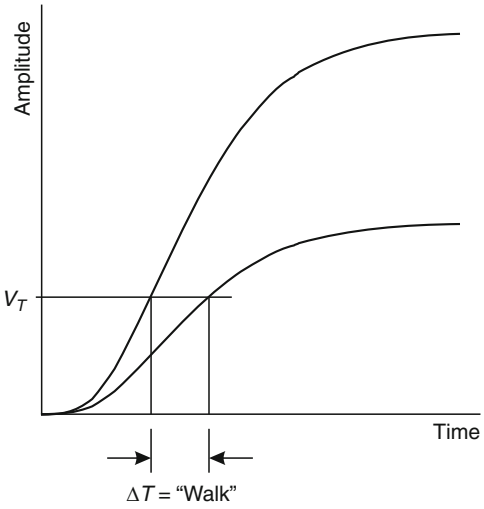


Fig. 13
The time at which a signal crosses a fixed threshold depends on the signal amplitude, leading to "time walk"

5 Digital Electronics

Analog signals utilize continuously variable properties of the pulse to impart information, such as the pulse amplitude or pulse shape. Digital signals have constant amplitude, but the presence of the signal at specific times is evaluated, i.e. whether the signal is in one of two states, "low" or "high." However, this still involves an analog process, as the presence of a signal is determined by the signal level exceeding a threshold at the proper time.

5.1 Logic Elements

Figure 14 illustrates several functions utilized in digital circuits ("logic" functions). An AND gate provides an output only when all inputs are high. An OR gives an output when any input is high. An eXclusive OR (XOR) responds when only one input is high. The same elements are commonly implemented with inverted outputs, then called NAND and NOR gates, for example. The D flip-flop is a bistable memory circuit that records the presence of a signal at the data input D when a signal transition occurs at the clock input CLK. This device is commonly called a latch. Inverted inputs and outputs are denoted by small circles or by superimposed bars, e.g. $\overline{Q}$ is the inverted output of a flip-flop, as shown in *Fig. 15*.

Logic circuits are fundamentally amplifiers, so they also suffer from bandwidth limitations. The pulse train of the AND gate in *Fig. 14* illustrates a common problem. The third pulse of input B is going low at the same time that input A is going high. Depending on the time overlap, this can yield a narrow output that may or may not be recognised by the following circuit. In an XOR this can occur when two pulses arrive nearly at the same time. The D flip-flop requires a minimum setup time for a level change at the D input to be recognised, so changes in the

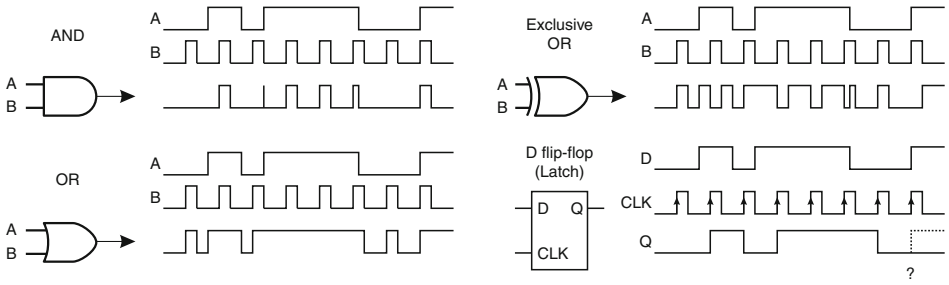


Fig. 14

Basic logic functions include gates (AND, OR, Exclusive OR) and flip-flops. The outputs of the AND and D flip-flop show how small shifts in relative timing between inputs can determine the output state

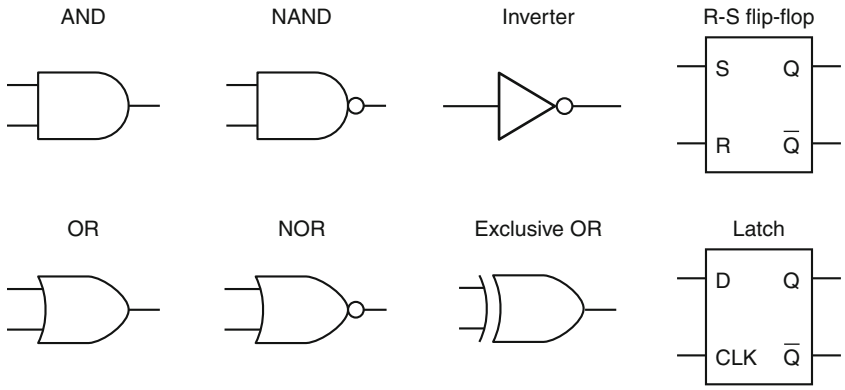


Fig. 15

Some common logic symbols. Inverted outputs are denoted by small circles or by a superimposed bar, as for the latch output $\overline{Q}$. Additional inputs can be added to gates as needed. An R–S flip-flop sets the Q output high in response to an S input. An R input resets the Q output to low

data level may not be recognised at the correct time. These marginal events may be extremely rare and perhaps go unnoticed. However, in complex systems the combination of "glitches" can make the system "hang up," necessitating a system reset. Data transmission protocols have been developed to detect such errors (parity checks, Hamming codes, etc.), so corrupted data can be rejected.

Some key aspects of logic systems can be understood by inspecting the circuit elements that are used to form logic functions. *Figure 16* shows simple inverter circuits using MOS transistors. In this context it is sufficient to know that in an NMOS transistor a conductive channel is formed when the input electrode is biased positive with respect to the channel. The input, called the "Gate" (G), is capacitively coupled to the output channel connected between the "Drain" (D) and "Source" (S) electrodes. In the NMOS inverter applying a positive voltage to the gate makes the output channel conduct, so the output level is low. A PMOS transistor

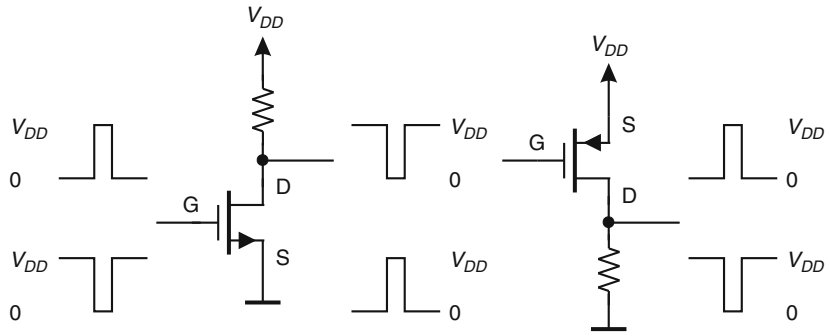


Fig. 16

In an NMOS inverter the transistor conducts when the input is high (*left*), whereas in a PMOS inverter the transistor conducts when the input is low (*right*). In both circuits the input pulse is inverted, whether the input swings high or low

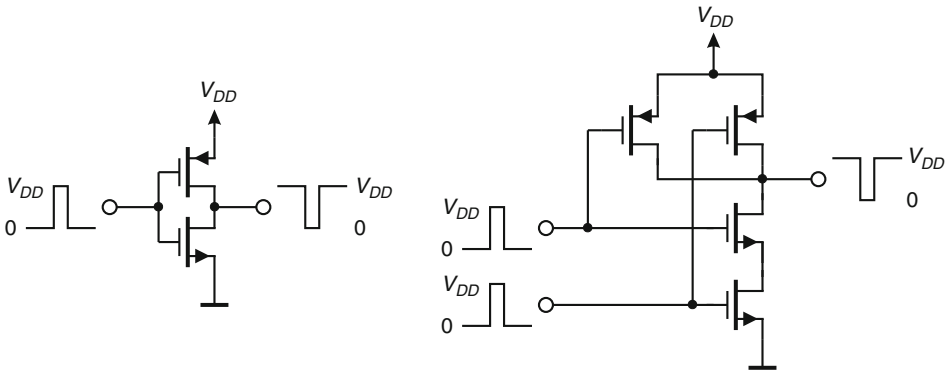


Fig. 17

A CMOS inverter (*left*) and NAND gate (*right*)

is the complementary device, where a conductive channel is formed when the gate is biased negative with respect to the source. Since the source is at positive potential, a low level at the inverter input yields a high level at the output. Regardless of the device and pulse polarity, the output pulse is always the inverse of the input.

NMOS and PMOS inverters draw current when in their "active" state. Combining NMOS and PMOS transistors in a complementary MOS (CMOS) circuit allows zero current draw in both the high and low states with a substantial reduction in power consumption. A CMOS inverter is shown in *Fig. 17*, which also shows how devices are combined to form a CMOS NAND gate. In the inverter, the lower (NMOS) transistor is turned off when the input is low, but the upper (PMOS) transistor is turned on, so the output is connected to V_{DD}, taking the output high. Since the current path from V_{DD} to ground is blocked by either the NMOS or PMOS device being off, the power dissipation is zero in both the high and low states. Current only flows during the level transition when both devices are on as the input level is at approximately $V_{DD}/2$. As a result, the power dissipation of CMOS logic is significantly less than in NMOS or

PMOS circuitry. However, the reduction in power only obtains in logic circuitry. CMOS analog amplifiers are not fundamentally more power efficient than NMOS or PMOS circuits, although CMOS allows more efficient circuit topologies.

5.2 Propagation Delays and Power Dissipation

Logic elements always operate in conjunction with other circuits, as illustrated in *Fig. 18*. The wiring resistance together with the total load capacitance increases the rise time of the logic pulse and as a result delays the time when the transition crosses the logic threshold. The energy dissipated in the wiring resistance R is

$$E = \int i^2(t) R \, \mathrm{d}t. \tag{14}$$

The current flow during one transition is

$$i(t) = \frac{V}{R} \exp\left(-\frac{t}{RC}\right), \tag{15}$$

so the dissipated energy per transition (either positive or negative)

$$E = \frac{V^2}{R} \int_0^\infty \exp\left(-\frac{2t}{RC}\right) \mathrm{d}t = \frac{1}{2} C V^2. \tag{16}$$

When pulses occur at a frequency f, the total power dissipated in both the positive and negative transitions is

$$P = f C V^2. \tag{17}$$

Thus, the power dissipation increases with clock frequency and the square of the logic swing.

The above expression is often derived from the energy stored in a capacitor. However, this energy will be returned when the capacitor is discharged, so after the leading and trailing edges of a pulse the net energy is zero. This is one of many examples where the wrong physics yields a correct result – until one digs deeper.

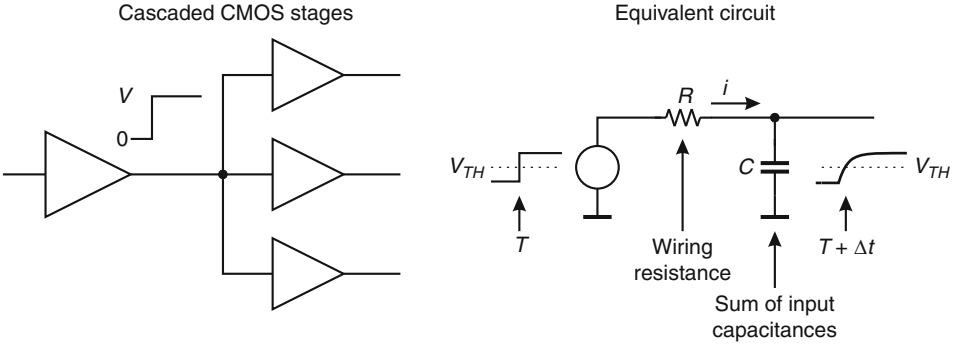


Fig. 18
The wiring resistance together with the distributed load capacitance delays the signal

Fast logic is time-critical. It relies on logic operations from multiple paths coming together at the right time. Valid results depend on maintaining minimum allowable overlaps and set-up times as illustrated in *Fig. 14*. Each logic circuit has a finite propagation delay, which depends on circuit loading, i.e. how many loads the circuit has to drive. In addition, as illustrated in *Fig. 18*, the wiring resistance and capacitive loads introduce delay. This depends on the number of circuits connected to a wire or trace, the length of the trace, and the dielectric constant of the substrate material. Relying on control of circuit and wiring delays to maintain timing requires great care, as it depends on circuit variations and temperature. In principle all of this can be simulated, but in complex systems there are too many combinations to test every one. A more robust solution is to use synchronous systems, where the timing of all transitions is determined by a master clock. Generally, this does not provide the utmost speed and requires some additional circuitry, but increases reliability. Nevertheless, clever designers frequently utilize asynchronous logic. Sometimes it succeeds … and sometimes it does not.

5.3 Logic Arrays

Commodity integrated circuits with basic logic blocks are readily available, e.g. with four NAND gates or two flip-flops in one package. These can be combined to form simple digital systems. However, complex logic systems are no longer designed using individual gates. Instead, logic functions are described in a high-level language (e.g. VHDL – VHSIC Hardware Description Language, VHSIC – Very High Speed Integrated Circuit), synthesized using design libraries, and implemented as custom ICs – application-specific ICs (ASICs) – or programmable logic arrays. In these implementations the digital circuitry no longer appears as an ensemble of inverters, gates, and flip-flops, but as an integrated logic block that provides specific outputs in response to various input combinations. This is illustrated in *Fig. 19*. Field Programmable Gate or logic Arrays (FPGAs) are a common example. A representative FPGA has 512 pads usable for inputs and outputs, ~10^6 gates, and ~100 K of memory. Modern design tools also account for propagation delays, wiring lengths, loads, and temperature dependence. The design software also generates "test vectors" that can be used to test finished parts. Properly implemented, complex digital designs can succeed on the first pass, whether as ASICs or as logic or gate arrays.

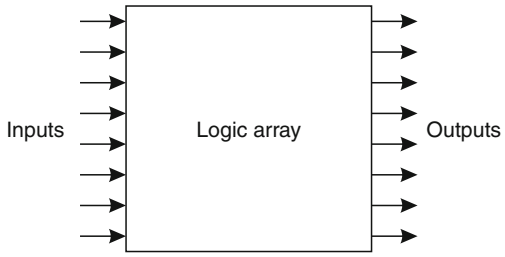


Fig. 19

Complex logic circuits are commonly implemented using logic arrays that as an integrated block provide the desired outputs in response to specific input combinations

6 Analog-to-Digital Conversion

For data storage and subsequent analysis, the analog signal at the shaper output must be digitized. Important parameters for analog-to-digital converters (ADCs or A/Ds) used in detector systems are:

1. Resolution: The "granularity" of the digitized output.
2. Differential nonlinearity: How uniform are the digitization increments?
3. Integral nonlinearity: Is the digital output proportional to the analog input?
4. Conversion time: How much time is required to convert an analog signal to a digital output?
5. Count-rate performance: How quickly can a new conversion commence after completion of a prior one without introducing deleterious artifacts?
6. Stability: Do the conversion parameters change with time?

Instrumentation ADCs used in industrial data acquisition and control systems share most of these requirements. However, detector systems place greater emphasis on differential nonlinearity and count-rate performance. The latter is important, as detector signals often occur randomly, in contrast to systems where signals are sampled at regular intervals. As in amplifiers, if the DC gain is not precisely equal to the high-frequency gain, the baseline will shift. Furthermore, following each pulse, it takes some time for the baseline to return to its quiescent level. For periodic signals of roughly equal amplitude, these baseline deviations will be the same for each pulse, but for a random sequence of pulses with varying amplitudes, the instantaneous baseline level will be different for each pulse and broaden the measured signal.

Another common characteristic is that the actual resolution of a digitizer is worse than expected based on the digital resolution. This can have several main causes, e.g. electronic noise, differential nonlinearity, and digital cross-talk. It can be checked easily by viewing the spectrum of a precision pulser whose amplitude is centered in one digitization channel, so only a single bit combination should result. ❯ *Figure 20* shows results of a 13-bit ADC where the pulser

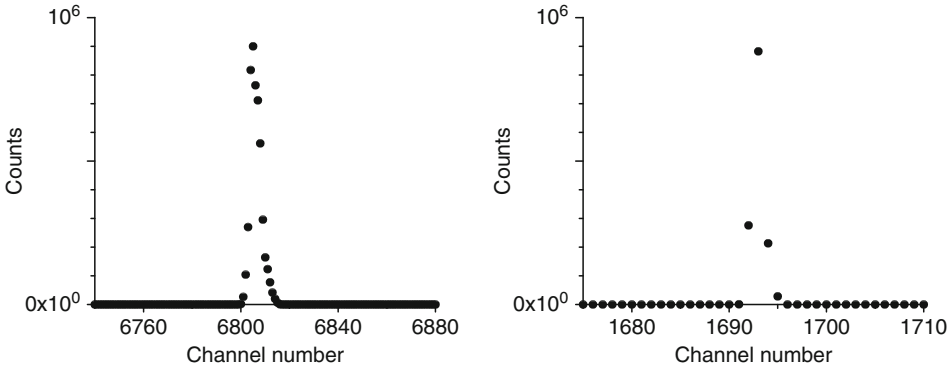


Fig. 20

Spectrum of a precision pulser centered within an ADC channel. The maximum number of counts per channel is about 10^6. In the 13-bit range (*left*), the signal is distributed over many channels. In the 11-bit range the spectrum is matches the digital resolution. Although this ADC can provide 13 bits of digital resolution, its analog resolution is only 10–11 bits

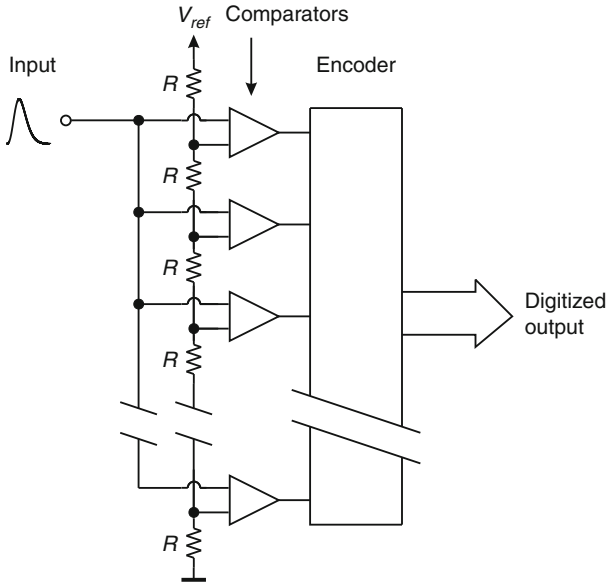


Fig. 21
Block diagram of a flash ADC

signal is distributed over >12 channels, whereas in the 11-bit range the digital resolution matches the analog resolution. Although this ADC can provide 13 bits of digital resolution, its analog resolution is only 10–11 bits, so the 12th and 13th bits are superfluous.

Conceptually, the simplest technique is flash conversion, illustrated in *Fig. 21*. The signal is fed in parallel to a bank of threshold comparators. The individual threshold levels are set by a resistive divider. The comparator outputs are encoded such that the output of the highest-level comparator that fires yields the correct bit pattern. The threshold levels can be set to provide a linear conversion characteristic where each bit corresponds to the same analog increment, or a nonlinear characteristic to provide increments proportional to the absolute level, which provides constant relative resolution over the range, for example.

The big advantage of this scheme is speed; conversion proceeds in one step, and conversion times <10 ns are readily achievable. The drawbacks are component count and power consumption, as one comparator is required per conversion bin. For example, an 8-bit converter requires 256 comparators. The conversion is always monotonic and differential nonlinearity is determined by the matching of the resistors in the threshold divider. Only relative matching is required, so this topology is a good match for monolithic integrated circuits. Flash ADCs are available with conversion rates >500 MS/s (megasamples per second) at 8-bit resolution and a power dissipation of about 5 W.

The most commonly used technique is the successive-approximation ADC, shown in *Fig. 22*. The input pulse is sent to a pulse stretcher, which follows the signal until it reaches its cusp and then holds the peak value. The stretcher output feeds a comparator, whose reference is provided by a digital-to-analog converter (DAC). The DAC is cycled beginning with the most significant bits. The corresponding bit is set when the comparator fires, i.e. the DAC output becomes less than the pulse height. Then the DAC cycles through the less significant bits,

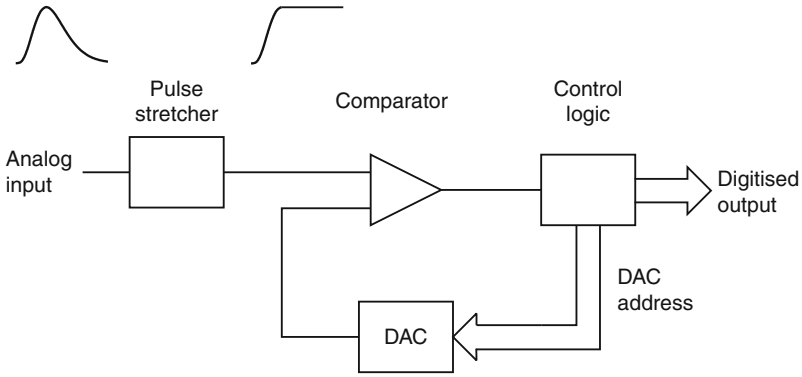


Fig. 22

Principle of a successive-approximation ADC. The DAC is controlled to sequentially add levels proportional to 2^n, 2^{n-1}, …, 2^0. The corresponding bit is set if the comparator output is high (DAC output < pulse height)

always setting the corresponding bit when the comparator fires. Thus, n-bit resolution requires n steps and yields 2^n bins. This technique makes efficient use of circuitry and is fairly fast. High-resolution devices (16–20 bits) with conversion times of order µs are readily available. Currently a 16-bit ADC with a conversion time of 1 µs (1 MS/s) requires about 100 mW.

A common limitation is differential nonlinearity (DNL) since the resistors that set the DAC levels must be extremely accurate. For DNL <1%, the resistor determining the 2^{12}-level in a 13-bit ADC must be accurate to $<2.4 \cdot 10^{-6}$. As a consequence, differential nonlinearity in high-resolution successive-approximation converters is typically 10–20% and often exceeds the 0.5 LSB (least significant bit) required to ensure monotonic response. DNL can be reduced by having multiple pulses of a given amplitude converted to a range of output bit combinations, so the deviations between channels average out. This is done by offsetting the DAC event by event and then correcting the digitized output accordingly.

The Wilkinson ADC (Wilkinson 1950) has traditionally been the mainstay of precision pulse digitization. The principle is shown in ❯ *Fig. 23*. The peak signal amplitude is acquired by a combined peak detector/pulse stretcher and transferred to a memory capacitor. The output of the peak detector initiates the conversion process:

1. The memory capacitor is disconnected from the stretcher.
2. A current source is switched on to linearly discharge the capacitor with current I_R.
3. And simultaneously, a counter is enabled to determine the number of clock pulses until the voltage on the capacitor reaches the baseline level V_{BL}.

The time required to discharge the capacitor is a linear function of pulse height, so the counter content provides the digitized pulse height. The clock pulses are provided by a crystal oscillator, so the time between pulses is extremely uniform and this circuit inherently provides excellent differential linearity. The drawback is the relatively long conversion time T_C, which is proportional to the pulse height, $T_C = n \cdot T_{clk}$, where the channel number n corresponds to the pulse height. For example, a clock frequency of 100 MHz provides a clock period $T_{clk} = 10$ ns and a maximum conversion time $T_C = 82$ µs for 13 bits ($n = 8192$). Clock frequencies of 100 MHz are

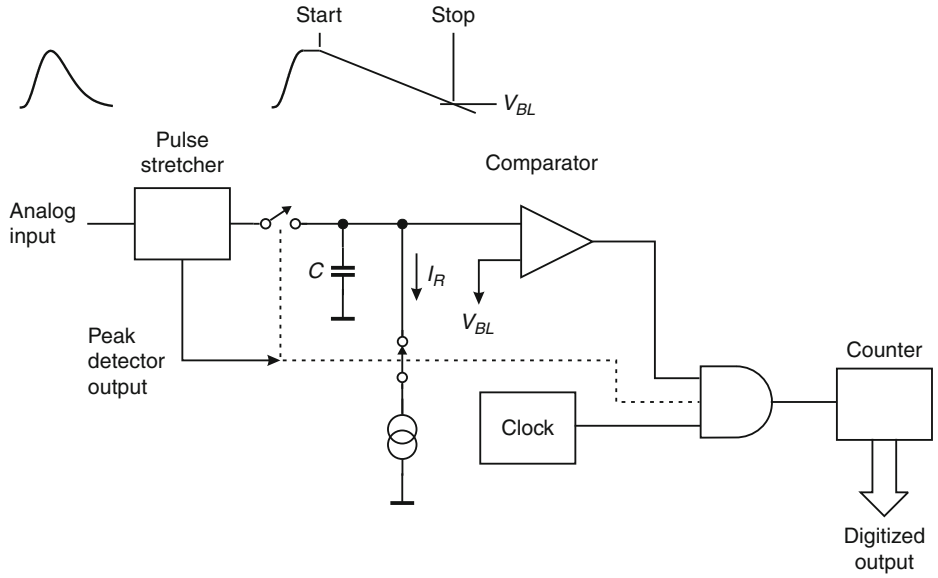


Fig. 23

Principle of a Wilkinson ADC. After the peak amplitude has been acquired, the output of the peak detector initiates the conversion process. The memory capacitor is discharged by a constant current while counting the clock pulses. When the capacitor is discharged to the baseline level V_{BL}, the comparator output goes low and the conversion is complete

typical, but >400 MHz have been implemented with excellent performance (DNL < 10^{-3}). This scheme makes efficient use of circuitry and allows low power dissipation. Wilkinson ADCs have been implemented in 128-channel readout ICs for silicon strip detectors (Garcia-Sciveres 1999). Each ADC added only 100 µm to the length of a channel and a power of 0.3 mW per readout channel.

Systems that utilize digital signal processing (*Fig. 4* and Sect. 8) require ADCs that are much faster than imposed by the event rate. Flash ADCs are one possibility, but another that requires less power is the pipelined ADC (see Spieler 2005, pp 208–209).

7 Time-to-Digital Converters (TDCs)

The combination of a clock generator with a counter is the simplest technique for time-to-digital conversion, as shown in *Fig. 24*. The clock pulses are counted between the start and stop signals, which yields a direct readout in real time. The limitation is the speed of the counter, which in current technology is limited to about 1 GHz, yielding a time resolution of 1 ns. Using the stop pulse to strobe the instantaneous counter status into a register provides multi-hit capability.

Analog techniques are commonly used in high-resolution digitizers to provide resolution in the range of ps to ns. The principle is to convert a time interval into a voltage by charging a capacitor through a switchable current source. The start pulse turns on the current source and the stop pulse turns it off. The resulting voltage on the capacitor C is $V = Q/C = I_T(T_{stop} - T_{start})/C$,

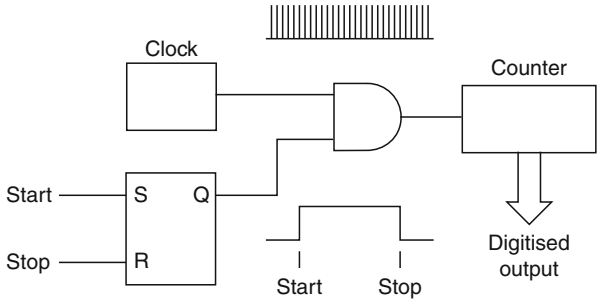


Fig. 24

The simplest form of time digitizer counts the number of clock pulses between the start and stop signals

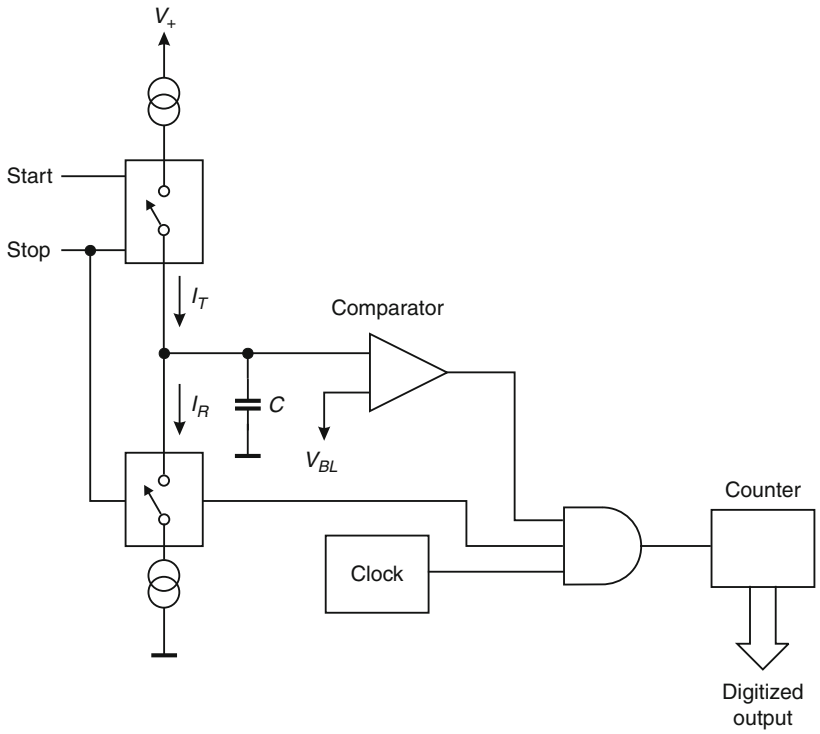


Fig. 25

Combining a time-to-amplitude converter with an ADC forms a time digitizer capable of ps resolution. The memory capacitor *C* is charged by the current I_T for the duration $T_{stop} - T_{start}$ and subsequently discharged by a Wilkinson ADC

which is digitized by an ADC. A convenient implementation switches the current source to a smaller discharge current I_R and uses a Wilkinson ADC for digitization, as illustrated in *Fig. 25*. This technique provides high resolution, but at the expense of dead time and multi-hit capability.

8 Digital Signal Processing

This section was adapted from Spieler (2005 pp 210–216). Up to now we have utilized analog techniques for pulse shaping. However, filtering can also be applied in the digital domain. This is a topic worthy of a book in itself, so this will only be a brief introduction designed to provide some perspective relevant to large-scale detector systems. For a more detailed discussion of digital signal processing techniques, see texts by Ifeachor and Jervis (1993), Oppenheimer and Schafer (1998), and others. For examples applied to detector pulse processing see Pullia et al. (2000) and Cardoso et al. (2004), which also give additional references.

First, the detector signal is sampled with a fast digitizer with sufficient resolution to reconstruct the pulse, as shown in *Fig. 26*. Subsequently, a digital signal processor (DSP) applies the appropriate algorithms to filter the pulse and extract the pulse height (*Fig. 27*). Digital signal processing allows great flexibility in implementing filtering functions, even adapting event by event. The software can be changed readily to adapt to a wide variety of operating conditions, and it is possible to implement filters that are impractical or even impossible using analog circuitry. However, this comes at the expense of increased circuit complexity and increased demands on the ADC compared to analog shaping.

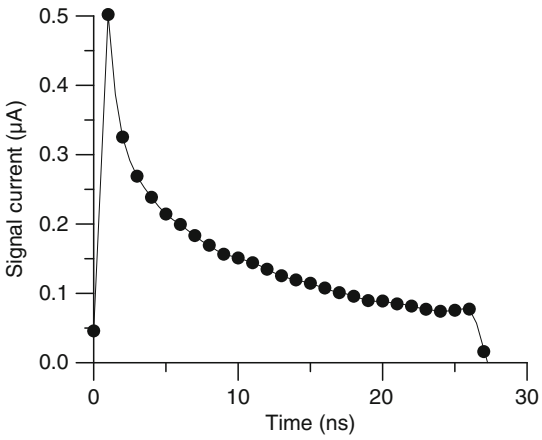


Fig. 26
Sampling a pulse to allow digital signal processing. The pulse shown is the current pulse from a silicon strip detector

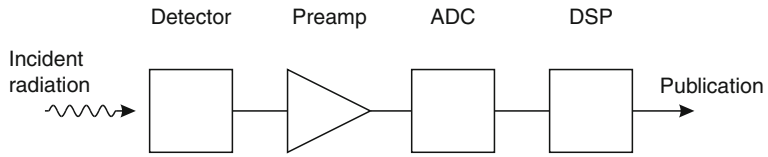


Fig. 27
Block diagram of a detector readout using digital signal processing

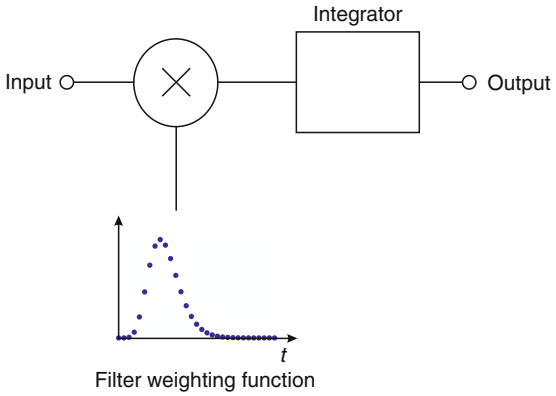


Fig. 28
In a simple digital filter the input signal is multiplied at each discrete time step by a filter weighting function

Figure 28 illustrates how a filter function can be implemented using digital techniques. The amplitude of the input signal is multiplied at each discrete time step by a filter weighting function. The filter function can be calculated in real time by the DSP or it can be stored as values in a look-up table. This process could be applied to either a continuous or a digitized input signal. Subsequently the samples are integrated. Since the amplitudes add coherently, whereas the noise components add in quadrature, this yields a net improvement in signal-to-noise ratio. It is also rather straightforward to show that the optimum signal-to-noise ratio obtains when the weighting function has the same shape as the input signal. This is an example of a "matched filter." However, this is only the optimum filter for retrieving the signal while retaining its shape. As we have seen, integrating the signal to extend its duration and then filtering decouples the choice of filter parameters from the original signal duration.

The simple scheme shown in *Fig. 28* requires that the time of the desired signals is known, so the weighting factors can be synchronized with the signal. This constraint is removed when the filtering is performed by convolution, so the DSP block in *Fig. 27* performs a sequence of multiplications and sums

$$S_o(n) = \sum_{k=0}^{N-1} W(k) \cdot S_i(n-k), \tag{18}$$

where S_o and S_i are the output and input signals and W is the weighting function that yields the desired pulse shape. This is analogous to pulse shaping in analog systems. In digital signal processing this is referred to as a finite impulse response (FIR) filter, similar to an infinite impulse response (IIR) filter, which takes the sum to infinity. Specialized digital signal processors optimized to perform these functions are available, but FPGAs also allow very efficient implementations (e.g. Dobbs 2008). Without special hardware, algorithms can be tested on a desktop computer using realistic detector pulses and noise spectra to assess artifacts in the output spectrum, e.g. using C++ functions (Embree and Danieli 1999).

The sample interval must be sufficiently small to capture the pulse structure. *Figure 29* shows the same pulse as in *Fig. 26*, but sampled at intervals of 4 ns instead of 1 ns. The sampling interval of 4 ns misses the initial peak.

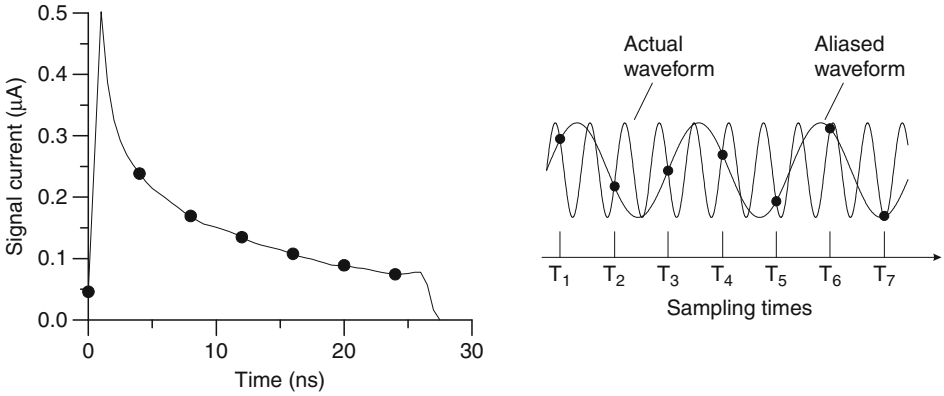


Fig. 29

Sampling at too low a rate does not preserve the full pulse structure (*left*) and also leads to "aliasing," i.e. reconstruction at lower frequencies

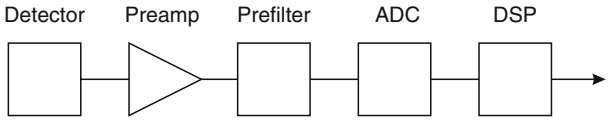


Fig. 30

A low-pass filter (prefilter) inserted in the ADC input prevents aliasing of high-frequency components into the desired frequency range

This illustrates the Nyquist criterion. The ADC must digitize at greater than twice the rate of the highest frequency component in the signal. Apart from missing information on the fast components of the pulse, undersampling introduces spurious artifacts. With too low a sampling rate, high-frequency components will be "aliased" to lower frequencies, as shown in *Fig. 29*.

To prevent aliasing, a low-pass filter must be introduced before the ADC. As a result, an additional analog block must be added to the signal processing chain (*Fig. 30*). When an input frequency f_i is sampled at a rate f_s, the output frequencies can be reconstructed as $f_i \pm k f_s$, where k is any integer value. Thus, the input is aliased to both lower and higher frequencies and the prefilter ("anti-aliasing filter") is needed to exclude both possibilities. Every sampling process is subject to aliasing – e.g. also 2D or 3D image processing.

Note that aliasing can be avoided at sampling rates lower then than the Nyquist criterion if the sampling is synchronized with the events, as can be achieved at collider accelerators, where events occur only at the collision time. Then the sampling time can be synchronized with the collision clock. In other applications where the pulse shape does not have to be reconstructed, lower sampling rates are sometimes also acceptable. When only the energy is to be measured by measuring the energy in successive time bins and then integrating them over the appropriate duration to obtain the total energy, the best technique is to integrate the signal within each time bin rather than merely sampling the amplitude. This can be done with a clocked integrator, i.e. a switched integrator that is synchronous with the digitizer clock frequency. Energy integration can be approximated by amplitude sampling if the pulse shape is known. With an exponential

decay with a known time constant, the amplitude measured at a time T also determines the amplitude at the time $T - \Delta T$. The uncertainty comes about in the time where the pulse rises and then assumes an exponential decay. If this occurs within one time bin, the difference in time between the beginning of the pulse and the sampling time is not known. This leads to an uncertainty that can be calculated and if necessary reduced by increasing the sampling rate.

The preamplifier is necessary to raise the level of the input noise sources such that the digitization noise of the ADC is negligible. The signal quantization inherent to the digitization process introduces quasi-random noise

$$\sigma_n = \frac{\Delta V}{\sqrt{12}}, \tag{19}$$

where ΔV is the signal increment corresponding to one bit. This quantization noise is increased by differential nonlinearity. When the Nyquist condition is fulfilled, the noise is spread nearly uniformly and extends to 1/2 the sampling frequency f_S, so the spectral noise density

$$e_n = \frac{\sigma_n}{\sqrt{\Delta f_n}} = \frac{\Delta V}{\sqrt{12}} \cdot \frac{1}{\sqrt{f_S/2}} = \frac{\Delta V}{\sqrt{6 f_S}}. \tag{20}$$

Sampling at a higher frequency spreads the total noise over a larger frequency range, so oversampling can be used to increase the effective resolution.

From this we see that the front-end electronics and ADC must exhibit the same precision as in an analog system, i.e. the baseline and other pulse-to-pulse amplitude fluctuations must be less than order $Q_n/10$, i.e. typically 10^{-4} in high-resolution systems. For 10 V full scale at the ADC input in a high-resolution gamma-ray detector system, this corresponds to <1 mV. In practice the effective resolution of ADCs suitable for these applications is commonly 2 bits worse than nominal, so this must be taken into account. At very high resolution the electronic noise of the ADC's input circuitry becomes the limit. For example, in a 24-bit ADC with a full-scale range of 10 V, one bit corresponds to a voltage difference of 0.4 nV. The thermal noise of a 50 Ω resistor in 1 Hz bandwidth is more than twice as large. Furthermore, the dynamic range requirements for the ADC may be more severe than in an analog filtered system, as can be seen from the rather high peak-to-average ratio of the pulse in ❯ *Fig. 26*. In any case, the ADC must provide high performance at short conversion times.

Today digital signal processing is technically feasible for some applications, e.g. detectors with moderate to long collection times (gamma and X-ray detectors), and systems are commercially available. Nevertheless, these systems tend to be complex and power-hungry.

However, in large-scale systems, e.g. LHC pixel detectors with $\approx 10^8$ channels or strip detectors with $\approx 10^7$ channels, the benefits are not so clear. Where intimate integration of sensors and electronics in a small volume is required, both circuit area and power dissipation are crucial considerations. Furthermore, these are special-purpose systems. The electronics are specifically tailored to the sensor and application and do not need to be modified during the course of the experiments (the inevitable upgrades notwithstanding). Furthermore, simple analog filters usually provide results that are only slightly inferior to the optimized filters that a DSP system would allow.

The benefits of digital signal processing are:

1. Flexibility in implementing filter functions.
2. Filters are possible that are impractical in hardware.
3. Filter parameters can be changed simply.

4. Tail cancellation and pileup rejection are easily incorporated.
5. Adaptive filtering can be used to compensate for pulse-shape variations.

Where is digital signal processing appropriate? It provides clear benefits in systems that are highly optimized for resolution, high counting rates, variable sensor pulse shapes, or pulse-shape discrimination.

Where is analog signal processing best (most efficient)? In systems that require fast time response, the high power requirements of high-speed ADCs are prohibitive. Systems that are not sensitive to pulse shape can use fixed shaper constants and rather simple filters, which can be either continuous or sampled. Finally, in high-density systems that require small circuit area and low power, analog filtering can efficiently transpose the relevant information to a frequency domain where digitization requirements are less demanding.

Given the dearth of good analog-circuit designers and no prospects for improvement, it is often claimed that digital signal processing is a better match to available skills and avoids the need to understand the wide range of details that a sophisticated analog system requires. This argument is specious; both types of systems require careful analog design. Nevertheless, progress in fast ADCs (precision, reduced power) will expand the range of DSP applications.

9 Summary

Signal processing is an important component for many radiation detectors. It can determine the accuracy of amplitude measurements (e.g. energy resolution and position sensing), the maximum event rate, the timing accuracy, and its implementation can also have a great effect on the required power, which is often important in large-scale systems. The requirements depend on the application, so these result from the interplay of detector parameters, electronic noise, and pulse shaping. Choosing the technique can sometimes simply rely on standard recipes, which is the common approach. However, novel measurements often require novel techniques whose implementation requires assessing the balance between the physics limits of the instrumentation and practical technology.

Detector systems depend on the interplay of numerous contributions. Implementing a demanding system does require the knowledge of electronic design details. However, before getting to those steps, developing the overall concept requires an understanding of the experimental requirements and how they translate into detector functions and these in turn must be translated into electronic parameters. What it takes to recognize the key functions in detectors and electronics and assess their effects is an understanding of basic physics and how it applies to practical applications. Computer simulations can predict overall performance, but they must include all critical aspects, which is not always the case. Simply because simulation results agree approximately with expectations does not guarantee that assumptions are correct. Simulation results should always be checked against separate calculations to assess individual contributions and they must also be compared with measurement results. Simulations and measurements should usually agree to a few percent. If they don't, the simulation may be wrong, the measurement may be faulty, or perhaps both. Identifying inconsistencies and finding their causes is often more educational than building something that simply works.

On a personal note, analyzing discrepancies between experimental results and my expectations over the past decades was extremely useful, and as a physicist this also led me to recognize the many engineering aspects that are critical. I also found that the key criteria and

the understanding of practical functions can be derived from basic physics, not just for the detectors, but also for the electronics. Deriving critical aspects from basics has also led to novel designs. Claims that existing experience shows that something cannot be done should not be accepted without going back to the basics. Many detector systems that not so long ago were "impossible" are now taken for granted.

References

ATLAS Collaboration (2008) The ATLAS experiment at the CERN Large Hadron Collider, 2008 JINST 3 S08003

Cardoso JM et al (2004) A high performance hardware reconfigurable hardware platform for digital pulse processing. IEEE Trans Nucl Sci 51(3): 921–925

Dobbs M, Bissonnette E, Spieler H (2008) Digital frequency domain multiplexer for mm-wavelength telescopes. IEEE Trans Nucl Sci 55(1):21–26

Embree PM, Danieli D (1999) C++ algorithms for digital signal processing. Prentice Hall PTR, Upper Saddle River. ISBN 0-13-179144-3, TK5102.9 E45

Garcia-Sciveres M et al (1999) The SVX3D integrated circuit for dead-timeless silicon strip readout. Nucl Instr Meth A435:58–64

Goulding FS (1972) Pulse shaping in low-noise nuclear amplifiers: a physical approach to noise analysis. Nucl Instr Meth 100:493–504

Goulding FS, Landis DA (1982) Signal processing for semiconductor detectors. IEEE Trans Nucl Sci 29(3):1125–1141

Ifeachor EC, Jervis BW (1993) Digital signal processing – a practical approach. Addison-Wesley, Wokingham. ISBN 0-201-54413-X, TK5102.I33

Oppenheimer AV, Schafer RW (1998) Discrete-time signal processing. Prentice Hall, Upper Saddle River, ISBN 0-13-754920-2, TK5102.9.067

Pullia A et al (2000) Quasi-optimum γ and X spectroscopy based on real-time digital techniques. Nucl Instr Meth A439:378–384

Radeka V (1972) Trapezoidal filtering of signals from large germanium detectors at high rates. Nucl Instr Meth 99:525–539

Radeka V (1974) Signal, noise and resolution in position-sensitive detectors. IEEE Trans Nucl Sci 21:51–64

Spieler H (1982) Fast timing methods for semiconductor detectors. IEEE Trans Nucl Sci 29(3):1142–1158

Spieler H (2005) Semiconductor detector systems. Oxford University Press, Oxford, ISBN 0-19-852784-5

Unno Y et al (2003) ATLAS silicon microstrip detector system (SCT). Nucl Instr Meth A511: 58–63

Wilkinson DH (1950) A stable ninety-nine channel pulse amplitude analyser for slow counting. Proc Cambridge Phil Soc 46(3): 508–518

4 Data Analysis

Günther Dissertori
Institute for Particle Physics, Zurich, Switzerland

1 Introduction ***84***

2 From Raw Data to Physics Objects ***85***
2.1 Basics of Track Finding 86
2.2 Energy Reconstruction in Calorimeters 88
2.3 Jet Algorithms 90
2.4 Further Higher-Level Algorithms 90
2.5 Simulations 91

3 Examples from e^+e^- and Hadron Colliders ***93***
3.1 Ratio of the Hadronic and Leptonic Cross Sections in e^+e^- Annihilations . . 93
3.2 Jet Production in Hadron Collisions 96

4 Computing and Software Aspects ***98***

5 Conclusion ***100***

Further Reading ***100***

References ***101***

C. Grupen, I. Buvat (eds.), *Handbook of Particle Detection and Imaging*,
DOI 10.1007/978-3-642-13271-1_4, © Springer-Verlag Berlin Heidelberg 2012

Abstract: The basic goal of data analysis is to establish a link between a set of measurements, in the form of electronically stored data using some format, and a theoretical model, which is intended to describe the phenomena at the origin of these measurements and usually is summarized by a set of equations with some parameters. The key elements of data analysis are data abstraction and data reduction. Abstraction means that the original set of raw measurements, e.g., a collection of electronic pulses induced by a particle passing through a detector, is converted ("reconstructed") into physical quantities and properties which can be assigned to the particle, such as its momentum or its energy. Typically, this process is accompanied by data reduction, i.e., the overall data volume is reduced when going from the original set of measurements to a compilation of reconstructed physical quantities.

In this chapter, I will describe the steps involved in order to achieve the above-mentioned data abstraction and reduction in the case of Particle Physics experiments. Examples will be given for measurements carried out at e^+e^- as well as hadron colliders. The basic concepts behind reconstruction algorithms, such as track finding in tracking detectors and energy measurements in calorimeters, will be discussed, along with higher-level algorithms such as particle-jet reconstruction. Finally, the typical software and computing environment of large collider experiments will be described, which is necessary in order to achieve the outlined goals of data analysis.

1 Introduction

The basic goal of scientific activity in the natural sciences is to inquire about the laws of nature. More concretely, we use experiments in order to investigate (measure) what *reality (nature) does*, and we develop models which are intended to reproduce and explain such measurements, as well as to predict the outcome of future experiments. The role of data analysis is to bridge this gap between reality and a theoretical model.

Such models are formulated using the language of mathematics, i.e., typically by a set of equations which depend on a priori unknown parameters. An example is the Standard Model of Particle Physics. Here we have a Lagrangian density as starting point, which is a function of (particle) fields and their derivatives, and further depends on parameters such as coupling constants for the various interactions and masses of the particles involved. This Lagrangian density is the starting point for deriving predictions for measurable observables, such as scattering cross sections, branching ratios of particle decays, etc., which consequently also depend on those parameters. Once these parameters are determined by comparing the experimental outcome to the prediction, the Lagrangian can be used to predict similar observables for other experiments. If these predictions turn out to be in agreement with observations, we start to talk about a consistent theory of nature, at least as long as no experiment is carried out which proves the model to be wrong.

On the experimental side, the outcome of an experiment is quantified in a set of numbers, basically the readings of a set of detection systems (*the detector*). These *raw data*, which for modern experiments can attain enormous sizes of order tera- or petabytes, are the necessarily imperfect measurements of a number of interactions in the detector. In Particle Physics experiments, these data are organized in unique subsets, called *events*. These characterize the outcome of a specific happening, such as the scattering of two particles, when counter-rotating beams in

a collider cross each other inside an experiment. This results in a group of particles and decay products, which induce the signals in the detector.

The aim of the following sections is to describe how data analysis is employed in order to confront theory with experiment by comparing the measured quantity, an observable, with its prediction from theory. Evidently, we have to bridge the gap between the raw data, a set of detector readings, and the theoretical formula for the observable. Before entering the heart of the matter, it is worth mentioning that a prior knowledge in basic Particle Physics, both theory and experiment, will be assumed. Good examples of textbooks at the introductory level are Perkins (2000) and Griffiths (2008).

2 From Raw Data to Physics Objects

In the following, we will learn about the basic steps of an analysis, from the triggering and detector readout, via the reconstruction of basic quantities and higher-level algorithms, to the detector simulation and the final comparison of data and theory. The basic themes accompanying us will be *data reduction* and *data abstraction*.

The data-analysis chain starts with the collection of data from many channels (up to millions) in the various sub-detectors. However, for modern experiments, it is not feasible to write all those data to permanent storage because of the sheer amount of storage space needed. Furthermore, a very large data volume induces considerable inefficiencies and long turnaround times for later analyses stages. Therefore, a *trigger system* takes a fast decision, based on a subset of readings from fast detector components, if an event will be fully read out or completely discarded (see also ❯ Chap. 2, "Electronics Part I" and ❯ Chap. 3, "Electronics Part II"). At this trigger stage, the data flow is reduced from rates in the kHz to MHz range, at which the raw detector data arrive from the detector front-end electronics to the trigger system, down to 10–100 Hz, at which the accepted events are written out to a first level of permanent storage. Such huge rate reductions can only be achieved by splitting the trigger decision sequence into various steps and using parallel processing and pipelining. This allows spanning the range from very short decision times, as achievable by logics implemented in dedicated and custom-designed electronics, to longer decision times needed in order to run sophisticated algorithms on large computer clusters. An important consequence of the sub-detector concept and the parallelized readout approach is the need to combine the data fragments from the various sub-detectors into a coherent set. The combination takes place before the final write-out to storage, such that the fragments are collected into the same unique event (now in terms of data storage) which they originally belong to. This is achieved by powerful network switches and so-called event builders.

The outcome of this *Data Acquisition (DAQ) or Online chain* is a set of raw data, basically the collected detector readings, organized in an event structure. The goal of the subsequent step, the *Offline Analysis chain*, is to further reduce the original data volume by reconstructing from the raw detector readings higher-level quantities (*Physics objects*), which are more directly related to concepts inherent in the theoretical models. This data reduction and data-abstraction step reduces the data volume from the tera- and petabyte level to more manageable gigabytes, and finally to kilobytes, which is the size of some computer graphics representing the distribution of a certain observable quantity. Ultimately, such a distribution is then compared to similar distributions, but now obtained with theoretical models. It is worth emphasizing that this reduction in data volume leads to a considerable reduction in turn-around time for data analyses, which are carried out on the higher-level quantities, thus to an increased working

efficiency for everybody involved. The reconstruction of Physics objects and the consequent data reduction and abstraction is best described by a few examples.

2.1 Basics of Track Finding

The basic steps in the reconstruction of charged particle tracks and the measurement of the particle's momentum are illustrated in ❯ *Fig. 1*. We assume that the passage of a charged particle is registered by a set of detector layers, where an analog signal is produced by the interaction of the particle with the detector medium. A widely used example is a tracking system composed of silicon microstrip detectors. The information stored in the analog pulses is digitized, i.e., if the pulse height is above a certain threshold the location of a *hit* and its exact time are stored. The three spatial coordinates of a hit can be retrieved from the known location of a sensor as well as from the location of the particular channel on a sensor which gave the signal (e.g., a strip or pixel on a modern silicon sensor). In the case of a modern Particle Physics experiment, typically there are many charged particles produced in a process, leading to many hits stored in the case of a single event (identified by the timing information). The next step consists in reconstructing the correct flight path of a particle by assigning a subset of all the found hits to a so-called *track* hypothesis. Details of such a track-finding step are described elsewhere (see also ❯ Chap. 12, "Tracking Detectors") in this textbook. In the simplest case, we would look for hits which are consistent with lying along a straight line. However, if the whole tracking

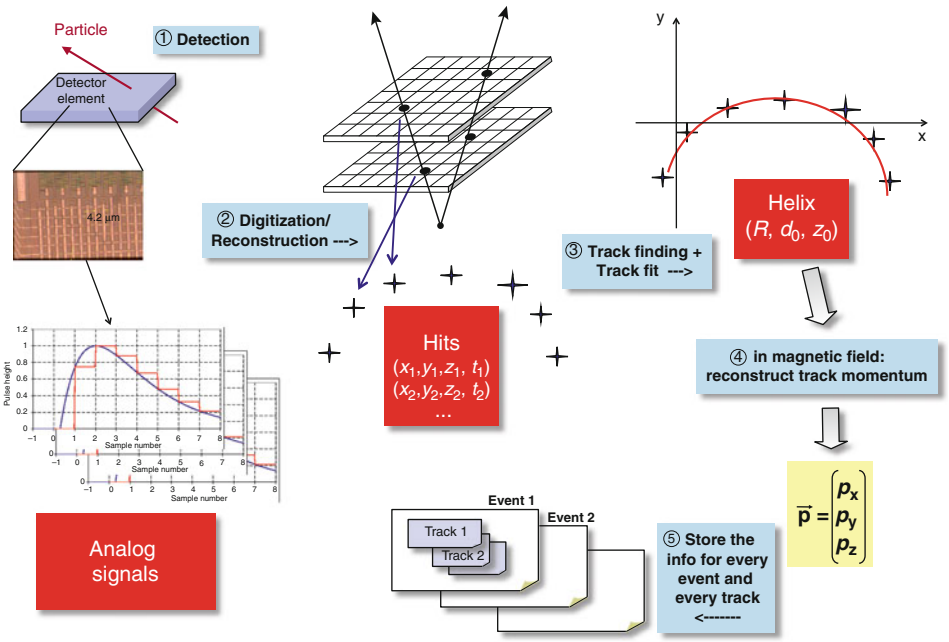


Fig. 1
Illustration of the steps necessary to reconstruct and store the momentum of a charged particle

detector is immersed in a homogeneous magnetic field, the underlying hypothesis for track finding is a helix, i.e., a circular path perpendicular to the direction of the magnetic field and a straight path parallel to it. After a set of hits are identified which are consistent with such a flight path, the relevant information can be reduced to the radius of the circular path as well as the distances of closest approach to the assumed origin, which in the case of collider experiments would be the center of the detector where the accelerator beams collide. Knowing the curvature of a track, the strength of the magnetic field and the track's direction at its origin, it is trivial to deduce the particle's momentum (we also assume the particle to carry one unit of electric charge) by basically using the expressions for a particle's trajectory under the influence of a Lorentz force. At this stage, the original information, a very large set of analog or digital signals, has been reduced to a set of three numbers, namely, the three components of the particle's momentum. By looking at the orientation of the helix curvature, also the charge sign can be identified. This information is then stored for all identified particle tracks in an event. Thus, the stored information is (1) reduced compared to the original set of measurements and (2) brought to a more abstract level, i.e., the number of charged particles found in an event, the sign of their charges, and their three momentum components. This basic information can then be retrieved in an efficient way afterwards, for usage in higher-level algorithms (such as vertex finding, resonance-mass reconstruction, jet algorithms) or for simply displaying the reconstructed event, as shown in ❯ *Fig. 2*.

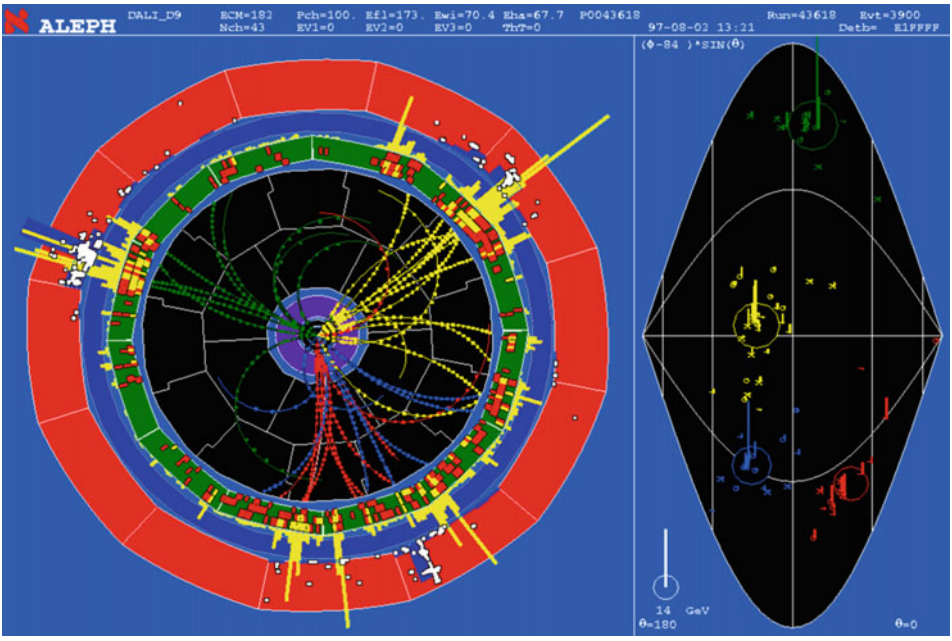


Fig. 2

Event display of an e^+e^- annihilation into hadrons as recorded by the ALEPH experiment at the LEP collider (see also ❯ Chap. 7, "Accelerators for Particle Physics"). The reconstructed charged-particle tracks, together with their assigned hits, are displayed, as well as the energy depositions in the calorimeters shown as histograms

2.2 Energy Reconstruction in Calorimeters

In addition to a tracking system, modern Particle Physics detectors are also equipped with calorimeter systems (Wigmans 2003). Here the basic aim is to measure the energy deposited by neutral particles, such as photons and neutrons, which do not leave tracks. From the measurement of the impact point and the extrapolation to the assumed particle's origin, which usually is the collision spot along the beam line, the four-momentum can be estimated. However, calorimeters are also used to measure the energy of charged particles, such as pions and electrons, as well as of particle jets, which will be discussed below in ❯ Sect. 2.3. Since the momentum resolution of a tracking system goes as $\sigma_p/p \propto p$, whereas the relative energy resolution of a calorimeter is parametrized as $\sigma_E/E \propto 1/\sqrt{E}$, we see that at very high energies it becomes more favorable to use a calorimeter to reconstruct the four-momentum of a particle, instead of a tracking detector.

In general, calorimeters are segmented in cells of certain sizes $\Delta\eta \times \Delta\phi$, where $\eta = -\ln\tan(\theta/2)$ is called *pseudo-rapidity*, and ϕ (θ) are the azimuthal (polar) angle w.r.t. the beam line. A fine segmentation has several advantages. Because of the shower induced by an impinging particle, the resulting energy deposits are spread out over several cells. Therefore, the exact position of the impact can be precisely reconstructed by an (energy-)weighted sum of the hit cell positions. Furthermore, high granularity can help to distinguish the shower pattern as induced by a single particle, e.g., a single isolated photon, or by two very nearby particles, such as two photons originating from the decay of a highly boosted neutral pion (π^0). Finally, the substructure of the energy deposits induced by a set of particles, which, e.g., stem from the same jet, can be useful to deduce certain properties of such a jet.

In ❯ *Fig. 3* (upper-left), we see a schematic drawing of an electron, which is first bent by a magnetic field before hitting the surface of a calorimeter. The resulting shower induces signals in a few cells. Furthermore, possible photons from earlier bremsstrahlung will lead to additional nearby clusters. The bending occurs in the orthogonal plane w.r.t. the magnetic field, which typically is parallel to the beam line. Therefore, an algorithm intended to reconstruct the total original energy of the electron should search for such additional photon clusters along a narrow road in the ϕ direction. On the upper right of the figure, the structure of a calorimeter (see also ❯ Chap. 20, "Calorimeters") cluster is represented in the $\eta-\phi$ plane, where a single cell corresponds to a concrete readout unit of the detector. Normally, the cell sizes are chosen to be similar to the expected lateral extension of a particle shower. Therefore, most of the signal will be concentrated in a few cells, with some tails around it. The basic goal of a reconstruction algorithm is to assign the right cells to the presumed shower and then to add up their energy deposits in order to determine the total energy. The difficulty lies in the fact that electronic noise, or other close-by particles, can lead to signals which are wrongly assigned to the original particle's cluster. A very simple algorithm would start by looking for the cell with the maximal energy deposit, which in addition has to be greater than some predefined threshold, as, e.g., visible in the lower part of ❯ *Fig. 3*. Next, the algorithm looks for adjacent cells with energy deposits above a second chosen threshold, which is intended to discriminate between real energy deposits and noise. If such a cell is found, it has to be checked if it already belongs to another cluster and if the cell previously added to the cluster has higher energy because of the assumption of a lateral falloff of the shower. The algorithm stops when no further adjacent cells are found which fulfill these criteria. The exact search sequence in the plane around the seed cell might be optimized for specific particle types, such as electrons which lose energy by bremsstrahlung mainly along the ϕ direction. Obviously, the exact values of the thresholds have to be optimized by the experimenter.

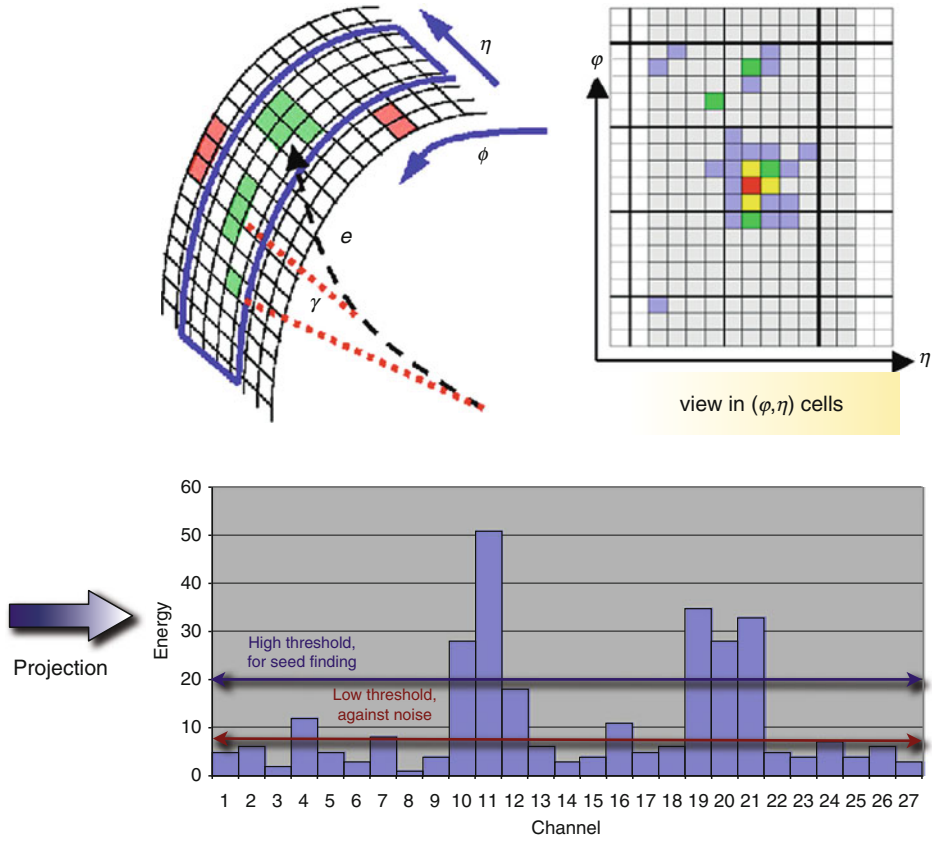


Fig. 3

Illustration of the cell structure of a shower, which is induced by a particle hitting the surface of a calorimeter

A too low noise threshold will cause the algorithm to pick up too much noise, leading to an overestimation of the true energy, whereas a too high threshold will lead to an underestimation since some of the real energy deposits may fall below the threshold and therefore will not be accounted for. It goes without saying that concrete implementations of such clustering algorithms go much beyond in sophistication than what is described here. However, again the basic principles are data reduction and abstraction since ultimately a number of detector readouts is reduced to a single value, namely, the estimated energy of a particle which has induced these signals.

It should be noted that this estimated energy not necessarily corresponds to the true energy of the particle. A *calibration correction* has to be applied in order to obtain such a correspondence. The calibration factor corrects for the calorimeter's response to a certain true energy deposit, i.e., typically not all of the energy deposited leads to a measurable signal. Also the effects of the thresholds discussed above need to be taken into account. Such calibration corrections can be obtained by exposing the calorimeter to particle beams of well-known type and energy.

Also kinematic constraints in the reconstruction of particle decays can be used, such as known masses of unstable particles, whose decay products induce the calorimeter signals.

2.3 Jet Algorithms

Because of the nature of strong interactions, partons (quarks or gluons) which are produced in a particle interaction will not be observed as such in an experiment. They are bound to form hadrons, mesons, and baryons, which will lead to detectable signals. Additional radiation of gluons by the original parton or gluons splitting into quark–antiquark pairs, which occur before the hadronization phase, will lead to a set of final-state hadrons which are close in phase space, e.g., observed in a limited angular region in the detector. Such sprays of particles, as easily seen in ❯ *Fig. 2*, are called *jets*. The exact definition of a jet, i.e., the exact assignment of detected particles to a single jet and the resulting determination of the jet's four-momentum, is to be implemented in a jet algorithm. A suitable algorithm will be able to reconstruct jets and related four-momenta which match to high precision the original parton's energy and direction. Since the inputs to such a jet algorithm will be energy deposits in a calorimeter, charged-particle tracks, or a combination thereof, we can view a reconstructed jet as a higher-level Physics object. Here the abstraction in terms of data interpretation is brought to a next level since the observable in question, the momentum of a parton, is only accessible via previous reconstruction of charged- and neutral-particle momenta and their correct assignment to a unique jet.

The literature of modern jet algorithms, as applied at lepton and hadron colliders, is rather rich. Here we cannot give full account of various implementations, but rather refer to Salam (2010). Usually, the basic idea behind a jet algorithm is to identify reconstructed objects (tracks, calorimeter clusters (see also ❯ Chaps. 6, "Particle Identification," ❯ 12, "Tracking Detectors," and ❯ 20, "Calorimeters")) which are close-by in phase space, search for such "neighbors" in an iterative way, and to combine them (in terms of summing up the momenta) with already assigned objects. Being close in phase space can be defined by a narrow angular cone in η–ϕ space or by a certain metric in momentum space, such as the relative transverse momentum of two objects.

2.4 Further Higher-Level Algorithms

Similarly to jet algorithms, the information obtained by reconstructed objects such as tracks and calorimeter clusters serves as input to other higher-level algorithms. Here we only list some of the most relevant ones, without going into any detail.

Particle identification spans quite a large spectrum. First there are basic signal patterns in large detector systems. These help to distinguish charged from neutral particles (an energy deposit in the calorimeter with or without a track pointing to it), electrons from pions by the almost complete absorption of the former in the early calorimeter layers, or muons from pions since the former are able to penetrate the complete calorimeter system and then leave further tracks in outer tracking stations, whereas pions are absorbed. In addition, specific measurements of the energy loss ($\mathrm{d}E/\mathrm{d}x$) or the time of flight, combined with other independent measurements such as energy or momentum, are employed to distinguish particles of different types, mostly of different masses.

By extrapolating a track to the surface of the calorimeter, a certain energy deposit around the extrapolated impact point can be compared to the particle's momentum as estimated from the tracking system. If the measured calorimeter energy is consistent with the momentum measurement from the tracker, the latter is used for assigning a momentum to this charged particle and the linked cluster is removed from the list of all calorimeter clusters. Remaining excess energy deposits or clusters, which are not assigned to any charged particle, are then identified to stem from neutral particles. Overall, in such a procedure called *energy* or *particle flow*, the maximum amount of information about an event can be reconstructed, with the aim to optimize the resolution of jets, of the total energy, or of the missing transverse energy in an event. The latter is simply the negative vector sum of the reconstructed transverse momenta of a list of objects (e.g., all calorimeter clusters, all tracks, all particle flow objects, or all identified jets and leptons). Since the total transverse momentum of the incoming beams in a collider experiment is zero, a nonvanishing missing transverse energy indicates the production of one or more (weakly interacting) particles which crossed the detector without leaving any signal. Such particles can be neutrinos as produced in W-boson decays, or newly predicted particles such as neutralinos in supersymmetric extensions of the Standard Model. Here it is important to have a very hermetic and well-instrumented detector since any normal particle passing across noninstrumented regions will induce an artificial missing energy.

Reconstructed tracks, in particular their extrapolations to the region close to the beam line, are used to search for evidence of the production of particles with lifetimes of order picoseconds, such as hadrons with heavy-quark content. This is achieved by either a direct reconstruction of secondary vertices, i.e., crossings of tracks which in terms of the track direction uncertainties are significantly distant from the main interaction point, or by measurements of a track's *impact parameter* and its significance. The impact parameter represents the closest approach of a track to the main interaction point. For tracks originating from decays of unstable particles, the impact parameter, on average, is significantly larger than for those directly coming from the main interaction. For example, hadronic jets originating from b quarks are tagged, with a certain efficiency and purity (see below for definitions of these quantities), by a suitable combination of the impact parameters of all tracks assigned to the jet or by the identification of secondary vertices within the jet (cf. ❯ *Fig. 4*).

The production of short-lived resonances can be identified by calculating invariant masses using the reconstructed momenta of the presumed decay products. Since usually these cannot be identified directly as such, simply all possible combinations of particles in an event are taken and an enhancement at the correct (resonance) mass value should appear above a so-called combinatorial background. Of course, if secondary vertices are found, only those tracks will be used in the invariant-mass calculation which are assigned to the specific vertex.

2.5 Simulations

A systematic and detailed simulation of the experiment, usually adopting Monte Carlo integration techniques, is an important component of basically all data analyses. Here both the fundamental processes and interactions, as they occur, e.g., in particle-beam collisions, are simulated according to some theoretical model, as well as the interactions with the detector elements of those particles which are produced in these processes. Thus, the real detector and DAQ chain is replaced by a simulation on large computer farms. However, the results of such simulations are stored in exactly the same format as the real data, which then allows to apply the same

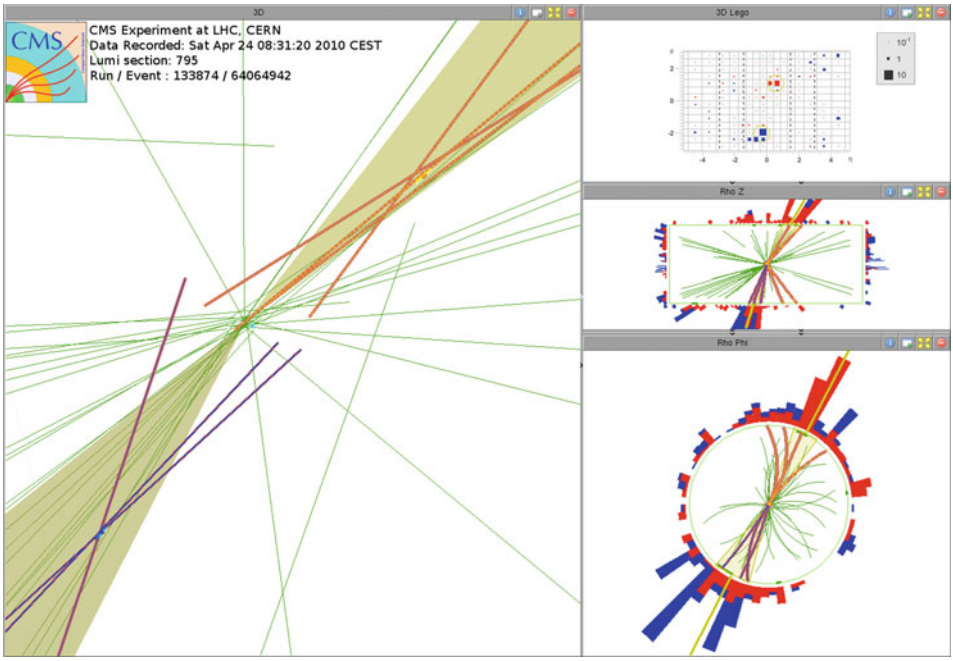


Fig. 4

Display of a multi-jet event as recorded with the CMS detector at the Large Hadron Collider (LHC), where secondary vertices are found in two of the jets

reconstruction and further analysis algorithms and to finally produce simulated analysis results. The difference is that here the *truth* is known, i.e., the underlying process and all its resulting particles, interactions, and kinematics. The purpose of such simulations is to understand the detector response, to determine resolution functions, efficiency, and acceptance corrections for lost particles because of non-instrumented detector regions or limitations of the reconstruction algorithms. Also backgrounds from similar processes or spurious detector effects can be analyzed. We will encounter some of these concepts below.

As such, Monte Carlo simulations are central tools for establishing the workflow of data analyses and monitoring their progress. By comparing simulated and real data distributions, we quantify the level of understanding of the underlying Physics model and of the detector response. Finally, such simulations are of paramount importance already at the early design stage of an experiment since they allow to optimize specific detector choices in terms of the final reach and precision of Physics measurements.

It is worth noting that the Monte Carlo method is much more general in scope than described here. Indeed, it is used in many fields of research, from basic mathematics and statistics to computer science, biology, and financial engineering, to mention only a few. Reference Fishman (1996) might serve as a starting point for the interested reader, but it is only one example from an obviously very rich literature on this topic.

3 Examples from e^+e^- and Hadron Colliders

In the following, we will describe further analysis steps, such as event selection, event counting, estimation of resolution, and calibration and the determination of statistical and systematic uncertainties. All these steps are carried out using the information available after reconstruction, as described above, and are necessary in order to finally compare a measurement to theoretical predictions. Again, it is best to take some concrete examples in order to elucidate the most important concepts.

3.1 Ratio of the Hadronic and Leptonic Cross Sections in e^+e^- Annihilations

A classic observable in e^+e^- annihilations is the ratio of the hadronic and leptonic cross sections (for further details see, e.g., Dissertori et al. (2003)). From *Fig. 5*, it is clear that the initial phase of multi-hadron production is very similar to the creation of muon pairs. Taking the ratio, everything cancels except the coupling strengths. In the energy range above the Upsilon

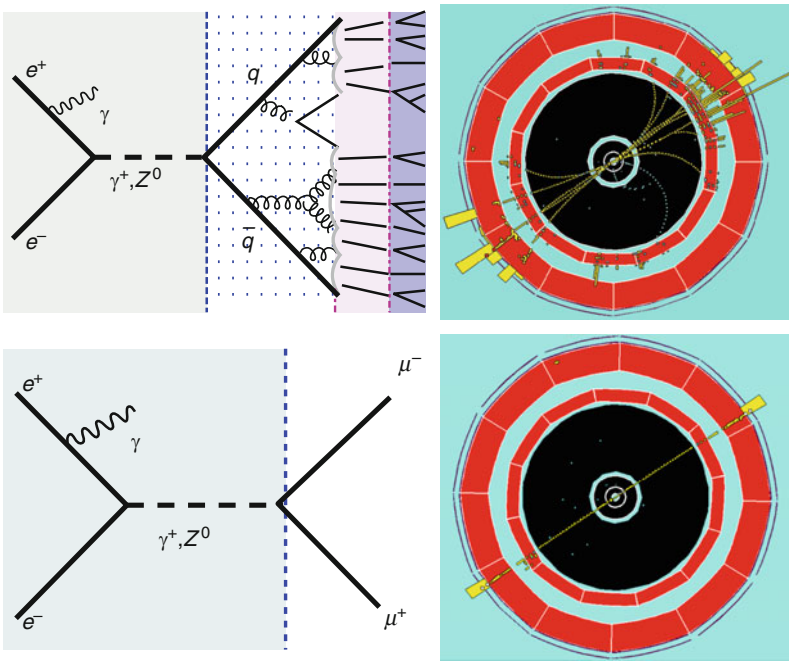


Fig. 5

Basic Feynman diagrams and corresponding event displays for e^+e^- annihilation into hadrons (*top*) and a muon–antimuon pair (*bottom*). The event displays are from collisions recorded by the ALEPH experiment at the LEP collider

resonances, where five quark flavors contribute, and below the Z resonance, where only virtual photon exchange is relevant, we expect at lowest order in perturbation theory that

$$R_\gamma = \frac{\sigma(e^+e^- \to \text{hadrons})}{\sigma(e^+e^- \to \mu^+\mu^-)} = \frac{\sigma_{\text{had}}}{\sigma_{\text{lep}}} = N_c \sum_q e_q^2 = N_c \frac{11}{9}. \quad (1)$$

Here N_c indicates the number of different color states of quarks and the sum runs over all flavors, which are kinematically allowed. For five flavors (up, down, strange, charm, and bottom) and their fractional charges e_q, we arrive at the result above. By comparing this theoretical prediction to the measurement, it should be possible to determine or at least constrain the number of colors, N_c.

In the simplest terms, the experimental task consists in counting how many events with hadronic or leptonic final state are registered over a certain data-taking period. Since we are dealing with a ratio, the accelerator luminosity $\mathcal{L}$ (see also ❯ Chap. 7, "Accelerators for Particle Physics") cancels out, i.e., $R_\gamma = \sigma_{\text{had}}/\sigma_{\text{lep}} = (N_{\text{had}}/\mathcal{L})/(N_{\text{lep}}/\mathcal{L}) = N_{\text{had}}/N_{\text{lep}}$. Now the basic question is how to identify if a certain event contains a hadronic or leptonic (i.e., $\mu^+\mu^-$) final state. Thus, we are left with the problem of event selection, its related efficiency, and backgrounds. More generally, we are dealing here with the concept of hypothesis testing, as indicated in ❯ *Fig. 6*. Using simulations we can identify observables, which characterize a specific event and which help to separate various hypotheses about the event's origin. In our specific case, it turns out that the number of reconstructed charged tracks, N_{tracks}, is a suitable quantity. Whereas for leptonic final states we expect only a few charged tracks, in most cases two as induced by the muon and the anti-muon, in the hadronic case much more charged tracks will be produced. Therefore, by applying a so-called *cut* on this observable, each measured event is classified as leptonic ($N_{\text{tracks}} < N_{\text{cut}}$) or hadronic ($N_{\text{tracks}} \geq N_{\text{cut}}$). In an ideal case, each real event of a certain category would also be identified as such. However, again indicated in ❯ *Fig. 6*, the expected distributions of the two hypotheses may have some overlap region. For example,

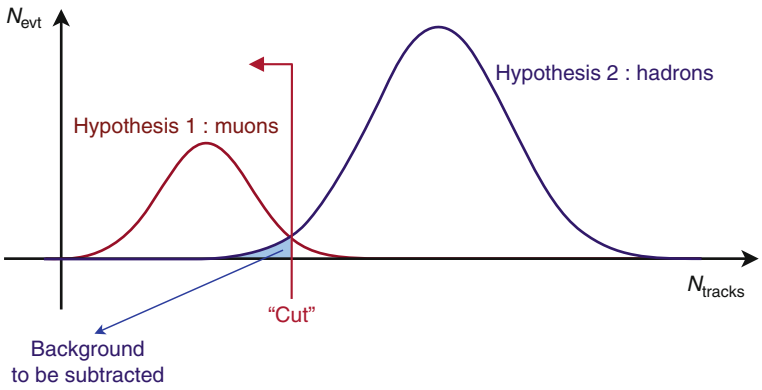


Fig. 6

Graphical representation of the event-selection principle, where distributions of observable quantities are analyzed with respect to their separation power between different hypotheses. In this case, the expected distributions of the number of charged tracks are displayed in a qualitative manner, for the hypotheses of e^+e^- annihilations into muon–antimuon pairs or into hadrons

by applying a cut not all of the true leptonic or hadronic events will be counted as such. This results in an *efficiency*, $\epsilon \leq 1$, to count a certain type of events, e.g., $N^{\text{meas}}_{\text{had/lep}} = \epsilon_{\text{had/lep}}\, N^{\text{true}}_{\text{had/lep}}$. We are interested in $N^{\text{true}}_{\text{had/lep}}$. Therefore, before taking the ratio of the event counts, these counting results $N^{\text{meas}}_{\text{had/lep}}$ have to be corrected for the respective efficiencies. Since in a simulation everything is known about an event, namely, both its true origin as well as its measured final state, typically such efficiencies are determined from there. However, sometimes it is possible to estimate selection efficiencies also from data alone, in particular if certain physical constraints exist. An example is the *Tag-and-Probe* approach, as applied to the determination of an experimental lepton reconstruction efficiency. Here the masses of well-known resonances, such as the Z boson or quarkonia, are employed as constraints. First such resonances are identified via their lepton–antilepton decays by applying a tight selection on one of the reconstructed lepton candidates and a rather loose selection on the second candidate. If the invariant mass of the two leptons falls within a chosen mass window around the true resonance mass, there is a very large probability that indeed both candidates are true leptons. At this stage, we can determine a relative reconstruction efficiency by counting how many of the loosely selected candidates remain after applying some further selection criteria. Hence, we are independent of any simulation or model prediction, at least modulo some overall efficiency for finding leptons after the loose identification.

Before leaving the discussion about efficiencies, it is worth noting that often the overall efficiency for selecting a class of events or particles is split into several components, e.g., $\epsilon = A \cdot \epsilon_{\text{tri}} \cdot \epsilon_{\text{rec}} \cdot \epsilon_{\text{cut}}$. The acceptance A describes the fraction of particles which are *detectable* in the first place. For example, if the detector covers only part of the whole solid angle, and particles can reach a detector element only for a momentum above a certain threshold, then the acceptance is the fraction of all produced particles which fall within this detectable phase-space region. The remaining factors describe the fractions of all detectable particles which (1) also give a trigger signal (ϵ_{tri}), (2) are reconstructed from the detector signals (ϵ_{rec}), and (3) remain after additional selection cuts (ϵ_{cut}).

Besides the efficiency aspect, for an event selection, we also have to worry about *backgrounds*. In our simple example, a background can arise from a tail of one of the hypothesis distributions, which extends beyond the region fixed by a cut value. More concretely, for a certain choice of N_{cut}, it can happen that hadronic events with a very small number of charged tracks are falsely classified as of leptonic type. Thus, the true number of leptonic events is given by $N^{\text{true}}_{\text{lep}} = (N^{\text{meas}}_{\text{lep}} - N_{\text{bckg}})/\epsilon_{\text{lep}}$. In many cases, the expected background contribution for a specific event selection is again determined from simulation. However, often quite some effort is undertaken in order to develop purely data-driven approaches to background estimations. This means that all necessary information is directly extracted from the data themselves, instead of relying on simulations and models. The simplest example is the determination of the background under a peak in an invariant-mass distribution, as induced by the presence of a particle resonance. Here the expected number of background events is extrapolated from the observed number in the so-called *sidebands*. These are regions in the distribution, which are distinct from the expected resonance mass and thus not affected by any peak-like enhancement. Unfortunately, a more extensive description of such approaches goes beyond the scope of this chapter.

Like every experimental measurement, the efficiency- and background-corrected numbers of hadronic and leptonic events have a statistical and a systematic uncertainty. The statistical part of the uncertainty will not be discussed any further here since it is covered quite extensively in the literature, see e.g., Cowan (1998), see also Chap. 5, "Statistics." The determination

of a systematic uncertainty is much less straightforward. Indeed, no generally valid prescriptions exist, and here only some general concepts and problems can be addressed. It is in the evaluation of systematic uncertainties where the careful (and sometimes subjective) judgment of the experimenter enters by trying to identify all possible sources of systematic effects which could influence the measurement. Typically, most of the time and efforts spent on a specific measurement go into the study of the systematic uncertainties. If we take our simple example and assume that the selection efficiency is obtained from a simulation, we have to investigate on the reliability of the estimated number. For example, it might be that the response of a detector element to charged particles is wrongly modeled in the simulation, which then directly translates into a bias or at least an uncertainty on the efficiency. An estimate of the related uncertainty might be obtained by comparing all relevant distributions in data and simulation, which directly or indirectly probe the response of the detector. From the overall agreement or disagreement of such distributions it might be possible to give bounds on the estimated efficiency. A traditional approach consists in changing the value of the cut since for a perfect simulation of the real experiment any value of chosen cut parameter should lead to the same final efficiency- and background-corrected result. Observed deviations are attributed to non-perfect simulations of the hypothesis distributions and quoted as uncertainty. It goes without saying that such uncertainties can and should not be interpreted in the same way as statistical errors.

3.2 Jet Production in Hadron Collisions

By discussing the measurement of jet production at hadron colliders, we can address further data-analysis aspects, which are relevant whenever a steeply falling distribution is measured. Therefore, the concepts discussed below apply to other research fields as well.

As a concrete example from Particle Physics, we study the determination of the inclusive jet cross section in hadron–hadron collisions. This is a measurement of the cross section (probability) that a jet is produced with a given transverse energy $E_T = E \sin\theta_{\text{jet}}$ within a certain pseudo-rapidity range $\Delta\eta_{\text{jet}}$ (cf. ❯ Sect. 2.2). Here E and θ_{jet} are the reconstructed jet energy and polar angle of the jet axis, respectively, as determined by a given jet algorithm (❯ Sect. 2.3). Such a measurement constitutes an important test of the theory of strong interactions, Quantum Chromodynamics (QCD). Furthermore, deviations from the QCD predictions in the high-energy tail of the measured spectrum could be signals for new physics, such as new contact interactions or quark substructure. An example of a measurement by the D0 collaboration at the Tevatron proton–antiproton collider (see also ❯ Chap. 7, "Accelerators for Particle Physics") is shown in ❯ *Fig. 7* and further discussed in Dissertori et al. (2003). We observe a steeply falling spectrum over many orders of magnitude, which has several experimental implications.

Whereas previously we only talked about measurements of overall event rates, now we are interested in measuring a spectrum, i.e., a cross section as function of some measurable quantity, in this case E_T. The first step consists in dividing up the accessible E_T range into so-called *bins*, which are simply contiguous intervals ΔE_T. At this stage, the measurement again can be viewed as a counting experiment, where for each event we count all reconstructed jets, which fall within a specific bin of E_T and a particular range of pseudo-rapidity $\Delta\eta_{\text{jet}}$. Thus,

$$\left\langle \frac{\mathrm{d}^2\sigma}{\mathrm{d}E_T\,\mathrm{d}\eta_{\text{jet}}} \right\rangle = \frac{N}{\Delta E_T\,\Delta\eta_{\text{jet}}\,\epsilon\,\mathcal{L}}\,, \tag{2}$$

where the left-hand side represents the differential cross section, either evaluated at the center of a specific bin or averaged over it. This distinction becomes more and more relevant the

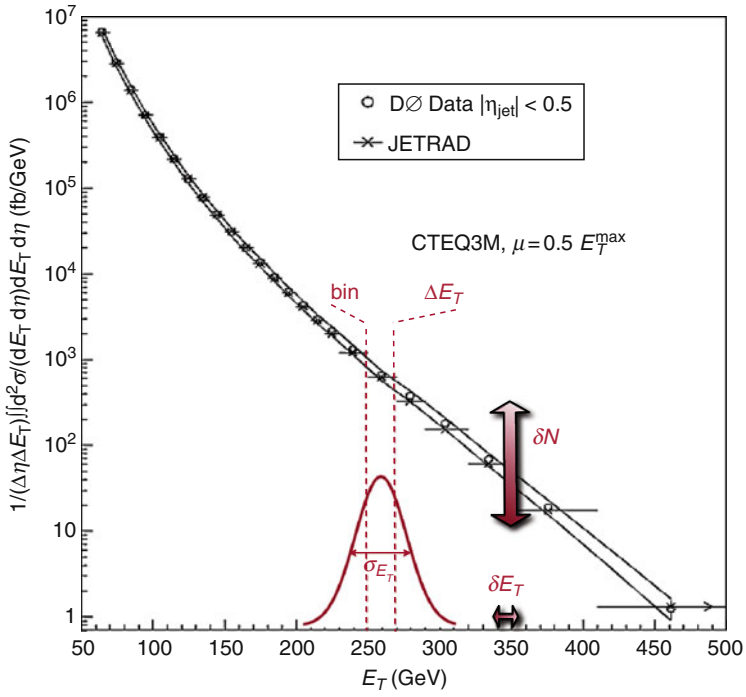


Fig. 7
Measurement of the inclusive jet cross section in proton–antiproton collisions, for $|\eta_{jet}| < 0.5$. Figure adapted from Blazey and Flaugher (1999)

steeper the slope of the spectrum is. On the right-hand side, we have the number of reconstructed jets N within a particular bin, divided by the bin sizes, the overall jet reconstruction efficiency ϵ, which also includes acceptance and trigger components as discussed earlier, and the integrated luminosity $\mathcal{L}$. As can be seen from *Fig. 7*, the bin size ΔE_T increases with increasing E_T, which ensures having a non-negligible number N of counted jets also in the tail of the distribution, thus avoiding large statistical fluctuations in that region.

However, as alluded to earlier, the apparently simple counting experiment is complicated further by the fact of having a steeply falling spectrum. A first issue is the precise knowledge of the absolute jet energy scale or, more generally, the precision with which the quantity on the x axis of a spectrum can be measured. Assume a differential cross section which follows a simple power law, i.e., $\mathrm{d}\sigma/\mathrm{d}E_T \propto E_T^{-n}$. Now it is obvious that any relative experimental uncertainty $\delta E_T/E_T$ on the absolute energy scale of a reconstructed jet immediately translates into an n-fold relative uncertainty on the counting result, $\delta N/N = n\,\delta E_T/E_T$. For example, with a power spectrum of $n = 6$, an energy reconstruction precision of 5% will translate into a cross-section uncertainty of 30%! This is the reason why quite some effort has to be investigated in understanding the overall energy calibration constants.

A second important aspect is the finite resolution on the reconstructed energy, which can lead to a distortion of the measured spectrum. For example, the resolution function can have a Gaussian shape,

$$R\left(E_T^{\text{meas}}, E_T^{\text{true}}\right) \propto \exp\left[-\frac{\left(E_T^{\text{meas}} - E_T^{\text{true}}\right)^2}{2\,\sigma_{E_T}^2}\right], \tag{3}$$

where E_T^{meas} $\left(E_T^{\text{true}}\right)$ represent the measured (true) transverse jet energies and the width of the Gaussian is given by σ_{E_T}. Then it follows that the counted number of jets, for a particular measured transverse energy E_T^{meas}, is given by the following convolution:

$$N\left(E_T^{\text{meas}}\right) = \int_0^\infty N\left(E_T^{\text{true}}\right)\, R\left(E_T^{\text{meas}}, E_T^{\text{true}}\right)\, \mathrm{d}E_T^{\text{true}}\,. \tag{4}$$

Since we are interested in $N\left(E_T^{\text{true}}\right)$, which is needed for a comparison to theoretical predictions, we have to solve this integral equation. More generally, this problem falls into the category of *distribution unfolding*, which is further complicated by the presence of statistical fluctuations and uncertainties on the measured quantities. An extensive discussion can be found in Cowan (1998). The distortion of the spectrum arises from the fact that a symmetric resolution function leads to a similar rate of entries fluctuating in and out of a given bin, but the number of bin entries varies strongly, even for adjacent bins. The issue can be somewhat mitigated by adjusting the bin size to the expected resolution in a given range of E_T. Nevertheless, considerable effort is needed for the understanding of the resolution function and the unfolding of the measured spectrum. Before closing this discussion, it is worth pointing out again that these analysis issues (absolute scale, resolution, unfolding) have to be addressed in every measurement, which attempts to determine a steeply falling spectrum of some physical quantity.

4 Computing and Software Aspects

After having learned about various aspects, which need to be taken into account when performing a data analysis, we finally discuss how such a process is implemented in practice, in particular in the context of large-scale experiments as those in modern Particle Physics. Here the sheer complexity of the experiment as such, of its hardware and software frameworks, as well as the large size and geographical spread of the involved experimental community require a well-coordinated and structured approach to the data-analysis problem. The basic components of this approach are depicted in ❯ *Fig. 8*. The reconstruction step, which translates the detector signals into higher-level objects (cf. ❯ Sect. 2), is not repeated by every single analysis, but rather carried out in the context of a coordinated and unique workflow. This means

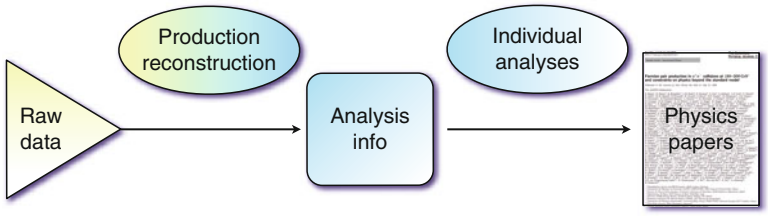


Fig. 8
Basic steps of a data-analysis chain, as implemented in the analysis frameworks of large-scale experiments

that a central software framework is put in place, which implements the raw data access, collects common algorithms such as track, calorimeter cluster, and jet finding, and organizes the storage and distribution of the reconstructed events. Since these issues are not specific to a particular analysis and all experimenters will ultimately rely on the same reconstructed information, it makes sense to organize this centrally. Also, having common algorithms makes it much easier to understand and debug their performance. At this analysis stage, the data access happens in a very coordinated manner, in the sense that only few individuals are responsible for running data-reconstruction jobs, on dedicated computer resources, at specified times, and using specified versions of the software.

The next step, which interprets the reconstructed data in terms of a specific Physics question, such as the measurement of an event rate or a spectrum, is carried out by individual analyzers. In large modern experiments, there can be hundreds and thousands of them, often organized in groups. Consequently, many different measurements are done in parallel and on different time scales. Here the data access is much more "chaotic," in the sense that periods of intense analysis activity (e.g., prior to important conferences) can be followed by more quiet times. Furthermore, a given individual or group will not access the whole data set, but only those fractions which are of relevance for their specific analysis. Typically, the required analysis code is developed on purpose. Ultimately, all these efforts are aimed at obtaining some new and unique Physics result, which can be published in a journal paper.

The challenge, which is faced by large experimental collaborations, is to build offline computing systems able to address these vastly different use cases and data-access patterns. The general approach to this challenge consists in partitioning the software frameworks and data/simulation production systems into individual components, as shown in *Fig. 9*. Such

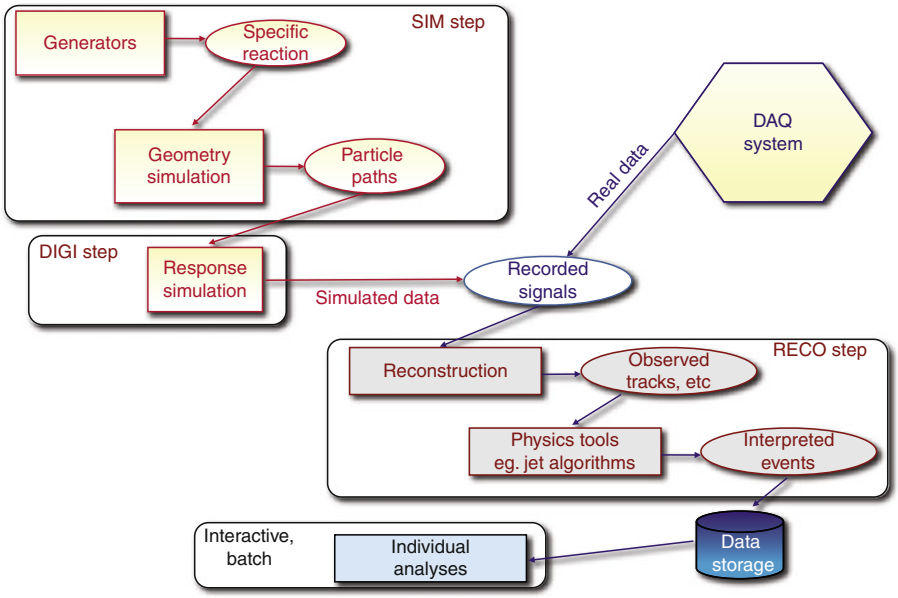


Fig. 9

Partitioning of software frameworks, which handle the simulation (SIM), digitization (DIGI), reconstruction (RECO), and analysis workflows of large experiments

a structure gives flexibility since the development, versioning, and deployment of the various components can occur asynchronously and on different time scales, by different experts. It considerably enhances the overall efficiency since a change or improvement in a given component not necessarily induces change requests for other components, which would require a repeat of the whole analysis chain. This is particularly true if there are dedicated data stores between the components. The overall management problem becomes tractable in such an approach, with its natural sharing of work load. Given the enormous size of modern experimental software frameworks, which contain millions of lines of code, and the related difficulty of building the executables, debugging, and maintaining these codes, it is obvious that a more monolithic approach would be a potential source of problems.

5 Conclusion

In this chapter, an attempt was made to describe the basic aspects of data analysis, which have to be addressed in modern Particle Physics experiments. Common recurring themes have been data reduction and data abstraction, in particular in the description of the steps, which lead from raw detector signals to higher-level physics objects. Some examples of reconstruction algorithms have been given, such as track, calorimeter cluster, and jet finding. Measurements from e^+e^- and hadron colliders have served as use cases for the discussion of further analysis aspects. Those comprised event selection techniques and the measurement of steeply falling spectra. Finally, the implementation of the analysis steps in modern software frameworks has been briefly addressed. Obviously, in the context of such a short review, it is not possible to go beyond the mere scratching of the tip of the iceberg. Therefore, it can only serve as a very first entry point into the large arena of Particle Physics data analysis. The author believes that the excitement lived in this arena will pay off for all the efforts spent on the way there.

Further Reading

This chapter is based on a set of lectures, which were given by the author and by B. Jacobsen in the context of the CERN Summer Student Lecture program, cf. http://cdsweb.cern.ch/collection/Summer%20Student%20Lectures.

A list of general textbooks on data analysis, statistical analysis, and data mining, which are mostly addressed at scientists and engineers, is given in Berthold and Hand (2003), Brandt (1998), Frühwirth et al. (2000), Lyman and Longnecker (2008), Nisbet et al. (1997), and Sivia and Skilling (2006).

As stated already in a few places in the text, many of the discussed concepts are central to data analysis also in other fields of research. For example, data analysis in medical applications is extensively discussed in Zupan et al. (1997), Lazar (2009), and Zaidi (2006), aspects relevant to biology can be found in Quinn and Keough (2002) and Roff (2006), and two examples from economic and social sciences are Koop (2009) and Elliott and Marsh (2009), respectively.

References

Berthold MR, Hand DJ (2003) Intelligent data analysis, 2nd edn. Springer, Berlin

Blazey GC, Flaugher BL (1999) Ann Rev Nucl Part Sci 49:633–685

Brandt S (1998) Data analysis: statistical and computational methods for scientists and engineers, 3rd edn. Springer, Berlin

Cowan G (1998) Statistical data analysis. Clarendon, Oxford

Dissertori G, Knowles IG, Schmelling M (2003) Quantum chromodynamics: high energy experiments and theory, 2nd edn. Clarendon, Oxford

Elliott J, Marsh C (2009) Exploring data: an introduction to data analysis for social scientists, 2nd edn. Polity, Oxford

Fishman G (1996) Monte Carlo: concepts, algorithms and applications. Springer, New York

Frühwirth R, Regler M, Bock RK, Grote H, Notz D (2000) Data analysis techniques for high-energy physics, 2nd edn. Cambridge University Press, Cambridge

Griffiths D (2008) Introduction to elementary particles, 2nd edn. Wiley-VCH, Weinheim

Koop G (2009) Analysis of economic data, 3rd edn. Wiley, New York

Lazar NA (2009) The statistical analysis of functional MRI data. Springer, New York

Lyman Ott R, Longnecker MT (2008) An introduction to statistical methods and data analysis, 6th edn. Duxbury, Belmont

Nisbet R, Elder IV J, Miner G (1997) Handbook of statistical analysis and data mining applications. Springer, New York

Perkins DH (2000) Introduction to high energy physics, 4th edn. Cambridge University Press, Cambridge

Quinn GP, Keough MJ (2002) Experimental design and data analysis for biologists. Cambridge University Press, Cambridge

Roff DA (2006) Introduction to computer-intensive methods of data analysis in biology. Cambridge University Press, Cambridge

Salam GP, preprint arXiv:0906.1833 [hep-ph], http://arxiv.org/abs/0906.1833

Sivia D, Skilling J (2006) Data analysis: A Bayesian tutorial, 2nd edn. Oxford University Press, Oxford

Wigmans R (2003) Calorimetry: energy measurement in particle physics. Clarendon, Oxford

Zaidi H (2006) Quantitative analysis in nuclear medicine imaging. Springer, New York

Zupan B, Keravnou E, Lavrac N (1997) Intelligent data analysis in medicine and pharmacology. Springer, Berlin

5 Statistics

Glen Cowan
Royal Holloway, University of London, Egham, Surrey, UK

1 Introduction . . . 104

2 Probability . . . 104

3 Random Variables . . . 105

4 Parameter Estimation . . . 108
4.1 Estimators for Mean, Variance, and Median . . . 108
4.2 The Method of Maximum Likelihood . . . 109
4.3 The Method of Least Squares . . . 110
4.4 The Bayesian Approach . . . 112

5 Statistical Tests . . . 114
5.1 Hypothesis Tests . . . 114
5.2 Significance Tests . . . 115
5.3 Bayesian Model Selection . . . 116

6 Intervals and Limits . . . 118
6.1 Bayesian Intervals . . . 118
6.2 Frequentist Confidence Intervals . . . 119
6.2.1 Profile Likelihood and Treatment of Nuisance Parameters . . . 120
6.2.2 Gaussian-Distributed Measurements . . . 120
6.2.3 Poisson or Binomial Data . . . 122

7 Conclusions . . . 123

References . . . 124

This chapter is largely excerpted and adapted from the reviews on Probability and Statistics in the Review of Particle Physics by the Particle Data Group (Nakamura et al. 2010).

C. Grupen, I. Buvat (eds.), *Handbook of Particle Detection and Imaging*,
DOI 10.1007/978-3-642-13271-1_5, © Springer-Verlag Berlin Heidelberg 2012

Abstract: In experimental particle physics as well as in many other fields, it has become increasingly important to analyze data in a manner that extracts the maximum information and takes into account all of the known uncertainties. This article reviews the most important statistical methods used to carry out this task. It begins with an overview of probability, as this forms the basis for quantifying uncertainty. The statistical methods considered include the general framework of statistical tests and parameter estimation, including methods for constructing intervals such as upper limits. Both frequentist and Bayesian approaches are described.

1 Introduction

This article presents an overview of statistical methods used in high-energy physics (HEP). In statistics, we are interested in using a sample of data to make inferences about a probabilistic model, for example, to assess the model's validity or to determine the values of its parameters. There are two main approaches to statistical inference, which we may call frequentist and Bayesian. These differ in their interpretation of probability. A review of probability and random variables is given in ❯ Sects. 2 and ❯ 3.

The most important statistical tools within the frequentist framework are parameter estimation, covered in ❯ Sect. 4, and statistical tests, discussed in ❯ Sect. 5. Frequentist confidence intervals, which are constructed so as to cover the true value of a parameter with a specified probability, are treated in ❯ Sect. 6.2. Bayesian methods for interval estimation are discussed in ❯ Sect. 6.1. These intervals quantify the degree of belief with which a parameter lies within a stated range. Intervals are discussed for the important cases of Gaussian-, binomial-, and Poisson-distributed data.

2 Probability

An abstract definition of *probability* can be given by considering a set S, called the sample space, and possible subsets $A, B, \ldots$, the interpretation of which is left open for now. The probability P is a real-valued function defined by the following axioms due to Kolmogorov (Kolmogorov 1993):

1. For every subset A in S, $P(A) \geq 0$.
2. For disjoint subsets (i.e., $A \cap B = \emptyset$), $P(A \cup B) = P(A) + P(B)$.
3. $P(S) = 1$.

In addition, one defines the conditional probability $P(A|B)$ (read P of A given B) as

$$P(A|B) = \frac{P(A \cap B)}{P(B)}. \tag{1}$$

From this definition and using the fact that $A \cap B$ and $B \cap A$ are the same, one obtains *Bayes' theorem*,

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}. \tag{2}$$

From the three axioms of probability and the definition of conditional probability, one obtains the *law of total probability*,

$$P(B) = \sum_i P(B|A_i)P(A_i), \tag{3}$$

for any subset B and for disjoint A_i with $\bigcup_i A_i = S$. This can be combined with Bayes' theorem (Eq. 2) to give

$$P(A|B) = \frac{P(B|A)P(A)}{\sum_i P(B|A_i)P(A_i)}, \tag{4}$$

where the subset A could, for example, be one of the A_i.

In the most commonly used interpretation of probability used in particle physics, the elements of the sample space correspond to outcomes of a repeatable experiment. The probability $P(A)$ is assigned a value equal to the limiting frequency of occurrence of A. This interpretation forms the basis of *frequentist statistics*.

The elements of the sample space can also be interpreted as *hypotheses*, i.e., statements that are either true or false, such as "The mass of the W boson lies between 80.3 and 80.5 GeV." In the frequency interpretation, such statements are either always or never true, i.e., the corresponding probabilities would be 0 or 1. Using *subjective probability*, however, $P(A)$ is interpreted as the degree of belief that the hypothesis A is true. Subjective probability is used in *Bayesian* (as opposed to frequentist) statistics. Bayes' theorem can be written

$$P(\text{theory}|\text{data}) \propto P(\text{data}|\text{theory})P(\text{theory}), \tag{5}$$

where "theory" represents some hypothesis and "data" is the outcome of the experiment. Here, $P(\text{theory})$ is the *prior* probability for the theory, which reflects the experimenter's degree of belief before carrying out the measurement, and $P(\text{data}|\text{theory})$ is the probability to have gotten the data actually obtained, given the theory, which is also called the *likelihood*.

Bayesian statistics provides no fundamental rule for obtaining the prior probability; this is necessarily subjective and may depend on previous measurements, theoretical prejudices, etc. Once this has been specified, however, Eq. 5 tells how the probability for the theory must be modified in the light of the new data to give the *posterior* probability, $P(\text{theory}|\text{data})$. As Eq. 5 is stated as a proportionality, the probability must be normalized by summing (or integrating) over all possible hypotheses.

3 Random Variables

A *random variable* is a numerical characteristic assigned to an element of the sample space. In the frequency interpretation of probability, it corresponds to an outcome of a repeatable experiment. Let x be a possible outcome of an observation. If x can take on any value from a continuous range, we write $f(x;\theta)\,\mathrm{d}x$ as the probability that the measurement's outcome lies between x and $x + \mathrm{d}x$. The function $f(x;\theta)$ is called the *probability density function* (p.d.f.), which may depend on one or more parameters θ. If x can take on only discrete values (e.g., the nonnegative integers), then $f(x;\theta)$ is itself a probability. The p.d.f. is always normalized to unit area (unit sum, if discrete). Both x and θ may have multiple components and are then often written as vectors.

The *cumulative distribution function* $F(x)$ is the probability for the random variable to be observed less than or equal to x:

$$F(x) = \int_{-\infty}^{x} f(x')\,\mathrm{d}x'. \tag{6}$$

Here and below, if x is discrete-valued, the integral is replaced by a sum. The endpoint x is expressly included in the integral or sum.

Any function of random variables is itself a random variable, with (in general) a different p.d.f. The *expectation value* or *mean* of any function $u(x)$ is

$$E[u(x)] = \int_{-\infty}^{\infty} u(x)\,f(x)\,\mathrm{d}x, \tag{7}$$

assuming the integral is finite. If $u(x)$ and $v(x)$ are any two functions of x, then $E[u+v] = E[u] + E[v]$. For constant values c and k, one finds $E[cu+k] = cE[u] + k$.

The nth moment of a random variable is

$$\alpha_n \equiv E[x^n] = \int_{-\infty}^{\infty} x^n f(x)\,\mathrm{d}x, \tag{8}$$

and the nth central moment of x (or moment about the mean, α_1) is

$$m_n \equiv E[(x-\alpha_1)^n] = \int_{-\infty}^{\infty} (x-\alpha_1)^n f(x)\,\mathrm{d}x. \tag{9}$$

The most commonly used moments are the mean μ and variance σ^2:

$$\mu \equiv \alpha_1, \tag{10}$$

$$\sigma^2 \equiv V[x] \equiv m_2 = \alpha_2 - \mu^2. \tag{11}$$

The mean is the location of the "center of mass" of the p.d.f., and the variance is a measure of the square of its width. Note that $V[cx+k] = c^2 V[x]$. It is often convenient to use the *standard deviation* of x, σ, defined as the square root of the variance.

Besides the mean, another useful indicator of the "middle" of the probability distribution is the *median*, x_{med}, defined by $F(x_{\mathrm{med}}) = 1/2$, i.e., half the probability lies above and half lies below x_{med}. (More rigorously, x_{med} is a median if $P(x \geq x_{\mathrm{med}}) \geq 1/2$ and $P(x \leq x_{\mathrm{med}}) \geq 1/2$. If only one value exists, it is called "*the* median.")

Let x and y be two random variables with a *joint* p.d.f. $f(x,y)$. The *marginal* p.d.f. of x (the distribution of x with y unobserved) is

$$f_1(x) = \int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}y, \tag{12}$$

and similarly for the marginal p.d.f. $f_2(y)$. The *conditional* p.d.f. of y given fixed x (with $f_1(x) \neq 0$) is defined by $f_3(y|x) = f(x,y)/f_1(x)$, and similarly $f_4(x|y) = f(x,y)/f_2(y)$. From these, we immediately obtain Bayes' theorem (see ❯ Eqs. 2 and ❯ 4),

$$f_4(x|y) = \frac{f_3(y|x)f_1(x)}{f_2(y)} = \frac{f_3(y|x)f_1(x)}{\int f_3(y|x')f_1(x')\,\mathrm{d}x'}. \tag{13}$$

The mean of x is

$$\mu_x = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x\,f(x,y)\,\mathrm{d}x\,\mathrm{d}y = \int_{-\infty}^{\infty} x\,f_1(x)\,\mathrm{d}x, \tag{14}$$

and similarly for y. The *covariance* of x and y is

$$\mathrm{cov}[x,y] = E[(x-\mu_x)(y-\mu_y)] = E[xy] - \mu_x\mu_y. \tag{15}$$

A dimensionless measure of the covariance of x and y is given by the *correlation coefficient*,

$$\rho_{xy} = \mathrm{cov}[x, y]/\sigma_x\sigma_y, \tag{16}$$

where σ_x and σ_y are the standard deviations of x and y. It can be shown that $-1 \le \rho_{xy} \le 1$.

Two random variables x and y are *independent* if and only if

$$f(x, y) = f_1(x)f_2(y). \tag{17}$$

If x and y are independent, then $\rho_{xy} = 0$; the converse is not necessarily true. If x and y are independent, then for any functions $u(x)$ and $v(y)$ one has $E[u(x)v(y)] = E[u(x)]E[v(y)]$, and also one finds $V[x + y] = V[x] + V[y]$. If x and y are not independent, $V[x + y] = V[x] + V[y] + 2\,\mathrm{cov}[x, y]$, and $E[uv]$ does not necessarily factorize.

Consider a set of n continuous random variables $\mathbf{x} = (x_1, \ldots, x_n)$ with joint p.d.f. $f(\mathbf{x})$, and a set of n new variables $\mathbf{y} = (y_1, \ldots, y_n)$ related to $\mathbf{x}$ by means of a function $\mathbf{y}(\mathbf{x})$ that is one-to-one, i.e., the inverse $\mathbf{x}(\mathbf{y})$ exists. The joint p.d.f. for $\mathbf{y}$ is given by

$$g(\mathbf{y}) = f(\mathbf{x}(\mathbf{y}))\,|J|, \tag{18}$$

where $|J|$ is the absolute value of the determinant of the square matrix $J_{ij} = \partial x_i/\partial y_j$ (the Jacobian determinant). If the transformation from $\mathbf{x}$ to $\mathbf{y}$ is not one-to-one, the $\mathbf{x}$ space must be broken into regions where the function $\mathbf{y}(\mathbf{x})$ can be inverted and the contributions to $g(\mathbf{y})$ from each region summed.

Several probability functions and p.d.f.s along with their properites are given in *Table 1*.

Table 1

Probability distributions, their mean values, and variances

	Distribution	Mean	Variance
Binomial	$f(r; N, p) = \frac{N!}{r!(N-r)!}\, p^r(1-p)^{N-r}$	Np	$Np(1-p)$
Poisson	$f(n; \nu) = \frac{\nu^n \mathrm{e}^{-\nu}}{n!}$	ν	ν
Uniform	$f(x; a, b) = \begin{cases} 1/(b-a) & a \le x \le b, \\ 0 & \text{otherwise.} \end{cases}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Gaussian	$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$	μ	σ^2
Multivariate Gaussian	$f(x; \boldsymbol{\mu}, V) = \frac{1}{(2\pi)^{n/2}\sqrt{\lvert V\rvert}} \times \exp\left[-\frac{1}{2}(x-\boldsymbol{\mu})^T V^{-1}(x-\boldsymbol{\mu})\right]$	$\boldsymbol{\mu}$	V_{ij}
Exponential	$f(x; \mu) = \frac{1}{\mu}\mathrm{e}^{-x/\mu}$	μ	μ^2
Chi-square	$f(z; n) = \frac{z^{n/2-1}\mathrm{e}^{-z/2}}{2^{n/2}\Gamma(n/2)}$	n	$2n$
Student's t	$f(t; n) = \frac{1}{\sqrt{n\pi}} \frac{\Gamma[(n+1)/2]}{\Gamma(n/2)} \times \left(1 + \frac{t^2}{n}\right)^{-(n+1)/2}$	0 $(n > 1)$	$n/(n-2)$ $(n > 2)$
Gamma	$f(x; \lambda, k) = \frac{x^{k-1}\lambda^k \mathrm{e}^{-\lambda x}}{\Gamma(k)}$	k/λ	k/λ^2
Beta	$f(x; \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

4 Parameter Estimation

Here, we review *point estimation* of parameters, first with an overview of the frequentist approach and its two most important methods, maximum likelihood and least squares, treated in ❯ Sects. 4.2 and ❯ 4.3. The Bayesian approach is outlined in ❯ Sect. 4.4.

An *estimator* $\hat{\theta}$ (written with a hat) is a function of the data whose value, the *estimate*, is intended as a meaningful guess for the value of the parameter θ. There is no fundamental rule dictating how an estimator must be constructed. One tries, therefore, to choose that estimator which has the best properties. The most important of these are (a) *consistency*, (b) *bias*, (c) *efficiency*, and (d) *robustness*.

(a) An estimator is said to be *consistent* if the estimate $\hat{\theta}$ converges to the true value θ as the amount of data increases. This property is so important that it is possessed by all commonly used estimators.

(b) The *bias*, $b = E[\,\hat{\theta}\,] - \theta$, is the difference between the expectation value of the estimator and the true value of the parameter. The expectation value is taken over a hypothetical set of similar experiments in which $\hat{\theta}$ is constructed in the same way. When $b = 0$, the estimator is said to be unbiased. The bias depends on the chosen metric, i.e., if $\hat{\theta}$ is an unbiased estimator of θ, then $\hat{\theta}^2$ is not in general an unbiased estimator for θ^2.

(c) *Efficiency* is the inverse of the ratio of the variance $V[\,\hat{\theta}\,]$ to the minimum possible variance for any estimator of θ. Under rather general conditions, the minimum variance for a single parameter θ is given by the Rao–Cramér–Frechet bound,

$$\sigma^2_{\text{min}} = -\left(1 + \frac{\partial b}{\partial \theta}\right)^2 \Big/ E\left[\frac{\partial^2 \ln L(\theta)}{\partial \theta^2}\right], \tag{19}$$

where is $L(\theta)$ is the *likelihood function* (see below). The *mean-squared error*,

$$\text{MSE} = E[(\hat{\theta} - \theta)^2] = V[\hat{\theta}] + b^2, \tag{20}$$

is a measure of an estimator's quality which combines the uncertainties due to bias and variance.

(d) *Robustness* is the property of being insensitive to departures from assumptions in the p.d.f., for example, owing to uncertainties in the distribution's tails.

Simultaneously optimizing for all the measures of estimator quality described above can lead to conflicting requirements. For example, there is in general a trade-off between bias and variance. For some common estimators, the properties above are known exactly. More generally, it is possible to evaluate them by Monte Carlo simulation. Note that they will often depend on the unknown θ.

4.1 Estimators for Mean, Variance, and Median

Suppose we have a set of N independent measurements, x_i, assumed to be unbiased measurements of the same unknown quantity μ with a common, but unknown, variance σ^2. Then

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i, \tag{21}$$

$$\widehat{\sigma^2} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \hat{\mu})^2 \tag{22}$$

are unbiased estimators of μ and σ^2. The variance of $\hat{\mu}$ is σ^2/N and the variance of $\widehat{\sigma^2}$ is

$$V\left[\widehat{\sigma^2}\right] = \frac{1}{N}\left(m_4 - \frac{N-3}{N-1}\sigma^4\right), \tag{23}$$

where m_4 is the fourth central moment of x. For Gaussian-distributed x_i, this becomes $2\sigma^4/(N-1)$ for any $N \geq 2$, and for large N, the standard deviation of $\hat{\sigma}$ (the "error of the error") is $\sigma/\sqrt{2N}$. Again, if the x_i are Gaussian, $\hat{\mu}$ is an efficient estimator for μ, and the estimators $\hat{\mu}$ and $\widehat{\sigma^2}$ are uncorrelated. Otherwise the arithmetic mean (Eq. 21) is not necessarily the most efficient estimator; this is discussed further in Sect. 8.7 of James (2007).

4.2 The Method of Maximum Likelihood

Suppose we have a set of N measured quantities $\mathbf{x} = (x_1, \ldots, x_N)$ described by a joint p.d.f. $f(\mathbf{x}; \boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ is set of n parameters whose values are unknown. The *likelihood function* is given by the p.d.f. evaluated with the data $\mathbf{x}$, but viewed as a function of the parameters, i.e., $L(\boldsymbol{\theta}) = f(\mathbf{x}; \boldsymbol{\theta})$. If the measurements x_i are statistically independent and each follow the p.d.f. $f(x; \boldsymbol{\theta})$, then the joint p.d.f. for $\mathbf{x}$ factorizes and the likelihood function is

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{N} f(x_i; \boldsymbol{\theta}). \tag{24}$$

The method of maximum likelihood takes the estimators $\hat{\boldsymbol{\theta}}$ to be those values of $\boldsymbol{\theta}$ that maximize $L(\boldsymbol{\theta})$.

Note that the likelihood function is *not* a p.d.f. for the parameters $\boldsymbol{\theta}$; in frequentist statistics this is not defined. In Bayesian statistics, one can obtain from the likelihood the posterior p.d.f. for $\boldsymbol{\theta}$, but this requires multiplying by a prior p.d.f. (see Sect. 6.1).

It is usually easier to work with $\ln L$, and since both are maximized for the same parameter values $\boldsymbol{\theta}$, the maximum likelihood (ML) estimators can be found by solving the *likelihood equations*

$$\frac{\partial \ln L}{\partial \theta_i} = 0, \qquad i = 1, \ldots, n. \tag{25}$$

Often the solution must be found numerically. Maximum-likelihood estimators are important because they are approximately unbiased and efficient for large data samples under quite general conditions, and the method has a wide range of applicability.

In evaluating the likelihood function, it is important that any normalization factors in the p.d.f. that involve $\boldsymbol{\theta}$ be included. However, we will only be interested in the maximum of L and in ratios of L at different values of the parameters; hence, any multiplicative factors that do not involve the parameters that we want to estimate may be dropped, including factors that depend on the data but not on $\boldsymbol{\theta}$.

Under a one-to-one change of parameters from $\boldsymbol{\theta}$ to $\boldsymbol{\eta}$, the ML estimators $\hat{\boldsymbol{\theta}}$ transform to $\boldsymbol{\eta}(\hat{\boldsymbol{\theta}})$. That is, the ML solution is invariant under change of parameters. However, other properties of ML estimators, in particular the bias, are not invariant under change of parameters.

Under requirements usually satisfied in practical analyses and for a sufficiently large data sample, the inverse V^{-1} of the covariance matrix $V_{ij} = \mathrm{cov}[\hat{\theta}_i, \hat{\theta}_j]$ for a set of ML estimators can be estimated by using

$$(\widehat{V}^{-1})_{ij} = -\left.\frac{\partial^2 \ln L}{\partial \theta_i \partial \theta_j}\right|_{\hat{\boldsymbol{\theta}}}. \tag{26}$$

In the large-sample limit (or in a linear model with Gaussian errors), L has a Gaussian form and $\ln L$ is (hyper)parabolic. In this case, it can be seen that a numerically equivalent way of determining s-standard-deviation errors is from the contour given by the $\boldsymbol{\theta}'$ such that

$$\ln L(\boldsymbol{\theta}') = \ln L_{\max} - s^2/2, \tag{27}$$

where $\ln L_{\max}$ is the value of $\ln L$ at the solution point (compare with Eq. 65). The extreme limits of this contour on the θ_i axis give an approximate s-standard-deviation confidence interval for θ_i (see Sect. 6.2).

In the case where the size n of the data sample $x_1, \ldots, x_n$ is small, the unbinned maximum-likelihood method, i.e., use of Eq. 24, is preferred since binning can only result in a loss of information, and hence larger statistical errors for the parameter estimates. The sample size n can be regarded as fixed, or the user can choose to treat it as a Poisson-distributed variable; this latter option is sometimes called "extended maximum likelihood" (see, e.g., Lyons 1986; Barlow 1990; Cowan 1998).

If the sample is large, it can be convenient to bin the values in a histogram, so that one obtains a vector of data $\mathbf{n} = (n_1, \ldots, n_N)$ with expectation values $\boldsymbol{\nu} = E[\mathbf{n}]$ and probabilities $f(\mathbf{n}; \boldsymbol{\nu})$. Then one may maximize the likelihood function based on the contents of the bins (so i labels bins). This is equivalent to maximizing the likelihood ratio $\lambda(\boldsymbol{\theta}) = f(\mathbf{n}; \boldsymbol{\nu}(\boldsymbol{\theta}))/f(\mathbf{n}; \mathbf{n})$ or to minimizing the equivalent quantity $-2\ln\lambda(\boldsymbol{\theta})$. For independent Poisson-distributed n_i this is (Cousins and Baker 1984)

$$-2\ln\lambda(\boldsymbol{\theta}) = 2\sum_{i=1}^{N}\left[\nu_i(\boldsymbol{\theta}) - n_i + n_i \ln\frac{n_i}{\nu_i(\boldsymbol{\theta})}\right], \tag{28}$$

where for bins with $n_i = 0$, the last term in Eq. 28 is zero. The expression Eq. 28 without the terms $\nu_i - n_i$ also gives $-2\ln\lambda(\boldsymbol{\theta})$ for multinomially distributed n_i, i.e., when the total number of entries is regarded as fixed. In the limit of zero bin width, maximizing Eq. 28 is equivalent to maximizing the unbinned likelihood function (Eq. 24).

A benefit of binning is that it allows for a goodness-of-fit test (see Sect. 5.2). According to Wilks' theorem, for sufficiently large ν_i and providing certain regularity conditions are met, the minimum of $-2\ln\lambda$ as defined by Eq. 28 follows a chi-square distribution (see, e.g., Stuart et al. 1999). If there are N bins and m fitted parameters, then the number of degrees of freedom for the chi-square distribution is $N - m$ if the data are treated as Poisson-distributed, and $N - m - 1$ if the n_i are multinomially distributed.

Suppose the n_i are Poisson-distributed and the overall normalization $\nu_{\text{tot}} = \sum_i \nu_i$ is taken as an adjustable parameter, so that $\nu_i = \nu_{\text{tot}} p_i(\boldsymbol{\theta})$, where the probability to be in the ith bin, $p_i(\boldsymbol{\theta})$, does not depend on ν_{tot}. Then by minimizing Eq. 28, one obtains that the area under the fitted function is equal to the sum of the histogram contents, i.e., $\sum_i \nu_i = \sum_i n_i$. This is not the case for parameter estimation methods based on a least-squares procedure with traditional weights (see, e.g., Cowan 1998).

4.3 The Method of Least Squares

The *method of least squares* (LS) coincides with the method of maximum likelihood in the following special case. Consider a set of N independent measurements y_i at known points x_i. The measurement y_i is assumed to be Gaussian distributed with mean $\mu(x_i; \boldsymbol{\theta})$ and known

variance σ_i^2. The goal is to construct estimators for the unknown parameters $\boldsymbol{\theta}$. The likelihood function contains the sum of squares

$$\chi^2(\boldsymbol{\theta}) = -2\ln L(\boldsymbol{\theta}) + \text{ constant } = \sum_{i=1}^{N} \frac{(y_i - \mu(x_i;\boldsymbol{\theta}))^2}{\sigma_i^2}. \tag{29}$$

The set of parameters $\boldsymbol{\theta}$ which maximize L is the same as those which minimize χ^2.

The minimum of Eq. 29 defines the least-squares estimators $\hat{\boldsymbol{\theta}}$ for the more general case where the y_i are not Gaussian distributed as long as they are independent. If they are not independent but rather have a covariance matrix $V_{ij} = \text{cov}[y_i, y_j]$, then the LS estimators are determined by the minimum of

$$\chi^2(\boldsymbol{\theta}) = (\mathbf{y} - \boldsymbol{\mu}(\boldsymbol{\theta}))^T V^{-1} (\mathbf{y} - \boldsymbol{\mu}(\boldsymbol{\theta})), \tag{30}$$

where $\mathbf{y} = (y_1, \ldots, y_N)$ is the vector of measurements, $\boldsymbol{\mu}(\boldsymbol{\theta})$ is the corresponding vector of predicted values (understood as a column vector in Eq. 30), and the superscript T denotes transposed (i.e., row) vector.

In many practical cases, one further restricts the problem to the situation where $\mu(x_i;\boldsymbol{\theta})$ is a linear function of the parameters, i.e.,

$$\mu(x_i;\boldsymbol{\theta}) = \sum_{j=1}^{m} \theta_j h_j(x_i). \tag{31}$$

Here, the $h_j(x)$ are m linearly independent functions, for example, $1, x, x^2, \ldots, x^{m-1}$, or Legendre polynomials. We require $m < N$, and at least m of the x_i must be distinct.

Minimizing χ^2 in this case with m parameters reduces to solving a system of m linear equations. Defining $H_{ij} = h_j(x_i)$ and minimizing χ^2 by setting its derivatives with respect to the θ_i equal to zero gives the LS estimators

$$\hat{\boldsymbol{\theta}} = (H^T V^{-1} H)^{-1} H^T V^{-1} \mathbf{y} \equiv D\mathbf{y}. \tag{32}$$

The covariance matrix for the estimators $U_{ij} = \text{cov}[\hat{\theta}_i, \hat{\theta}_j]$ is given by

$$U = DVD^T = (H^T V^{-1} H)^{-1}, \tag{33}$$

or equivalently, its inverse U^{-1} can be found from

$$(U^{-1})_{ij} = \frac{1}{2} \frac{\partial^2 \chi^2}{\partial\theta_i \partial\theta_j}\bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = \sum_{k,l=1}^{N} h_i(x_k)(V^{-1})_{kl} h_j(x_l). \tag{34}$$

The LS estimators can also be found from the expression

$$\hat{\boldsymbol{\theta}} = U\mathbf{g}, \tag{35}$$

where the vector $\mathbf{g}$ is defined by

$$g_i = \sum_{j,k=1}^{N} y_j h_i(x_k)(V^{-1})_{jk}. \tag{36}$$

For the case of uncorrelated y_i, for example, one can use Eq. 35 with

$$(U^{-1})_{ij} = \sum_{k=1}^{N} \frac{h_i(x_k) h_j(x_k)}{\sigma_k^2}, \tag{37}$$

$$g_i = \sum_{k=1}^{N} \frac{y_k h_i(x_k)}{\sigma_k^2}. \tag{38}$$

Expanding $\chi^2(\boldsymbol{\theta})$ about $\hat{\boldsymbol{\theta}}$, one finds that the contour in parameter space defined by

$$\chi^2(\boldsymbol{\theta}) = \chi^2(\hat{\boldsymbol{\theta}}) + 1 = \chi^2_{\min} + 1 \tag{39}$$

has tangent planes located at approximately plus-or-minus-one standard deviation $\sigma_{\hat{\theta}}$ from the LS estimates $\hat{\boldsymbol{\theta}}$.

In constructing the quantity $\chi^2(\boldsymbol{\theta})$, one requires the variances or, in the case of correlated measurements, the covariance matrix. Often, these quantities are not known a priori and must be estimated from the data; an important example is where the measured value y_i represents a counted number of events in the bin of a histogram. If, for example, y_i represents a Poisson variable, for which the variance is equal to the mean, then one can either estimate the variance from the predicted value, $\mu(x_i; \boldsymbol{\theta})$, or from the observed number itself, y_i. In the first option, the variances become functions of the fitted parameters, which may lead to calculational difficulties. The second option can be undefined if y_i is zero, and in both cases for small y_i, the variance will be poorly estimated. In either case, one should constrain the normalization of the fitted curve to the correct value, i.e., one should determine the area under the fitted curve directly from the number of entries in the histogram (see Cowan 1998, Sect. 7.4). A further alternative is to use the method of maximum likelihood; for binned data this can be done by minimizing ❯ Eq. 28.

As the minimum value of the χ^2 represents the level of agreement between the measurements and the fitted function, it can be used for assessing the goodness-of-fit; this is discussed further in ❯ Sect. 5.2.

4.4 The Bayesian Approach

In the frequentist methods discussed above, probability is associated only with data, not with the value of a parameter. This is no longer the case in Bayesian statistics, however, which we introduce in this section. Bayesian methods are considered further in ❯ Sect. 6.1 for interval estimation and in ❯ Sect. 5.3 for model selection. For general introductions to Bayesian statistics see, for example, O'Hagan and Forster 2004; Sivia and Skilling 2006; Gregory 2005; Bernardo and Smith 2000.

Suppose the outcome of an experiment is characterized by a vector of data $\mathbf{x}$, whose probability distribution depends on an unknown parameter (or parameters) $\boldsymbol{\theta}$ that we wish to determine. In Bayesian statistics, all knowledge about $\boldsymbol{\theta}$ is summarized by the posterior p.d.f. $p(\boldsymbol{\theta}|\mathbf{x})$, which gives the degree of belief for $\boldsymbol{\theta}$ to take on values in a certain region given the data $\mathbf{x}$. It is obtained by using Bayes' theorem,

$$p(\boldsymbol{\theta}|\mathbf{x}) = \frac{L(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int L(\mathbf{x}|\boldsymbol{\theta}')\pi(\boldsymbol{\theta}')\,\mathrm{d}\boldsymbol{\theta}'}, \tag{40}$$

where $L(\mathbf{x}|\boldsymbol{\theta})$ is the likelihood function, i.e., the joint p.d.f. for the data viewed as a function of $\boldsymbol{\theta}$, evaluated with the data actually obtained in the experiment, and $\pi(\boldsymbol{\theta})$ is the prior p.d.f. for $\boldsymbol{\theta}$. Note that the denominator in ❯ Eq. 40 serves to normalize the posterior p.d.f. to unity.

As it can be difficult to report the full posterior p.d.f. $p(\boldsymbol{\theta}|\mathbf{x})$, one would usually summarize it with statistics such as the mean (or median) and the covariance matrix. In addition, one may construct intervals with a given probability content, as is discussed in ❯ Sect. 6.1 on Bayesian interval estimation.

Bayesian statistics supplies no unique rule for determining the prior $\pi(\boldsymbol{\theta})$; in a subjective Bayesian analysis this reflects the experimenter's degree of belief (or state of knowledge) about $\boldsymbol{\theta}$ before the measurement was carried out. For the result to be of value to the broader community,

whose members may not share these beliefs, it is important to carry out a sensitivity analysis, i.e., to show how the result changes under a reasonable variation of the prior probabilities.

One might like to construct $\pi(\boldsymbol{\theta})$ to represent complete ignorance about the parameters by setting it equal to a constant. A problem here is that if the prior p.d.f. is flat in $\boldsymbol{\theta}$, then it is not flat for a nonlinear function of $\boldsymbol{\theta}$, and so a different parametrization of the problem would lead in general to a non-equivalent posterior p.d.f.

For the special case of a constant prior, one can see from Bayes' theorem (Eq. 40) that the posterior is proportional to the likelihood, and therefore the mode (peak position) of the posterior is equal to the ML estimator. The posterior mode, however, will change in general upon a transformation of parameter. A summary statistic other than the mode may be used as the Bayesian estimator, such as the median, which is invariant under parameter transformation. But this will not in general coincide with the ML estimator.

The difficult and subjective nature of encoding personal knowledge into priors has led to what is called *objective Bayesian statistics*, where prior probabilities are based not on an actual degree of belief but rather derived from formal rules. These give, for example, priors which are invariant under a transformation of parameters or which result in a maximum gain in information for a given set of measurements. For an extensive review see, for example, Kass and Wasserman 1996.

An important procedure for deriving objective priors is due to Jeffreys. According to *Jeffreys' rule*, one takes the prior as

$$\pi(\boldsymbol{\theta}) \propto \sqrt{\det(I(\boldsymbol{\theta}))}, \tag{41}$$

where

$$I_{ij}(\boldsymbol{\theta}) = -E\left[\frac{\partial^2 \ln L(\mathbf{x}|\boldsymbol{\theta})}{\partial\theta_i \partial\theta_j}\right] = -\int \frac{\partial^2 \ln L(\mathbf{x}|\boldsymbol{\theta})}{\partial\theta_i \partial\theta_j} L(\mathbf{x}|\boldsymbol{\theta})\, \mathrm{d}\mathbf{x} \tag{42}$$

is the *Fisher information matrix*. One can show that the Jeffreys prior leads to inference that is invariant under a transformation of parameters.

Neither the constant nor $1/\sqrt{\mu}$ priors can be normalized to unit area and are said to be *improper*. This can be allowed because the prior always appears multiplied by the likelihood function, and if the likelihood falls off sufficiently quickly, then one may have a normalizable posterior density.

Bayesian statistics provides a framework for incorporating systematic uncertainties into a result. Suppose, for example, that a model depends not only on parameters of interest $\boldsymbol{\theta}$, but on *nuisance parameters* $\boldsymbol{\nu}$, whose values are known with some limited accuracy. For a single nuisance parameter ν, for example, one might have a p.d.f. centered about its nominal value with a certain standard deviation σ_ν. Often a Gaussian p.d.f. provides a reasonable model for one's degree of belief about a nuisance parameter; in other cases, more complicated shapes may be appropriate. If, for example, the parameter represents a nonnegative quantity, then a log-normal or gamma p.d.f. can be a more natural choice than a Gaussian truncated at zero. The likelihood function, prior, and posterior p.d.f.s then all depend on both $\boldsymbol{\theta}$ and $\boldsymbol{\nu}$, and are related by Bayes' theorem, as usual. One can obtain the posterior p.d.f. for $\boldsymbol{\theta}$ alone by integrating over the nuisance parameters, i.e.,

$$p(\boldsymbol{\theta}|\mathbf{x}) = \int p(\boldsymbol{\theta}, \boldsymbol{\nu}|\mathbf{x})\, \mathrm{d}\boldsymbol{\nu}. \tag{43}$$

Such integrals can often not be carried out in closed form, and if the number of nuisance parameters is large, then they can be difficult to compute with standard Monte Carlo methods. *Markov Chain Monte Carlo* (MCMC) is often used for computing integrals of this type.

5 Statistical Tests

In addition to estimating parameters, one often wants to assess the validity of certain statements concerning the data's underlying distribution. Frequentist *Hypothesis tests*, described in Sect. 5.1, provide a rule for accepting or rejecting hypotheses depending on the outcome of a measurement. In *significance tests*, covered in Sect. 5.2, one gives the probability to obtain a level of incompatibility with a certain hypothesis that is greater than or equal to the level observed with the actual data. In the Bayesian approach, the corresponding procedure is referred to as model selection, which is based fundamentally on the probabilities of competing hypotheses. In Sect. 5.3, we describe a related construct called the Bayes factor, which can be used to quantify the degree to which the data prefer one or another hypothesis.

5.1 Hypothesis Tests

Consider an experiment whose outcome is characterized by a vector of data $\mathbf{x}$. A *hypothesis* is a statement about the distribution of $\mathbf{x}$. It could, for example, define completely the p.d.f. for the data (a simple hypothesis), or it could specify only the functional form of the p.d.f., with the values of one or more parameters left open (a composite hypothesis).

A *statistical test* is a rule that states for which values of $\mathbf{x}$ a given hypothesis (often called the null hypothesis, H_0) should be rejected in favour of its alternative H_1. This is done by defining a region of $\mathbf{x}$ space called the critical region; if the outcome of the experiment lands in this region, H_0 is rejected, otherwise it is accepted.

Rejecting H_0, if it is true, is called an error of the first kind. The probability for this to occur is called the *size* or *significance level* of the test, α, which is chosen to be equal to some prespecified value. It can also happen that H_0 is false and the true hypothesis is the alternative, H_1. If H_0 is accepted in such a case, this is called an error of the second kind, which will have some probability β. The quantity $1 - \beta$ is called the *power* of the test relative to H_1.

In high-energy physics, the components of $\mathbf{x}$ might represent the measured properties of candidate events, and the acceptance region is defined by the cuts that one imposes in order to select events of a certain desired type. Here, H_0 could represent the background hypothesis and the alternative H_1 could represent the sought after signal.

Often, rather than using the full set of quantities $\mathbf{x}$, it is convenient to define a *test statistic*, t, which can be a single number, or in any case a vector with fewer components than $\mathbf{x}$. Each hypothesis for the distribution of $\mathbf{x}$ will determine a distribution for t, and the acceptance region in $\mathbf{x}$ space will correspond to a specific range of values of t.

Often one tries to construct a test to maximize power for a given significance level, i.e., to maximize the signal efficiency for a given significance level. The *Neyman–Pearson lemma* states that this is done by defining the critical region for the test of the background hypothesis H_0 (i.e., the acceptance region for signal, H_1) such that, for $\mathbf{x}$ in that region, the ratio of p.d.f.s for the hypotheses H_1 and H_0,

$$\lambda(\mathbf{x}) = \frac{f(\mathbf{x}|H_1)}{f(\mathbf{x}|H_0)}, \tag{44}$$

is greater than a given constant, the value of which is chosen to give the desired signal efficiency. Here, H_0 and H_1 must be simple hypotheses, i.e., they should not contain undetermined parameters. The lemma is equivalent to the statement that Eq. 44 represents the test statistic with

which one may obtain the highest signal efficiency for a given purity for the selected sample. It can be difficult in practice, however, to determine $\lambda(\mathbf{x})$, since this requires knowledge of the joint p.d.f.s $f(\mathbf{x}|H_0)$ and $f(\mathbf{x}|H_1)$.

In the usual case where the likelihood ratio (Eq. 44) cannot be used explicitly, there exist a variety of other multivariate classifiers that effectively separate different types of events. Methods often used in HEP include *neural networks* or *Fisher discriminants*. Recently, further classification methods from machine-learning have been applied in HEP analyses; these include *probability density estimation (PDE)* techniques, *kernel-based PDE* (*KDE* or *Parzen window*), *support vector machines*, and *decision trees*. Techniques such as "boosting" and "bagging" can be applied to combine a number of classifiers into a stronger one with greater stability with respect to fluctuations in the training data. Descriptions of these methods can be found in Hastie et al. (2009), Webb (2002), Kuncheva (2004), and Proceedings of the PHYSTAT conference series (PHYSTAT Conference Series). Software for HEP includes the `TMVA` (Höcker et al. 2007) and `StatPatternRecognition` (Narsky 2005) packages.

5.2 Significance Tests

Often, one wants to quantify the level of agreement between the data and a hypothesis without explicit reference to alternative hypotheses. This can be done by defining a statistic t, which is a function of the data whose value reflects in some way the level of agreement between the data and the hypothesis. The user must decide what values of the statistic correspond to better or worse levels of agreement with the hypothesis in question; for many goodness-of-fit statistics, there is an obvious choice.

The hypothesis in question, say, H_0, will determine the p.d.f. $g(t|H_0)$ for the statistic. The significance of a discrepancy between the data and what one expects under the assumption of H_0 is quantified by giving the p-value, defined as the probability to find t in the region of equal or lesser compatibility with H_0 than the level of compatibility observed with the actual data. For example, if t is defined such that large values correspond to poor agreement with the hypothesis, then the p-value would be

$$p = \int_{t_{\text{obs}}}^{\infty} g(t|H_0)\,\mathrm{d}t, \tag{45}$$

where t_{obs} is the value of the statistic obtained in the actual experiment. The p-value should not be confused with the size (significance level) of a test or the confidence level of a confidence interval (Sect. 6), both of which are prespecified constants.

The p-value is a function of the data, and is therefore itself a random variable. If the hypothesis used to compute the p-value is true, then for continuous data, p will be uniformly distributed between zero and one. Note that the p-value is not the probability for the hypothesis; in frequentist statistics, this is not defined. Rather, the p-value is the probability, under the assumption of a hypothesis H_0, of obtaining data at least as incompatible with H_0 as the data actually observed.

When searching for a new phenomenon, one tries to reject the hypothesis H_0 that the data are consistent with known, for example, Standard Model processes. If the p-value of H_0 is sufficiently low, then one is willing to accept that some alternative hypothesis is true. Often, one converts the p-value into an equivalent significance Z, defined so that a Z standard deviation upward fluctuation of a Gaussian random variable would have an upper tail area equal to p, i.e.,

$$Z = \Phi^{-1}(1 - p). \tag{46}$$

Here, Φ is the cumulative distribution of the Standard Gaussian, and Φ^{-1} is its inverse (quantile) function. Often in HEP, the level of significance where an effect is said to qualify as a discovery is $Z = 5$, i.e., a 5σ effect, corresponding to a p-value of 2.87×10^{-7}. One's actual degree of belief that a new process is present, however, will depend in general on other factors as well, such as the plausibility of the new signal hypothesis and the degree to which it can describe the data, one's confidence in the model that led to the observed p-value, and possible corrections for multiple observations out of which one focuses on the smallest p-value obtained (the "look-elsewhere effect"). For a review of how to incorporate systematic uncertainties into p-values see, for example, Demortier 2007.

When estimating parameters using the method of least squares, one obtains the minimum value of the quantity χ^2 from Eq. 29. This statistic can be used to test the *goodness-of-fit*, i.e., the test provides a measure of the significance of a discrepancy between the data and the hypothesized functional form used in the fit. It may also happen that no parameters are estimated from the data, but that one simply wants to compare a histogram, for example, a vector of Poisson-distributed numbers $\mathbf{n} = (n_1, \ldots, n_N)$, with a hypothesis for their expectation values $\nu_i = E[n_i]$. As the distribution is Poisson with variances $\sigma_i^2 = \nu_i$, the quantity χ^2 of Eq. 29 becomes *Pearson's chi-square statistic*,

$$\chi^2 = \sum_{i=1}^{N} \frac{(n_i - \nu_i)^2}{\nu_i}. \tag{47}$$

If the hypothesis $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_N)$ is correct, and if the expected values ν_i in Eq. 47 are sufficiently large (in practice, this will be a good approximation if all $\nu_i > 5$), then the χ^2 statistic will follow the chi-square p.d.f. with the number of degrees of freedom equal to the number of measurements N minus the number of fitted parameters. The minimized χ^2 from Eq. 29 also has this property if the measurements y_i are Gaussian.

Alternatively, one may fit parameters and evaluate goodness-of-fit by minimizing $-2 \ln \lambda$ from Eq. 28. One finds that the distribution of this statistic approaches the asymptotic limit faster than does Pearson's χ^2, and thus computing the p-value with the chi-square p.d.f. will in general be better justified (see Cousins and Baker 1984 and references therein).

Assuming the goodness-of-fit statistic follows a chi-square p.d.f., the p-value for the hypothesis is then

$$p = \int_{\chi^2}^{\infty} f(z; n_\mathrm{d})\, \mathrm{d}z, \tag{48}$$

where $f(z; n_\mathrm{d})$ is the chi-square p.d.f. and n_d is the appropriate number of degrees of freedom. If the conditions for using the chi-square p.d.f. do not hold, the statistic can still be defined as before, but its p.d.f. must be determined by other means in order to obtain the p-value, for example, using a Monte Carlo calculation.

Since the mean of the chi-square distribution is equal to n_d, one expects in a "reasonable" experiment to obtain $\chi^2 \approx n_\mathrm{d}$. Hence, the quantity χ^2/n_d is sometimes reported. Since the p.d.f. of χ^2/n_d depends on n_d, however, one must report n_d as well if one wishes to determine the p-value.

5.3 Bayesian Model Selection

In Bayesian statistics, all of one's knowledge about a model is contained in its posterior probability, which one obtains using Bayes' theorem (Eq. 40). Thus, one could reject a hypothesis

H if its posterior probability $P(H|\mathbf{x})$ is sufficiently small. The difficulty here is that $P(H|\mathbf{x})$ is proportional to the prior probability $P(H)$, and there will not be a consensus about the prior probabilities for the existence of new phenomena. Nevertheless, one can construct a quantity called the Bayes factor (described below), which can be used to quantify the degree to which the data prefer one hypothesis over another and is independent of their prior probabilities.

Consider two models (hypotheses), H_i and H_j, described by vectors of parameters $\boldsymbol{\theta}_i$ and $\boldsymbol{\theta}_j$, respectively. Some of the components will be common to both models, and others may be distinct. The full prior probability for each model can be written in the form

$$\pi(H_i, \boldsymbol{\theta}_i) = P(H_i)\pi(\boldsymbol{\theta}_i|H_i). \tag{49}$$

Here, $P(H_i)$ is the overall prior probability for H_i, and $\pi(\boldsymbol{\theta}_i|H_i)$ is the normalized p.d.f. of its parameters. For each model, the posterior probability is found using Bayes' theorem,

$$P(H_i|\mathbf{x}) = \frac{\int L(\mathbf{x}|\boldsymbol{\theta}_i, H_i)P(H_i)\pi(\boldsymbol{\theta}_i|H_i)\,\mathrm{d}\boldsymbol{\theta}_i}{P(\mathbf{x})}, \tag{50}$$

where the integration is carried out over the internal parameters $\boldsymbol{\theta}_i$ of the model. The ratio of posterior probabilities for the models is therefore

$$\frac{P(H_i|\mathbf{x})}{P(H_j|\mathbf{x})} = \frac{\int L(\mathbf{x}|\boldsymbol{\theta}_i, H_i)\pi(\boldsymbol{\theta}_i|H_i)\,\mathrm{d}\boldsymbol{\theta}_i}{\int L(\mathbf{x}|\boldsymbol{\theta}_j, H_j)\pi(\boldsymbol{\theta}_j|H_j)\,\mathrm{d}\boldsymbol{\theta}_j}\frac{P(H_i)}{P(H_j)}. \tag{51}$$

The *Bayes factor* is defined as

$$B_{ij} = \frac{\int L(\mathbf{x}|\boldsymbol{\theta}_i, H_i)\pi(\boldsymbol{\theta}_i|H_i)\,\mathrm{d}\boldsymbol{\theta}_i}{\int L(\mathbf{x}|\boldsymbol{\theta}_j, H_j)\pi(\boldsymbol{\theta}_j|H_j)\,\mathrm{d}\boldsymbol{\theta}_j}. \tag{52}$$

This gives what the ratio of posterior probabilities for models i and j would be if the overall prior probabilities for the two models were equal. If the models have no nuisance parameters, i.e., no internal parameters described by priors, then the Bayes factor is simply the likelihood ratio. The Bayes factor, therefore, shows by how much the probability ratio of model i to model j changes in the light of the data, and thus can be viewed as a numerical measure of evidence supplied by the data in favour of one hypothesis over the other.

Although the Bayes factor is by construction independent of the overall prior probabilities $P(H_i)$ and $P(H_j)$, it does require priors for all internal parameters of a model, i.e., one needs the functions $\pi(\boldsymbol{\theta}_i|H_i)$ and $\pi(\boldsymbol{\theta}_j|H_j)$. In a Bayesian analysis, where one is only interested in the posterior p.d.f. of a parameter, it may be acceptable to take an unnormalizable function for the prior (an improper prior) as long as the product of likelihood and prior can be normalized. But improper priors are only defined up to an arbitrary multiplicative constant, which does not cancel in the ratio of ❯ Eq. 52. Furthermore, although the range of a constant normalized prior is unimportant for parameter determination (provided it is wider than the likelihood), this is not so for the Bayes factor when such a prior is used for only one of the hypotheses. So to compute a Bayes factor, all internal parameters must be described by normalized priors that represent meaningful probabilities over the entire range where they are defined.

An exception to this rule may be considered when the identical parameter appears in the models for both numerator and denominator of the Bayes factor. In this case, one can argue that the arbitrary constants would cancel. One must exercise some caution, however, as parameters with the same name and physical meaning may still play different roles in the two models.

Both integrals in ❯ Eq. 52 are of the form

$$m = \int L(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}, \tag{53}$$

which is called the *marginal likelihood* (or in some fields called the *evidence*). A review of Bayes factors including a discussion of computational issues is Kass and Raftery (1995).

6 Intervals and Limits

When the goal of an experiment is to determine a parameter θ, the result is usually expressed by quoting, in addition to the point estimate, some sort of interval which reflects the statistical precision of the measurement. In the simplest case, this can be given by the parameter's estimated value $\hat{\theta}$ plus or minus an estimate of the standard deviation of $\hat{\theta}$, $\sigma_{\hat{\theta}}$. If, however, the p.d.f. of the estimator is not Gaussian or if there are physical boundaries on the possible values of the parameter, then one usually quotes instead an interval according to one of the procedures described below.

The choice of method may be influenced by practical considerations such as ease of producing an interval from the results of several measurements. Of course, the experimenter is not restricted to quoting a single interval or limit; one may choose, for example, first to communicate the result with a confidence interval having certain frequentist properties, and then in addition to draw conclusions about a parameter using Bayesian statistics. It is recommended, however, that there be a clear separation between these two aspects of reporting a result. In the remainder of this section, we assess the extent to which various types of intervals achieve the goals stated here.

6.1 Bayesian Intervals

As described in Sect. 4.4, a Bayesian posterior probability may be used to determine regions that will have a given probability of containing the true value of a parameter. In the single-parameter case, for example, an interval (called a Bayesian or credible interval) $[\theta_{\rm lo}, \theta_{\rm up}]$ can be determined which contains a given fraction $1 - \alpha$ of the posterior probability, i.e.,

$$1 - \alpha = \int_{\theta_{\rm lo}}^{\theta_{\rm up}} p(\theta|\mathbf{x})\, \mathrm{d}\theta. \tag{54}$$

Sometimes an upper or lower limit is desired, i.e., $\theta_{\rm lo}$ can be set to zero or $\theta_{\rm up}$ to infinity. In other cases, one might choose $\theta_{\rm lo}$ and $\theta_{\rm up}$ such that $p(\theta|\mathbf{x})$ is higher everywhere inside the interval than outside; these are called *highest posterior density* (HPD) intervals. Note that HPD intervals are not invariant under a nonlinear transformation of the parameter.

If a parameter is constrained to be nonnegative, then the prior p.d.f. can simply be set to zero for negative values. An important example is the case of a Poisson variable n, which counts signal events with unknown mean s, as well as background with mean b, assumed known. For the signal mean s, one often uses the prior

$$\pi(s) = \begin{cases} 0 & s < 0 \\ 1 & s \geq 0 \end{cases}. \tag{55}$$

This prior is regarded as providing an interval whose frequentist properties can be studied, rather than as representing a degree of belief. In the absence of a clear discovery, (e.g., if $n = 0$ or if in any case n is compatible with the expected background), one usually wishes to place

an upper limit on s (see, however, Feldman and Cousins 1998 on "flip-flopping" concerning frequentist coverage). Using the likelihood function for Poisson-distributed n,

$$L(n|s) = \frac{(s+b)^n}{n!} e^{-(s+b)}, \tag{56}$$

along with the prior (Eq. 55) in Eq. 40 gives the posterior density for s. An upper limit s_{up} at confidence level (or here, rather, credibility level) $1-\alpha$ can be obtained by requiring

$$1-\alpha = \int_{-\infty}^{s_{up}} p(s|n)\,\mathrm{d}s = \frac{\int_{-\infty}^{s_{up}} L(n|s)\,\pi(s)\,\mathrm{d}s}{\int_{-\infty}^{\infty} L(n|s)\,\pi(s)\,\mathrm{d}s}, \tag{57}$$

where the lower limit of integration is effectively zero because of the cutoff in $\pi(s)$. By relating the integrals in Eq. 57 to incomplete gamma functions, the equation reduces to

$$\alpha = e^{-s_{up}} \frac{\sum_{m=0}^{n} (s_{up}+b)^m/m!}{\sum_{m=0}^{n} b^m/m!}. \tag{58}$$

This must be solved numerically for the limit s_{up}. It so happens that for the case of $b = 0$, the upper limits obtained in this way coincide numerically with the values of the frequentist upper limits discussed in Sect. 6.2. The frequentist properties of confidence intervals for the Poisson mean obtained in this way are discussed in Cousins (1995) and Roe and Woodroofe (2001).

As in any Bayesian analysis, it is important to show how the result would change if one uses different prior probabilities. For example, one could consider the Jeffreys prior as described in Sect. 4.4. For this problem, one finds the Jeffreys prior $\pi(s) \propto 1/\sqrt{s+b}$ for $s \geq 0$ and zero otherwise. As with the constant prior, one would not regard this as representing one's prior beliefs about s, both because it is improper and also as it depends on b. Rather, it is used with Bayes' theorem to produce an interval whose frequentist properties can be studied.

6.2 Frequentist Confidence Intervals

The frequentist approach to interval estimation is based on the the concept of a confidence interval. These are constructed so as to contain the true value of the parameter with a minimum specified probability, called the confidence level. To construct a confidence interval, consider a test (see Sect. 5) of the hypothesis that the parameter's true value is θ (assume one constructs a test for all physical values of θ). One then excludes all values of θ where the hypothesis would be rejected at a significance level less than α. The remaining values constitute the confidence interval at confidence level $\mathrm{CL} = 1-\alpha$.

In this procedure, one is still free to choose the test to be used. One possibility is use a test statistic based on the *likelihood ratio*,

$$\lambda = \frac{f(x;\theta)}{f(x;\hat{\theta})}, \tag{59}$$

where $\hat{\theta}$ is the value of the parameter which, out of all allowed values, maximizes $f(x;\theta)$. This results in the intervals described by Feldman and Cousins (1998).

6.2.1 Profile Likelihood and Treatment of Nuisance Parameters

As mentioned in Sect. 6.1, one may have a model containing parameters that must be determined from data, but which are not of any interest in the final result (nuisance parameters). Suppose the likelihood $L(\boldsymbol{\theta}, \boldsymbol{\nu})$ depends on parameters of interest $\boldsymbol{\theta}$ and nuisance parameters $\boldsymbol{\nu}$. The nuisance parameters can be effectively removed from the problem by constructing the *profile likelihood*, defined by

$$L_{\mathrm{p}}(\boldsymbol{\theta}) = L(\boldsymbol{\theta}, \widehat{\hat{\boldsymbol{\nu}}}(\boldsymbol{\theta})), \tag{60}$$

where $\widehat{\hat{\boldsymbol{\nu}}}(\boldsymbol{\theta})$ is given by the $\boldsymbol{\nu}$ that maximizes the likelihood for fixed $\boldsymbol{\theta}$. The profile likelihood may then be used to construct tests of or intervals for the parameters of interest. For example, one may construct the profile likelihood ratio,

$$\lambda_{\mathrm{p}}(\boldsymbol{\theta}) = \frac{L_{\mathrm{p}}(\boldsymbol{\theta})}{L(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\nu}})}, \tag{61}$$

where $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\nu}}$ are the ML estimators. The ratio λ_{p} can be used in place of the likelihood ratio Eq. 59 for inference about $\boldsymbol{\theta}$. The resulting intervals for the parameters of interest are not guaranteed to have the exact coverage probability for all values of the nuisance parameters, but, in cases of practical interest, the approximation is found to be very good.

6.2.2 Gaussian-Distributed Measurements

One often encounters the case where the data consist of a single random value x modeled as following a Gaussian distribution with unknown mean μ. This is often the case when x represents an estimator for a parameter and one has a sufficiently large data sample. Using the observed value of x, one can easily constuct a confidence interval for μ. Assuming the Gaussian distribution has a known standard deviation σ, the quantity

$$1 - \alpha = \frac{1}{\sqrt{2\pi}\sigma} \int_{\mu-\delta}^{\mu+\delta} \mathrm{e}^{-(x-\mu)^2/2\sigma^2}\, \mathrm{d}x = \mathrm{erf}\left(\frac{\delta}{\sqrt{2}\,\sigma}\right) \tag{62}$$

is the probability that the measured value x will fall within $\pm\delta$ of the true value μ. From the symmetry of the Gaussian with respect to x and μ, this is also the probability for the interval $x \pm \delta$ to include μ.

Figure 1 shows a $\delta = 1.64\sigma$ confidence interval unshaded. The choice $\delta = \sigma$ gives an interval called the *standard error* which has $1 - \alpha = 68.27\%$ if σ is known. Values of α for other frequently used choices of δ are given in *Table 2*. We can set a one-sided (upper or lower) limit by excluding values of μ above $x + \delta$ (or below $x - \delta$). The values of α for such limits are half the values in *Table 2*.

The relation in Eq. 62 can be reexpressed using the cumulative distribution function for the chi-square distribution as

$$\alpha = 1 - F(\chi^2; n), \tag{63}$$

for $\chi^2 = (\delta/\sigma)^2$ and $n = 1$ degree of freedom.

For the case of n parameter estimates $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_n)$, one requires the full covariance matrix $V_{ij} = \mathrm{cov}[\hat{\theta}_i, \hat{\theta}_j]$, which can be estimated as described in Sects. 4.2 and 4.3. Under fairly general conditions with the methods of maximum likelihood or least squares in the large-sample limit, the estimators will be distributed according to a multivariate Gaussian centered

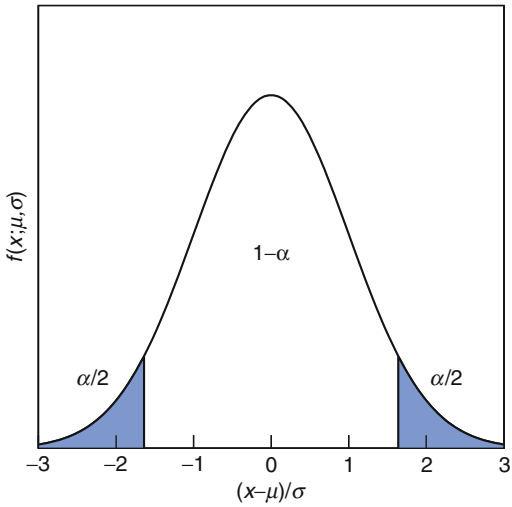


Fig. 1

Illustration of a symmetric 90% confidence interval (*unshaded*) for a measurement of a single quantity with Gaussian errors. Integrated probabilities, defined by α, are as shown

Table 2

Area of the tails α outside $\pm\delta$ from the mean of a Gaussian distribution

α	δ	α	δ
0.3173	1σ	0.2	1.28σ
4.55×10^{-2}	2σ	0.1	1.64σ
2.7×10^{-3}	3σ	0.05	1.96σ
6.3×10^{-5}	4σ	0.01	2.58σ
5.7×10^{-7}	5σ	0.001	3.29σ
2.0×10^{-9}	6σ	10^{-4}	3.89σ

about the true (unknown) values $\boldsymbol{\theta}$, and furthermore, the likelihood function itself takes on a Gaussian shape.

The standard error ellipse for the pair $(\hat{\theta}_i, \hat{\theta}_j)$ is shown in *Fig. 2*, corresponding to a contour $\chi^2 = \chi^2_{\min} + 1$ or $\ln L = \ln L_{\max} - 1/2$. The ellipse is centered about the estimated values $\hat{\boldsymbol{\theta}}$, and the tangents to the ellipse give the standard deviations of the estimators, σ_i and σ_j. The angle of the major axis of the ellipse is given by

$$\tan 2\phi = \frac{2\rho_{ij}\sigma_i\sigma_j}{\sigma_j^2 - \sigma_i^2}, \tag{64}$$

where $\rho_{ij} = \mathrm{cov}[\hat{\theta}_i, \hat{\theta}_j]/\sigma_i\sigma_j$ is the correlation coefficient.

As in the single-variable case, because of the symmetry of the Gaussian function between $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\theta}}$, one finds that contours of constant $\ln L$ or χ^2 cover the true values with a certain, fixed probability. That is, the confidence region is determined by

$$\ln L(\boldsymbol{\theta}) \geq \ln L_{\max} - \Delta \ln L, \tag{65}$$

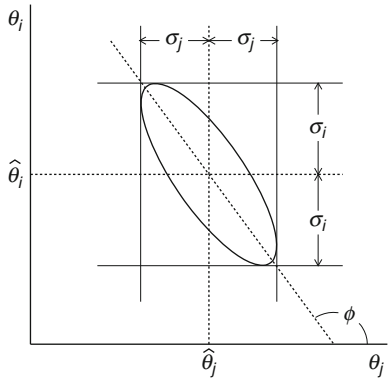


Fig. 2

Standard error ellipse for the estimators $\hat{\theta}_i$ and $\hat{\theta}_j$. In this case, the correlation is negative

Table 3

$\Delta\chi^2$ or $2\Delta \ln L$ corresponding to a coverage probability $1-\alpha$ in the large-data-sample limit, for joint estimation of m parameters

$(1-\alpha)$ (%)	$m = 1$	$m = 2$	$m = 3$
68.27	1.00	2.30	3.53
90.0	2.71	4.61	6.25
95.0	3.84	5.99	7.82
95.45	4.00	6.18	8.03
99.0	6.63	9.21	11.34
99.73	9.00	11.83	14.16

or, where a χ^2 has been defined for use with the method of least squares,

$$\chi^2(\boldsymbol{\theta}) \leq \chi^2_{\min} + \Delta\chi^2. \tag{66}$$

Values of $\Delta\chi^2$ or $2\Delta \ln L$ are given in *Table 3* for several values of the coverage probability and number of fitted parameters.

For finite data samples, the probability for the regions determined by Eqs. 65 or 66 to cover the true value of $\boldsymbol{\theta}$ will depend on $\boldsymbol{\theta}$, so these are not exact confidence regions according to our previous definition. Nevertheless, they can still have a coverage probability only weakly dependent on the true parameter and approximately as given in *Table 3*. In any case, the coverage probability of the intervals or regions obtained according to this procedure can in principle be determined as a function of the true parameter(s), for example, using a Monte Carlo calculation.

6.2.3 Poisson or Binomial Data

Another important class of measurements consists of counting a certain number of events, n. In this section, we will assume these are all events of the desired type, i.e., there is no background.

If n represents the number of events produced in a reaction with cross section σ, say, in a fixed integrated luminosity $\mathcal{L}$, then it follows a Poisson distribution with mean $\nu = \sigma\mathcal{L}$. If, on the other hand, one has selected a larger sample of N events and found n of them to have a particular property, then n follows a binomial distribution where the parameter p gives the probability for the event to possess the property in question. This is appropriate, for example, for estimates of branching ratios or selection efficiencies based on a given total number of events.

For the case of Poisson-distributed n, the lower and upper limits on the mean value ν can be found from

$$\nu_{\mathrm{lo}} = \tfrac{1}{2} F^{-1}_{\chi^2}(\alpha_{\mathrm{lo}}; 2n), \tag{67}$$

$$\nu_{\mathrm{up}} = \tfrac{1}{2} F^{-1}_{\chi^2}(1-\alpha_{\mathrm{up}}; 2(n+1)), \tag{68}$$

where the upper and lower limits are at confidence levels of $1-\alpha_{\mathrm{lo}}$ and $1-\alpha_{\mathrm{up}}$, respectively, and $F^{-1}_{\chi^2}$ is the *quantile* of the chi-square distribution (inverse of the cumulative distribution). For central confidence intervals at confidence level $1-\alpha$, set $\alpha_{\mathrm{lo}} = \alpha_{\mathrm{up}} = \alpha/2$.

For the case of binomially distributed n successes out of N trials with probability of success p, the upper and lower limits on p are found to be

$$p_{\mathrm{lo}} = \frac{nF^{-1}_F[\alpha_{\mathrm{lo}}; 2n, 2(N-n+1)]}{N-n+1+nF^{-1}_F[\alpha_{\mathrm{lo}}; 2n, 2(N-n+1)]}, \tag{69}$$

$$p_{\mathrm{up}} = \frac{(n+1)F^{-1}_F[1-\alpha_{\mathrm{up}}; 2(n+1), 2(N-n)]}{(N-n)+(n+1)F^{-1}_F[1-\alpha_{\mathrm{up}}; 2(n+1), 2(N-n)]}. \tag{70}$$

Here, F^{-1}_F is the quantile of the F distribution (also called the Fisher–Snedecor distribution; see James 2007).

7 Conclusions

Given the high cost and complexity of experiments in Particle Physics, it has become increasingly important to use data analysis methods that extract the maximum information from the data in a way that takes into account all of the known uncertainties in the measurement. Here, the key is to construct a probabilistic model which is sufficiently accurate to be regarded as correct, and this can be achieved with a model containing a sufficient number of adjustable parameters. The accuracy achieved by including additional parameters must be balanced against the price, however, of reducing one's sensitivity to the parameters of interest, such as those which may point to a potential discovery.

The two primary schools of statistical inference – frequentist and Bayesian – provide different but related approaches to this task. In the fortunate case where the information from the data overwhelms any prior knowledge, the two approaches appear to coalesce, although the interpretation of the results remains distinct. One can of course use both approaches; if they point to the same conclusion, then this can only increase one's confidence. If they do not, then one will have to find out why, and this can also lead to important realisations, such as unexpected

sensitivity to prior information or to specific model assumptions. In either case, the result of an experiment should be presented along with sufficient information so that it can be incorporated into future analyses.

Acknowledgments

The author is indebted to the Particle Data Group for permission to adapt the material from the *Review of Particle Physics* (Nakamura et al. 2010) for this article, as well as to Dr Tilo Stroh and Professor Claus Grupen for their support and assistance.

References

Nakamura K et al (2010) (Particle Data Group), J Phys G37, 075021

Baker S, Cousins R (1984) Nucl Inst Meth 221:437 (For a reveiw)

Barlow R (1990) Nucl Inst Meth A297:496

Bernardo JM, Smith AFM (2000) Bayesian theory. Wiley, Chichester

Cousins RD (1995) Am J Phys 63:398

Cowan G (1998) Statistical data analysis. Oxford University Press, New York

Demortier L (2007) P-values and nuisance parameters. In: Proceedings of PHYSTAT 2007, CERN-2008-001, p 23

Feldman GJ, Cousins RD (1998) Phys Rev D57:3873

Gregory PC (2005) Bayesian logical data analysis for the physical sciences, Cambridge University Press, Cambrdige

Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning, 2nd edn. Springer, New York

Höcker A et al (2007) TMVA users guide, physics/0703039; software available from tmva.sf.net

James FE (2007) Statistical methods in experimental physics, 2nd edn. World Scientific, Singapore

Kass RE, Raftery AE (1995) Bayes factors. J Am Stat Assoc 90(430):773–795

Kass RE, Wasserman L (1996) The selection of prior distributions by formal rules. J Am Stat Assoc 91(435): 1343–1370

Kolmogorov AN (1993) Grundbegriffe der Wahrscheinlichkeitsrechnung. Springer, Berlin; Foundations of the theory of probability, 2nd edn. Chelsea, New York 1956

Kuncheva LI (2004) Combining pattern classifiers. Wiley, New York

Links to the Proceedings of the PHYSTAT conference series, (Durham 2002, Stanford 2003, Oxford 2005, and Geneva 2007) can be found at phystat.org

Lyons L (1986) Statistics for nuclear and particle PHysicists. Cambridge University Press, New York

Narsky I (2005) StatPatternRecognition: a C++ package for statistical analysis of high energy physics data, physics/0507143 (2005); software available from sourceforge.net/projects/statpatrec

O'Hagan A, Forster JJ (2004) Bayesian inference. In: Kendall's advanced theory of statistics, vol 2B, 2nd edn. Arnold, London

Roe BP, Woodroofe MB (2001) Phys Rev D63:13009

Sivia D, Skilling J (2006) Data analysis: a Bayesian tutorial, Oxford University Press, New York

Stuart A, Ord JK, Arnold S (1999) Kendall's advanced theory of statistics. In: Kendall M, Stuart A (eds) Classical inference and the linear model, vol 2A, 6th edn. Oxford University Press, New York

Webb A (2002) Statistical pattern recognition, 2nd edn. Wiley, New York

6 Particle Identification

Jürgen Engelfried
Universidad Autónoma de San Luis Potosí, Manuel Nava #6, San Luis Potosí, Mexico

1 ***Introduction*** ***126***

2 ***Radiation by Charged Particles*** ***127***

3 ***Particle Identification in Calorimeters*** ***127***

4 ***Time-of-Flight (TOF) Measurements*** ***128***

5 ***Specific Energy Loss*** $\mathbf{dE/dx}$ ***128***

6 ***Transition Radiation Detectors (TRDs)*** ***130***

7 ***Cherenkov Radiation*** ***132***
7.1 Threshold Cherenkov Detectors 133
7.2 Ring Imaging 133

8 ***Muon Identification*** ***135***

9 ***Neutrinos*** ***135***

10 ***Conclusions*** ***135***

11 ***Cross-References*** ***135***

References ***136***

C. Grupen, I. Buvat (eds.), *Handbook of Particle Detection and Imaging*,
DOI 10.1007/978-3-642-13271-1_6, © Springer-Verlag Berlin Heidelberg 2012

Abstract: We present here the devices and techniques used to identify particles. Most detectors and detector systems are also described in other parts of this book; here we emphasize their use in the identification of particles.

1 Introduction

The identification of particles involved in a high-energy or nuclear reaction is of utmost importance. Only knowing all (or at least most) of the 4-vectors of the particles, the underlying physics reaction can be studied enabling the distinction of the signal from the background.

In general, there are two major applications for particle identification: (1) identification of beam particles, especially for fixed-target experiments, and (2) identification of decay products. If the (vector-)momentum of the particle is known, either selected naturally by the beam-line elements in a small momentum bin, or by means of a magnetic spectrometer, particle identification necessitates the measurement of some other kinematic variable, like the velocity, the total energy, or the specific energy loss $\mathrm{d}E/\mathrm{d}x$.

Velocity-dependent effects used in particle identification devices in addition to the already mentioned $\mathrm{d}E/\mathrm{d}x$, include the measurement of the time of flight (TOF), the Cherenkov and transition radiation.

To identify long-lived (but still weakly decaying) neutral particles like the hyperons Λ^0 and Ξ^0, and short-lived particles (τ, charm, beauty, resonances), the determination of the 4-vector of all decay products is necessary to be able to calculate the invariant mass of the final state, suppress background, and identify the original particle.

Particle identification thus reduces to identify all stable (or practically stable) particles: p, n, $K^{\pm}$, K^0_{L}, $\pi^{\pm}$, $e^{\pm}$, $\mu^{\pm}$, and γ.

In the case of electrically neutral particles, only the total energy in a calorimeter can be measured, complicating somewhat the identification process. If no charged track is pointing to where the signal originated in the calorimeter, we could conclude that the signal was produced by a neutral particle. Depending on the type of calorimeter (electromagnetic or hadronic) not too many possibilities are usually left, but some ambiguities remain.

The different detector systems have to be arranged correctly in order to measure all the different properties needed; for example, calorimeters have to be positioned after the tracking, transition radiation, and Cherenkov detectors, since they absorb the particles totally and no further measurements are possible (with the exception of muons). The selection of a specific technique used in a particular application not only depends on the type of particle (like charged, neutral, hadron), but also on the momentum range for which an identification of the particle is desired.

As already mentioned, in this chapter we describe the use of detectors and detector systems which are also discussed in other chapters. We will repeat here some of the aspects with emphasis on particle identification.

At the end of the chapter, we will summarize the different options discussed as to identify the different particles.

2 Radiation by Charged Particles

The interaction of charges particles with matter is dominated by the electromagnetic interaction, via the exchange of virtual or real photons. Virtual photons are absorbed by the atoms of the material which leads to ionization and/or excitation of the atoms. Depending on the type of acceleration and/or condition, the resulting radiation and effects carry different names:

1. $|\boldsymbol{v}| > c/n$: *Cherenkov radiation*
2. $\boldsymbol{v}/c_{\mathrm{ph}} = \boldsymbol{v} \cdot n/c$ changes
   (a) $|\boldsymbol{v}|$ changes: *bremsstrahlung*
   (b) Direction of $\boldsymbol{v}$ changes: *synchrotron radiation*
   (c) n changes: *transition radiation*
3. Ionization and/or excitation of matter: specific energy loss $\mathrm{d}E/\mathrm{d}x$

Here $\boldsymbol{v}$ is the velocity of the particle, c the speed of light in vacuum, c_{ph} the phase velocity in a medium, and n the refractive index of the medium. Any one of these conditions can occur separately or in any combination.

Any one of these effects has a different dependency on the velocity, the charge, and the mass of the particle, and on the properties of the material the particle is passing and interacting with. In the following, we will show how some of them can be used for the task of particle identification.

3 Particle Identification in Calorimeters

Calorimeters (see Chap. 20, "Calorimeters") are discussed in detail within this book. They are used to determine the total energy of the particle or a jet. Depending on the material(s) used and how they are arranged (homogeneous, sampling), calorimeters are usually subdivided into "electromagnetic" and "hadronic," the idea being that particles interacting mostly electromagnetically (electrons and photons) are absorbed in the "electromagnetic" part, and particles interacting strongly (like $\pi^{\pm}$, p, n, $K^{\pm}$, K^0_{L}, …) in the "hadronic" part. It has also to be taken into account that charged hadrons lose part of their energy (but not all) in the electromagnetic section.

In hadronic calorimeters with sufficient segmentation, individual particles (as opposed to whole jets) can be detected and their individual total energy can be measured. The distinction between a charged and neutral hadron is performed if the track of a charged particle, measured previously with some tracking system, can be associated with the shower in the calorimeter or not. Further identification, for example the separation of p, π^+, or K^+, is not possible within an hadronic calorimeter and other techniques (see later) have to be used to achieve this.

Electrons and photons are distinguished also with the charged-track-pointing method in an electromagnetic calorimeter. In this device, in addition one can compare for a charged track the deposited (measured) energy over the previously measured momentum of a particle (called "E/p"). If $E/(cp) \sim 1$, a (relativistic) electron originated the signal, for hadrons one has $E/(cp) < 1$ (deposition of only a part of the total energy), and muons produce an energy deposit compatible with minimum ionization only.

Calorimeters used for particle identification need to have in addition to an excellent energy resolution a fine segmentation, so individual particles can be measured and charged tracks can be associated with the calorimeter signals.

4 Time-of-Flight (TOF) Measurements

This is the most straightforward method for measuring the velocity and thus the identification of charged particles: The time difference $L/(\beta c)$ between the signals of two (usually scintillation or gas) counters at a known distance L is measured, thus the difference Δt of such times for two particles with different masses but same momentum is

$$\Delta t = \frac{L}{\beta_1 c} - \frac{L}{\beta_2 c} = \frac{L}{c}\left[\sqrt{1+\frac{m_1^2 c^2}{p^2}} - \sqrt{1+\frac{m_2^2 c^2}{p^2}}\right].$$

Here p is the momentum of the particles determined by a magnetic spectrometer, β_i are the velocities in units of c, and m_i the masses of two particles, respectively. In the relativistic limit ($p^2 \gg m^2 c^2$) this reduces to $\Delta t \approx (m_1^2 - m_2^2)Lc/(2p^2)$.

Time resolutions in running systems of about 100 ps have been achieved (Klempt 1999). With the maximum distance possible between the two detectors (≈10 m for measuring decay products, ≈100 m in a beam line) kaons and pions can be separated up to a few GeV/c. At higher rate and/or more than one particles hitting the same detector elements the time difference measurement is ambiguous and the method will not work anymore.

A current example for the use of this technique is the ALICE TOF detector (The Alice Collaboration 1995, 2002, 2008), based on Multigap Resistive Plate Chambers (MRPCs); it should reach timing resolutions of about 65 ps.

A new development based on DIRC-like (see Chap. 18, "Cherenkov Counters") devices with fast photon detectors is proposed as an upgrade to the BELLE detector as well as for the new detector at the Frascati B-Factory. Here time resolutions as low as 5 ps have been reported in prototype devices (Inami et al. 2006; Korpar et al. 2007; Va'vra 2011; Va'vra et al. 2008).

5 Specific Energy Loss dE/dx

An excellent derivation of the specific energy loss in material is given in Grupen (1996). A more exact treatment, summarized in great detail in Particle Data Group (2010), gives as the final result for the mean rate of energy loss dE in a small distance dx (the "Bethe–Bloch formula")

$$-\left\langle\frac{\mathrm{d}E}{\mathrm{d}x}\right\rangle = 2\pi\frac{Z N_\mathrm{A}}{A}\frac{2 r_e^2 m_e c^2}{\beta^2} z^2 \left[\frac{1}{2}\ln\frac{2 m_e c^2 \gamma^2 T_\mathrm{max}}{I^2} - \beta^2 - \frac{\delta}{2}\right]. \tag{1}$$

This equation describes the energy loss for $0.1 \le \beta\gamma \le 1{,}000$ with an accuracy of a few percent. The density correction at high energies can be described by

$$\frac{\delta}{2} = \ln\frac{\hbar\omega_p}{I} + \ln\beta\gamma - \frac{1}{2}, \quad \hbar\omega_p = \sqrt{4\pi N_e r_e^2}\,\frac{m_e c^2}{\alpha}, \tag{2}$$

where N_e is the electron density and ω_p the plasma frequency of the absorbing material, and α the Sommerfeld fine-structure constant. $T_{\max}$, the maximum kinetic energy which can be imparted to a free electron in a single collision, is given by

$$T_{\max} = \frac{2m_e c^2 \beta^2 \gamma^2}{1 + 2\gamma m_e/M + (m_e/M)^2} \tag{3}$$

with a low-energy ($2\gamma m_e/M \ll 1$) approximation of $T_{\max} = 2m_e c^2 \beta^2 \gamma^2$.

In *Fig. 1* (taken from Particle Data Group (2010)), we show examples for energy loss in different materials. The minimum energy loss of all particles in nearly all materials occurs at $3 \le \beta\gamma \le 4$, and is in the range of MeV/(g/cm^2) (examples: helium: $-\langle dE/dx \rangle = 1.94$ MeV/(g/cm^2), uranium: 1.08 MeV/(g/cm^2)) with the exception of hydrogen, in which particles experience a larger energy loss ($Z/A = 1$). Due to the $\ln\gamma$ term the energy loss increases for relativistic particles and reaches the so-called Fermi plateau, limited by the density effect. In gases, the plateau is typically ≈60% higher than the minimum.

Due to the increased energy loss at smaller $\beta\gamma$, particles will deposit most of their energy at the end of their track, just before they will be completely stopped. This "Bragg peak" is used for the treatment of deep-seated tumors, selecting the particle type and initial energy to optimize the energy loss close to the tumor location, avoiding too much damage to tissue in front of the tumor. This is described in more detail in Chap. 47, "Tumor Therapy with Ion Beams" of this book.

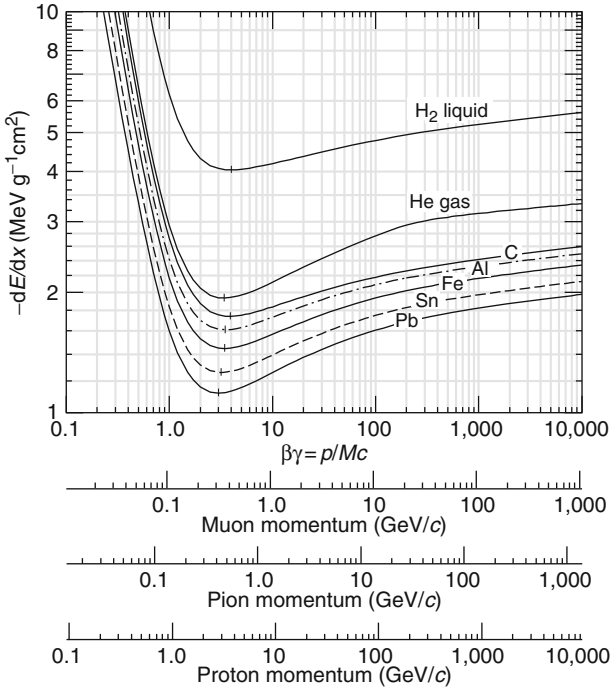


Fig. 1
Mean energy loss for different materials (Particle Data Group 2010)

The energy loss is distributed asymmetrically around the mean energy loss described by the Bethe–Bloch formula (Eq. 1); the distribution can be approximated by the Landau distribution $\Omega(\lambda)$,

$$\Omega(\lambda) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}(\lambda + \mathrm{e}^{-\lambda})}, \tag{4}$$

where $\lambda = \left[\left(\frac{\mathrm{d}E}{\mathrm{d}x}\right) - \left(\frac{\mathrm{d}E}{\mathrm{d}x}\right)^{\mathrm{m.p.}}\right] / (0.123\,\mathrm{keV})$, with $\left(\frac{\mathrm{d}E}{\mathrm{d}x}\right)^{\mathrm{m.p.}}$ being the most probable energy loss. In gases and thin absorbers, the Landau fluctuations have to be considered; for example, a particle with $\beta\gamma = 4$ in Argon experiences a most probable energy loss of $\left(\frac{\mathrm{d}E}{\mathrm{d}x}\right)^{\mathrm{m.p.}} = 1.2\,\mathrm{keV/cm}$ and a mean energy loss of $\left\langle\frac{\mathrm{d}E}{\mathrm{d}x}\right\rangle = 2.69\,\mathrm{keV/cm}$.

The Landau fluctuation presents a problem when using the energy loss for particle identification. To obtain the mean (or most probable) value of the energy loss one has to sample often (typically hundred times or more) and use adequate algorithms (e.g., "truncated mean") to determine the correct averages.

Only in the so-called relativistic rise, after the minimum and well before the Fermi plateau, are the curves for different particles sufficiently separated to be useful for particle identification. This limits the range to particle momenta up to a few GeV/c. A detailed description can be found in Blum et al. (2008). More historic examples for successful application of this particle identification method can be found in the literature (Fischer et al. 1986; Heintze 1982; Nygren 1978). The ALICE TPC (The Alice Collabortion 2000, 2010) is a recent example for the use of this technique.

6 Transition Radiation Detectors (TRDs)

Transition radiation is emitted due to the reformation of the particle field if it crosses a boundary between two media with different dielectric properties. The total energy S of radiation emitted at a single interface is given by

$$S = \frac{\alpha\hbar Z^2}{3} \frac{(\omega_1 - \omega_2)^2}{\omega_1 + \omega_2} \gamma, \tag{5}$$

where $\alpha = 1/137$ (the fine-structure constant), ω_1 and ω_2 are the plasma frequencies, Z is the atomic number of the foil material, and $\gamma = E/(mc^2)$ (the relativistic γ). Typical values for the plasma frequencies are $\omega_1 = 0.7\,\mathrm{eV}$ for air, $\omega_2 = 20\,\mathrm{eV}$ for polypropylene. The spectral (ω) and angular (ϑ) dependence of the transition radiation is given by

$$\frac{\mathrm{d}^2 S}{\mathrm{d}\vartheta\,\mathrm{d}\omega} = \frac{2e^2}{\pi c} \left(\frac{\vartheta}{\gamma^{-2} + \vartheta^2 + \omega_1^2/\omega^2} - \frac{\vartheta}{\gamma^{-2} + \vartheta^2 + \omega_2^2/\omega^2} \right)^2. \tag{6}$$

Most of the radiation is emitted in a cone with half angle $1/\gamma$ in the forward direction of the particle. A detailed description of the spectrum of transition radiation can be found in Paul (1991).

A minimum foil thickness is needed for the particle field to reach a new equilibrium inside the medium. Since there are two transitions (into and out of the foil) which are equal ($\omega_1 \to \omega_2$ and $\omega_2 \to \omega_1$) interference effects are seen as maxima and minima in the spectrum. Spacing the foils in the stack at equal distances, one can expect in addition interference between the amplitudes of different foils. As the number of foils increases, reabsorption of the radiation

($\propto Z^5$) is observed, so usually low-Z materials like Mylar, CH_2, carbon fibers, or lithium are used as foils, with a thickness of 30 μm and foil distances of 300 μm.

Typical photon energies are in the few-keV range. The number of X-ray photons emitted is proportional to γ (Eq. 5), but only for $\gamma \gtrsim 1{,}000$ photons are emitted. This limits this technique to the identification of electrons as decay products (see the introduction to this chapter) and to the separation of heavy and light particles in high-momentum (several hundred GeV/c) secondary high-energy beamlines for fixed-target experiments.

As mentioned before, the X-ray photons from transition radiation are emitted forward under a small angle to the particle track, so whatever detector is used to detect the X rays, it will be traversed by the particle itself as well, leaving an additional $\mathrm{d}E/\mathrm{d}x$ energy loss in the detector. A typical $\mathrm{d}E/\mathrm{d}x$ in gas detectors is some keV/cm and is Landau distributed; signals from $\mathrm{d}E/\mathrm{d}x$ and the X rays are very similar. A detector consists typically (Brückner et al. 1996; Errede et al. 1989, 1991; Paul 1991; Terentiev et al. 1995) of a "thin" (to minimize the $\mathrm{d}E/\mathrm{d}x$ signal) MWPC, with xenon or krypton as counting-gas additions, and several (10–30) radiator/chamber units to be able to determine effects due to the Landau distribution.

To separate the signals, and to discriminate between the (Landau-distributed) $\mathrm{d}E/\mathrm{d}x$ and the absorbed X-ray signal, two different analysis methods, which also translate to specifications on the readout electronics, are used: Charge Integration and Cluster Counting. The Charge Integration method measures the total charge of the ionization in every unit, and counts how many units detected a total charge above some given threshold. The Cluster Counting method employs the different spatial ionization distribution of the two sources (point for X ray, distributed for $\mathrm{d}E/\mathrm{d}x$) to separate them.

In *Fig. 2*, we show an example of a TRD detector which identifies the beam particles in the SELEX (Russ 1995; Russ et al. 1987) experiment. It consists of ten units, each of them with three wire planes, and is operated in the Charge Integration mode. The electronics used is seen on the bottom. The obtained signals (either clusters or charge) are compared with a likelihood

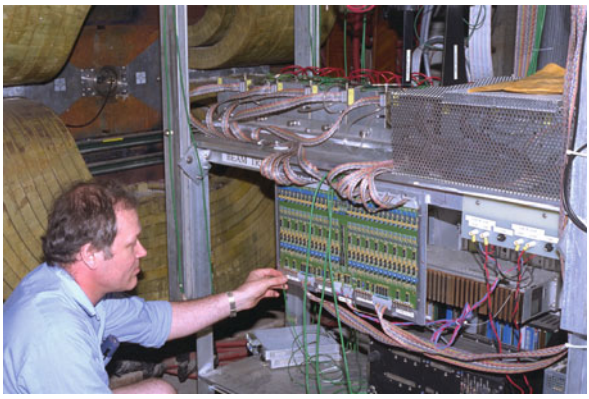


Fig. 2
The SELEX (Russ 1995; Russ et al. 1987) Beam Transition Radiation Detector (Terentiev et al. 1995) separates Σ^- from π^- in a 600 GeV/c fixed-target beam line. The exit of the magnetic channel can be seen on the left. The detector consists of 10 radiator/chamber units. The readout electronics is mounted below the detector

method to different particle hypotheses, with known momentum. The most probable hypothesis is selected for the identification.

The most sophisticated use of this technology is applied within the ATLAS (Aad et al. 2008) detector at LHC. The Transition Radiation Tracker (TRT) (Abat et al. 2008a, b) is based on the use of Straw Detectors, which can operate at high rates due to their small diameter (4 mm) and the isolation of the sense wires within individual gas volumes. Electron identification capability is added by employing xenon gas to detect transition radiation photons created in a radiator between the straws. In total (barrel and endcap) 370,000 straws are used. The detector, as says its name, doubles as a tracking device, measuring the drift time for every signal. First performance results are available in Olivito (2010).

7 Cherenkov Radiation

Even though the basic idea of determining the velocity of charged particles via measuring the Cherenkov angle was proposed in 1960 (Roberts 1960), and in 1977 a first prototype was successfully operated (Séguinot and Ypsilantis 1977), it was only during the last two decades that Ring Imaging Cherenkov (RICH) detectors were successfully used in experiments. A very useful collection of review articles and detailed descriptions can be found in the proceedings of seven international workshops on this type of detectors, which were held in 1993 (Bari, Italy) (Nappi and Ypsilantis 1994), 1995 (Uppsala, Sweden) (Ekelöf 1996), 1998 (Ein Gedi, Israel) (Breskin et al. 1999), 2002 (Pylos, Greece) (Ekelof et al. 2003), 2004 (Playa del Carmen, Mexico) (Engelfried and Paic 2005), 2007 (Trieste, Italy) (Bressan et al. 2008), and 2010 (Cassis, France) (Hallewell et al. 2011), respectively.

Charged particles with a velocity $|\boldsymbol{v}|$ larger than the speed of light in a medium with refractive index n will emit Cherenkov radiation under an angle θ_c, given by Cherenkov (1937)

$$\cos\theta_c = \frac{1}{\beta n} \tag{7}$$

with $\beta = v/c$, c being the speed of light in vacuum. The threshold velocity v_{thres} is given by

$$\beta_{\text{thres}} = \frac{v_{\text{thres}}}{c} \geq \frac{1}{n}, \quad \text{or} \quad \gamma_{\text{thres}} = \frac{n}{\sqrt{n^2-1}}, \tag{8}$$

with a maximum angle of $\theta_c^{\max} = \arccos(1/n)$. For water $\theta_c^{\max} = 42°$, for neon at 1 atm $\theta_c^{\max} = 11$ mrad.

The number of photons N emitted per energy interval $\mathrm{d}E$ and path length $\mathrm{d}l$ is given by Frank and Tamm (1937),

$$\frac{\mathrm{d}^2N}{\mathrm{d}E\,\mathrm{d}l} = \frac{\alpha}{\hbar c}\left(1 - \frac{1}{(\beta n)^2}\right) = \frac{\alpha}{\hbar c}\sin^2\theta\,, \tag{9}$$

or, expressed for a wavelength interval $\mathrm{d}\lambda$,

$$\frac{\mathrm{d}^2N}{\mathrm{d}\lambda\,\mathrm{d}l} = \frac{2\pi\alpha}{\lambda^2}\sin^2\theta\,. \tag{10}$$

Mostly gas radiators are used in Cherenkov counters; but also solid (quartz) and liquid radiators can be found. Water Cherenkov counters were originally developed to set limits on the

lifetime of protons, but got converted for (solar) neutrino detection. One example is (Super-) Kamiokande. Cherenkov effect in water is also used in the tanks of the Auger Experiment to detect muons produced in cosmic-ray air showers.

7.1 Threshold Cherenkov Detectors

A first (obvious) application are threshold Cherenkov counters. For a fixed momentum, and only two particle types to separate, one chooses a medium with appropriate index of refraction, with a threshold between the velocities of the two particles. If light is detected, one can conclude that the lighter particle has passed. For more than two particle types and/or a wider momentum range to cover, several counters with different thresholds have to be employed. Since these counters have radiator lengths of several meters each, in practice no more than three counters are used.

7.2 Ring Imaging

By measuring the Cherenkov angle θ_c one can in principle determine the velocity of the particle, which will, together with the momentum p obtained via a magnetic spectrometer, lead to the determination of the mass and therefore to the identification of the particle.

Neglecting multiple scattering and energy loss in the medium, all the Cherenkov light (in one plane) is parallel, and can therefore be focused (for small θ_c) with a spherical mirror (radius R) onto a point. Since the emission is symmetrical in the azimuthal angle around the particle trajectory, this leads to a ring of radius r in the focus, which is itself a sphere with radius $R/2$. The radius r is given by

$$r = \frac{R}{2} \tan\theta_c \approx \frac{R}{2}\sqrt{2 - \frac{2}{n}\sqrt{1 + \frac{m^2c^2}{p^2}}}\,. \tag{11}$$

The angular separation between two particles of masses m_1 and m_2 (in the small angle, relativistic, and small $(n-1)$ approximation) is given by

$$\theta_c\,\Delta\theta_c = \frac{m_1^2 - m_2^2}{2p^2}\,. \tag{12}$$

All these features can obviously been measured, as is shown in ❯ *Fig. 3*, were nearly 98 million "negative" tracks, with their momentum and ring radius measured, were used (Engelfried et al. 1998, 1999a, b, 2003). A similar plot was obtained for "positive" tracks, 16 particles and antiparticles can be identified.

A recent review of some historical developments on RICH detectors is given in Engelfried (2011, in print). The first RICH detectors were used in the 1980s, with mixed results. Today there is a wide range of detectors, with gas, liquid, and solid radiators in use and in development for new experiments.

The refractive index of the radiator material defines the momentum range in which particle identification can be performed. Most (but not all) photon detectors operate in the ultraviolet or even VUV region, since due to the higher photon energy it is generally easier to effectively detect the photon. Since the RICHes are usually placed downstream of a magnetic field and

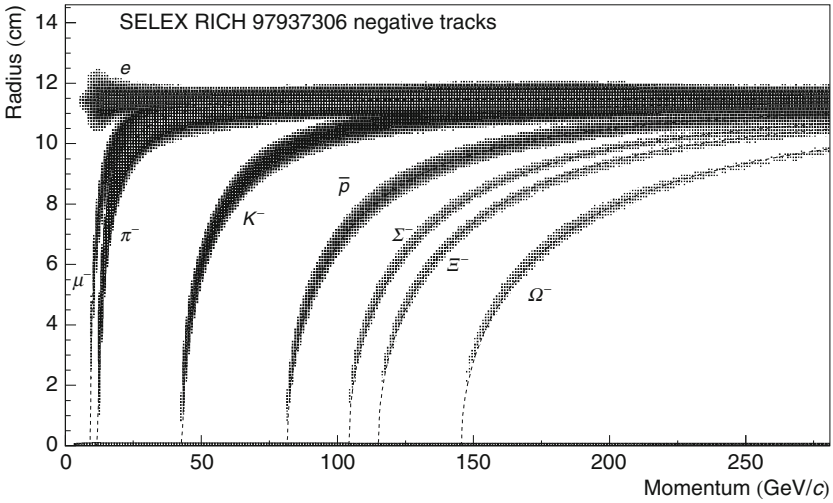


Fig. 3

Radius of rings versus momentum measured by the SELEX RICH (Engelfried et al. 1998, 1999a, b, 2003). Nearly 98 million "negative" tracks were used. We can clearly identify eight different particle types. The lines are absolute predicted ring radii for μ^-, π^-, K^-, $\bar{p}$, Σ^-, Ξ^-, and Ω^-

lower-momentum particles are deflected and do not reach the RICH detector, a typical selection for the pion threshold (and thus n) is the minimum momentum of the particles reaching the detector. After the refractive index (or the gas) is selected, one has to check if the resolution at high momentum (see Eq. 12) is good enough for the experiment. If not, there are several possible solutions, all of them adding additional complexity to the detector: better spatial resolution in the photon detector, more than one radiator. Depending on the refractive index, π/K separation can be performed from a few GeV/c to well above 150 GeV/c.

Early RICH detectors used photon-sensitive vapors like TEA or TMAE for the conversion of the Cherenkov photon to one electron, and a MWPC, a TPC, or TEC to detect the electron. A window separating the radiator from the photon detector is necessary, and has to be transparent to the photons which have enough energy to ionize the vapor. Also the counting gas in the chamber has to be transparent for the relevant photons. In the classical geometry the charged particle itself could pass through the chamber, leaving a huge $\mathrm{d}E/\mathrm{d}x$ signal (typically a few hundred electrons) in the chamber and disturbing the measurement of the single photoelectron.

Also the classical photon detector, the photomultiplier, is used in sizes up to 40 cm diameter in water Cherenkov detectors; and as small as 13 mm in RICH detectors. PMTs are used when the number of pixels required is not too large; the largest PMT system (BaBar) had about 10,000 channels.

Other photon detectors used or planned to be used are: Multi-Anode PMTs, Microchannel Plates, Hybrids (photocathode with Silicon Strip/Pixel Detector), CsI photocathode with a "GEM" to detect the electron, and solid-state (silicon) devices.

Recent experiments using RICH detectors include LHCb and ALICE at the LHC, COMPASS at CERN, PHENIX. There are several experiments developing new detectors, like NA62, BELLE II, CBM, PANDA, WASA-at-COSY, CLAS.

8 Muon Identification

In a typical setup – in case of a fixed-target experiment linear, in case of a collider experiment cylindrical – all particles except muons and neutrinos are absorbed in the calorimeters. To identify muons, one additional system of tracking detectors is used after/outside the calorimeters. If there are any signals, this indicates the presence of a muon. More details of these systems can be found in Chap. 19, "Muon Spectrometers" of this book.

9 Neutrinos

Neutrinos only interact weakly; in experiments where neutrinos appear in the final state (e.g., in semileptonic decays) it is impossible to measure the neutrinos directly. The only possibility to deduce that there are neutrinos is by measuring the entire final state, at least in some projection (like in transverse momentum), and conclude from the momentum imbalance that some undetectable particle escaped. This requires a very hermetic detector. Direct detection of neutrinos is discussed in Chap. 14, "Neutrino Detectors."

10 Conclusions

We summarize here the different techniques used to identify particles.

p, $K^{\pm}$, and $\pi^{\pm}$ are identified in the few GeV/c momentum region by $\mathrm{d}E/\mathrm{d}x$, time-of-flight, or Cherenkov detectors with solid (usually quartz) radiator; at higher momenta up to several hundred GeV/c by Ring Imaging Cherenkov Counters (RICH).

Photons are measured in electromagnetic calorimeters, where electrons as well deposit all their energy; while hadron on the contrary only deposit part of it. Hadrons (the already mentioned p, $\pi^{\pm}$, $K^{\pm}$, but also n, K_{L}^{0}) are measured in hadron calorimeters. Neutral particles are identified by the missing track impact.

Highly relativistic particles (electrons from reactions, or beam particles) are identified with the help of transition radiation.

Finally muons are identified as the only particles passing all materials.

11 Cross-References

- Chapter 1, "Interactions of Particles and Radiation with Matter"
- Chapter 11, "Gaseous Detectors"
- Chapter 12, "Tracking Detectors"
- Chapter 13, "Photon Detectors"
- Chapter 15, "Scintillation Counters"
- Chapter 18, "Cherenkov Counters"
- Chapter 19, "Muon Spectrometers"
- Chapter 20, "Calorimeters"
- Chapter 24, "Indirect Detection of Cosmic Rays"
- Chapter 31, "Neutron Detection"

References

Aad G et al. ATLAS Collaboration (2008) The ATLAS experiment at the CERN Large Hadron Collider. JINST 3, S08003

Abat E et al. ATLAS TRT Collaboration (2008a) The ATLAS TRT barrel detector. JINST 3, P02014

Abat E et al. The ATLAS TRT end-cap detectors (2008b) JINST 3, P10003

Blum W et al (2008) Particle detection with drift chambers, 2nd edn. Springer-Verlag, Berlin

Breskin A, Chechik R, Ypsilantis T (eds) Proceedings of the third international workshop on ring imaging Cherenkov detectors. Nucl Instr Meth A 433(1–2):1–578

Bressan A, Dalla Torre S, Gobbo B, Tessarotto F (eds) (2008) Proceedings of the VI international workshop on ring imaging Cherenkov detectors. Nucl Instr Meth A 595:1–292

Breuker H et al (1987) Particle identification with the OPAL jet chamber in the region of the relativistic rise. Nucl Instr Meth A 260:329–342

Brückner W et al (1996) The transition radiation detector in the hyperon beam experiment WA89 at CERN. Nucl Instr Meth A 378:451–457

Cherenkov PA (1937) Visible radiation produced by electrons moving in a medium with velocities exceeding that of light. Phys Rev 52:378–379

Ekelöf T (ed) (1996) Proceedings of the second international workshop on ring imaging Cherenkov detectors. Nucl Instr Meth A 371(1–2):1–343

Ekelof T, Resvanis LK, Seguinot J (eds) Proceedings of the IV international workshop on ring imaging Cherenkov detectors. Nucl Instr Meth A 502(1):1–326

Engelfried J (2011) Cherenkov light imaging – Fundamentals and recent developments. Nucl Instr Meth A 639:1–6

Engelfried J et al (1998) The E781 (SELEX) RICH detector. Nucl Instr Meth A 409:439–442

Engelfried J et al (1999a) The SELEX phototube RICH detector. Nucl Instr Meth A 431:53–69 (hep-ex/9811001)

Engelfried J et al (1999b) The RICH detector of the SELEX experiment. Nucl Instr Meth A 433: 149–152

Engelfried J et al (2003) SELEX RICH performance and physics results. Nucl Instr Meth A 502:285–288

Engelfried J, Paic G (eds) (2005) Proceedings of the V international workshop on ring imaging Cherenkov detectors. Nucl Instr Meth A 553(1–2):1–380

Errede D et al (1989) Design and performance characteristics of the E769 beamline transition radiation detector. IEEE Trans Nucl Sci 36:106–111

Errede D et al (1991) Use of a transition radiation detector in a beam of high-energy hadrons. Nucl Instr Meth A 309:386–400

Fischer HM et al (1986) The OPAL jet chamber full scale prototype. Nucl Instr Meth A 252: 331–342

Frank I, Tamm I (1937) Coherent visible radiation of fast electrons passing through matter. CR Acad Sci URSS 14:109–114

Grupen C (1996) Particle detectors. Cambridge University Press, New York

Hallewell G et al (eds) (2011) Proceedings of the VII international workshop on ring imaging Cherenkov detectors. Nucl Instr Meth A (2011) (in print)

Heintze J (1982) The jet chamber of the JADE experiment. Nucl Instr Meth A 196:293–297

Heuer RD, Wagner A (1988) The OPAL jet chamber. Nucl Instr Meth A 265:11–19

Inami K et al (2006) A 5-ps TOF-counter with an MCP-PMT. Nucl Instr Meth A 560: 303–308

Klempt W (1999) Review of particle identification by time-of-flight techniques. Nucl Instr Meth A 433:542–553

Korpar S et al (2007) Proximity focusing RICH with TOF capabilities. Nucl Instr Meth A 572: 432–433

Nakamura K et al, Particle Data Group (2010) Review of particle physics. J Phys G: Nucl Part Phys 37: 075021. doi: 10.1088/0954-3899/37/7A/075021

Nappi E, Ypsilantis T (eds) (1994) Proceedings of the first workshop on ring imaging Cherenkov detectors. Nucl Instr Meth A 343(1):1–326

Nygren DR, Marx JN (1978) The time projection chamber. Phys Today 31:46

Olivito D, ATLAS Collaboration (2010) Performance of the ATLAS transition radiation tracker readout with cosmic rays and first high energy collisions at the LHC. JINST 5, C11006

Paul S (1991) Particle identification using transition radiation detectors. CERN-PPE-91-199, CERN, Geneva

Roberts A (1960) A new type of Cherenkov detector for the accurate measurement of particle velocity and direction. Nucl Instr Meth 9(1):55–66

Russ J (1995) Fermilab hyperon program: present and future plans. Nucl Phys A 585:39–47

Russ J et al (1987) A proposal to construct SELEX. Fermilab P781, http://lss.fnal.gov/cgi-bin/find_paper.pl?proposal-0781

Séguinot J, Ypsilantis T (1977) Photo-ionization and Cherenkov ring imaging. Nucl Instr Meth 142:377–391

Terentiev N et al. (1995) E781 beam transition radiation detector. SELEX internal Note H-746, http://www-selex.fnal.gov/

The ALICE Collaboration (1995) A large ion collider experiment – Technical proposal. CERN/LHCC-95-71, CERN, Geneva

The ALICE Collaboration (2000) TPC technical design report. CERN/LHCC 200001, CERN, Geneva

The ALICE Collaboration (2002) Addendum to TOF technical design report. CERN/LHCC 2002-016, CERN, Geneva

The ALICE Collaboration (2008) The ALICE experiment at the CERN LHC. JINST 3, S08002

The ALICE Collaboration (2010) The ALICE TPC, a large 3-dimensional tracking device with fast readout for ultra-high mulitiplicity events. Nucl Instr Meth A 622:316–367

Va'vra J (2011) PID techniques: alternatives to RICH method. Nucl Instr Meth A 639: 193–201

Va'vra J et al (2008) A high-resolution TOF detector: a possible way to compete with a RICH detector. Nucl Instr Meth A 595:270–273

7 Accelerators for Particle Physics

Helmut Burkhardt
CERN, Geneva, Switzerland

1 ***Introduction*** ***140***

2 ***Basic Concepts and Units*** ***140***

3 ***Magnet Lattice*** ***142***
3.1 Dispersion and Chromaticity 144

4 ***Sources and Pre-injectors*** ***145***

5 ***RF Acceleration*** ***145***

6 ***Ring Accelerators*** ***146***

7 ***Phase Stability*** ***147***
7.1 Applications of Accelerators 148

8 ***Fixed-Target Accelerators and Colliders*** ***149***

9 ***Energy and Luminosity*** ***151***

10 ***Vacuum and Beam Lifetime*** ***153***

11 ***Synchrotron Radiation*** ***154***

12 ***The Highest Energies*** ***155***

13 ***Conclusion*** ***157***

14 ***Cross-References*** ***157***

References ***158***

C. Grupen, I. Buvat (eds.), *Handbook of Particle Detection and Imaging*,
DOI 10.1007/978-3-642-13271-1_7, © Springer-Verlag Berlin Heidelberg 2012

Abstract: Beams of high-energy particles with well-defined properties are very important both for fundamental research and applied sciences. Particle accelerators are the devices that allow to produce these high-energy particle beams.

High-energy particle accelerators have a length of many kilometers and are the largest scientific tools used today. We give a short overview over the main types of accelerators, and in particular synchrotrons, storage rings, and linear accelerators, their main properties and fields of application.

The concepts and basic formulas are illustrated and discussed using the main parameters of the largest existing or planned high-energy accelerators.

1 Introduction

Particle accelerators are the devices used to provide high-energy charged particles with well-defined properties for research and medical and industrial applications.

This chapter starts with a description of the basic concepts. Electric fields are used to accelerate the charged particles, i.e., electrons, protons, or ions. Magnets are employed to guide the particles along the design path through the accelerators.

This is followed by sections on how particles are obtained from sources and on the advantages of using radio frequency rather than static electric fields for the acceleration.

The remainder of the text focuses mostly on high-energy accelerators for particles physics. In this context, the accelerator is often called the *machine*, meaning the device that provides the particles and the particles' collisions. The collisions are observed and studied by the *experiments*, which refer to the large detectors and collaboration of physicists who detect, study, and analyze the particle collisions.

2 Basic Concepts and Units

We will briefly recall basic laws and expressions that are important for accelerator physics, and refer to standard textbooks like Jackson (1998) for a more detailed discussion.

The force $\mathbf{F}$ acting on a particle with charge q in an electromagnetic field is

$$\mathbf{F} = q\,(\mathbf{E} + \mathbf{v} \times \mathbf{B}) \tag{1}$$

and generally referred to as Lorentz force. The term $q\,\mathbf{E}$ is the electric force and the term $q\,\mathbf{v} \times \mathbf{B}$ the magnetic force.

By convention, the electric field $\mathbf{E}$ points from positive to negative charges and is measured in units of volt per meter (V/m). $\mathbf{B}$ is the magnetic field measured in tesla (T); q is the electric charge, measured in coulomb (C). When we are dealing with elementary particles, it will be more practical to specify the charge in multiples of the elementary charge e, where $e = 1.6021765 \times 10^{-19}$ C.

The equations of motion in an electromagnetic field can be obtained by combining Eq. 1 with Newton's second law for the relativistic momentum,

$$\mathbf{F} = \dot{\mathbf{p}} = \frac{\mathrm{d}(m\gamma\mathbf{v})}{\mathrm{d}t}\,. \tag{2}$$

Particles in accelerators reach velocities v approaching the speed of light c and relativistic effects are of primary importance. Unless stated otherwise, we will always use the exact relativistic expressions.

Two standard dimensionless quantities of relativistic dynamics are the velocity in units of the speed of light $\beta = v/c$, and the Lorentz factor

$$\gamma = \frac{1}{\sqrt{1 - v^2/c^2}} = \frac{1}{\sqrt{1 - \beta^2}}\,. \tag{3}$$

The relativistic momentum of a particle with rest mass m is $\mathbf{p} = m\,\gamma\,\mathbf{v}$ and the total (relativistic) energy E_{tot} and kinetic energy E_{kin} are

$$E_{\text{tot}} = \gamma\, mc^2 = \sqrt{p^2c^2 + (mc^2)^2}\,, \qquad E_{\text{kin}} = (\gamma - 1)\, mc^2. \tag{4}$$

Depending on the value of γ we distinguish three different domains

1. $\gamma \approx 1$, nonrelativistic, $v \ll c$
2. $\gamma > 1$ relativistic
3. $\gamma \gg 1$ ultra-relativistic, $\beta \approx 1$

In the nonrelativistic case, we can expand ❯ Eqs. 3, 4 in powers of β and get as low-energy limit the familiar classical kinetic energy

$$E_{\text{kin}} \approx \frac{1}{2} mc^2\beta^2 = \frac{1}{2} mv^2.$$

In the ultrarelativistic case the mass becomes negligible, and we get $E_{\text{tot}} \approx E_{\text{kin}} \approx pc$.

We first consider the case of direct (DC) acceleration in a static electric field, which is parallel to the direction of motion as sketched in ❯ *Fig. 1*. A charged particle, in this case an electron of charge $q = -e$, is accelerated in the direction from the negatively charged cathode toward the positive anode. This type of acceleration is called direct high voltage or DC HV acceleration.

The energy gain U in an electric field of potential (voltage) V is

$$U = qV. \tag{5}$$

In accelerator and particle physics, it is convenient to use as energy unit the electron volt eV, where $1\,\text{eV} = 1.60217653 \times 10^{-19}$ joule, and to express the mass unit in terms of the energy unit.

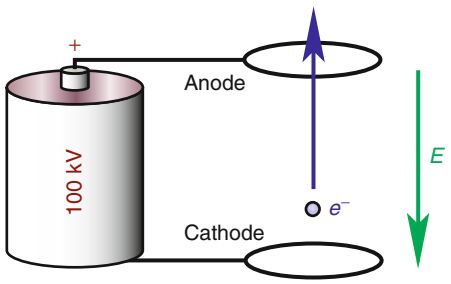


Fig. 1
Principle of acceleration by an electric field

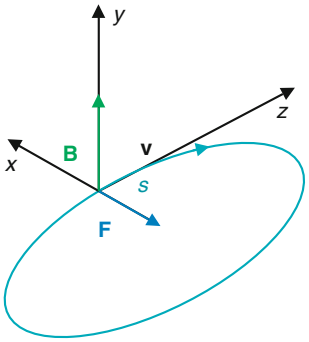


Fig. 2
Coordinates and motion in a magnetic field

For the electron, we have $m_e c^2 = 0.5109989\,\text{MeV}$ and for the proton $m_p c^2 = 938.272\,\text{MeV}$. As simple numerical example, we consider the accelerator sketched in *Fig. 1*, in which a single charged particle gains 100 keV = 0.1 MeV. If we start with an electron at rest, we would get to $E_{\text{tot}} = 0.611\,\text{MeV}$, with $\gamma = 1.1957$ and $\beta = 0.548$.

Static or quasi-static magnetic fields are used in accelerators to guide the particle motion along the design path. To illustrate this, we will look at a particle moving with velocity **v** along a path s, perpendicular to a homogeneous magnetic field **B** as shown in *Fig. 2*. The magnetic force is perpendicular to the direction of motion. It changes only the direction of the particle. The absolute value of the momentum and the energy are conserved. As a result, the particle moves on a circular path, with radius ρ, where

$$\rho = \frac{p}{qB}\,. \tag{6}$$

For $B = 1\,\text{T}$, $q = e$, and $p = 1\,\text{GeV}/c$ we get as radius $\rho = 3.336\,\text{m}$.

3 Magnet Lattice

We will now shortly describe the main principles of beam dynamics in accelerators. The aim here is to present a short, simple illustration of the basic formulas and effects. More thorough descriptions can be found in the literature. The basic principles were developed around the middle of the last century and summarized in the historic paper of Courant and Snyder (1958) and are well described in textbooks on accelerators (Lee 2004; Conte and MacKay 2008).

Quadrupole magnets provide fields that increase with distance from the center of the magnet. They are used to focus particles transversely to the beam direction. The combination of the dipoles used as bending magnets, quadrupoles and higher-order multipole magnets (sextupoles, octupoles, etc.) is called the magnetic lattice and determines the optical (guiding) properties for the particle beam. A standard configuration is the FODO lattice, sketched in *Fig. 3*. It consists of a sequence of F (horizontally focusing) quadrupoles, bending magnets ("O"), and D (horizontally de-focusing) quadrupole magnets. *Figure 3* also shows as dashed line a trajectory of a particle with a transverse offset. A quadrupole that focuses horizontally will

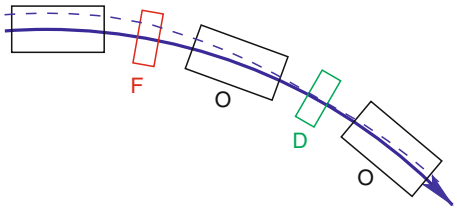


Fig. 3
Schematic view of a ring section with a FODO lattice

be de-focusing vertically and vice versa. The FODO lattice is an example of an alternating gradient lattice. The alternation between focusing and de-focusing elements has an overall focusing effect.

This can be qualitatively understood as follows. Particles are on average further off axis in F quadrupoles such that the focusing effect dominates. In one transverse (horizontal or vertical) plane, quadrupoles act like optical lenses. Two lenses with focal length f_1, f_2 at a distance D act together like a lens with focal length f, where $1/f = 1/f_1 + 1/f_2 - D/(f_1 f_2)$. In the alternating lattice, we have $f_2 = -f_1$, which together is focusing with $1/f = +D/f_1^2$.

The formulas given in this section are written in terms of the horizontal displacement x. The equivalent expressions also hold for the vertical displacement given by the y coordinate.

The basic equation of motion of particles in an accelerator lattice is of the form

$$x''(s) + k(s)\,x(s) = 0\,, \tag{7}$$

which is a second-order differential equation known as Mathieu–Hill equation from mathematical studies for mechanical applications in the nineteenth century (Mathieu 1868; Hill 1886). It can be regarded as generalized oscillator equation in which the restoring force k is a function of the independent variable s, the distance along the equilibrium orbit measured from some fixed reference point. $k(s)$ is the "restoring force" in the motion, given here by the focusing properties of the quadrupoles in the lattice. In a ring that is built up of regular cells with length L, k will be periodic, $k(L+s) = k(s)$. The total circumference C is the sum of the lengths of all cells.

The solution of Eq. 7 can be written as

$$x(s) = \sqrt{\epsilon\beta(s)}\cos(\mu(s) + \phi), \tag{8}$$

where ϵ and ϕ (initial phase) are integration constants determined by initial conditions. $\beta(s)$ is known as β function and describes the focusing properties of the magnetic lattice and

$$\mu(s) = \int_0^s \frac{\mathrm{d}s}{\beta(s)} \tag{9}$$

is the phase advance. The phase advance, divided by 2π, is also called the betatron wave number or tune $Q = \frac{\mu}{2\pi}$.

From Eq. 7 we can see that the amplitude of the oscillations is given by $\sqrt{\epsilon\beta(s)}$. This can be generalized to average quantities for a multiparticle beam. The r.m.s. beam size at the position s in the machine is

$$\sigma(s) = \sqrt{\epsilon\beta(s)}\,, \tag{10}$$

where ϵ is now an average beam quantity called emittance, which represents the beam size in phase space.

Collider optics are designed with minima of the β function in the collision regions. This decreases the beam sizes at the collisions points and increases the probability for collisions.

In an ideal machine, the transverse motion in the horizontal and vertical planes are independent. Such a machine is called fully decoupled. This allows for very different β functions and emittances in the horizontal and vertical plane. This is in particular important for electron rings to decrease the vertical emittance even in the presence of a large horizontal emittance from synchrotron radiation in the (horizontal) bending plane.

3.1 Dispersion and Chromaticity

We now discuss shortly how the optics changes with energy. More precisely, we consider the motion of particles with a relative momentum deviation $\Delta p/p$ from the average (design) momentum.

In a bending magnet, particles with a positive momentum deviation travel on a larger circle. For horizontal bending, we get a horizontal offset

$$\Delta x = D_x \frac{\Delta p}{p} . \tag{11}$$

The proportionality factor D_x is called the horizontal dispersion. It will be positive for dipoles and can be to some extent adjusted and reduced using quadrupoles. Nonzero dispersion also results in a coupling of the transverse and longitudinal motion in a ring: a particle at higher energy traveling on a larger radius will take a longer time to complete a full circle and appear shifted in time and longitudinal position compared to a particle with the average momentum.

Particles with a positive momentum deviation will also be less focused by quadrupoles, resulting in a decrease in tune,

$$\Delta Q = Q' \frac{\Delta p}{p} . \tag{12}$$

The proportionality factor is called chromaticity. The natural (quadrupole-generated) chromaticity is negative, $Q' < 0$, and roughly equal to the tune contribution from the regular arcs, $Q' \approx -Q$.

Higher-order magnets can be used to correct for aberrations by their feed-down effect, which is proportional to the transverse offset Δx from the magnet axis:

- An offset Δx in quadrupole results in a dipole (bending) magnet component proportional to the offset Δx.
- An offset Δx in a sextupole results in a focusing (quadrupole) magnet component proportional to the offset Δx.

Example: a sextupole magnet installed in the machine in a place with dispersion can provide extra focusing proportional to the offset Δx and compensate the negative natural chromaticity of the quadrupoles.

For machines with many FODO cells, the quadrupole strengths are often chosen such that the phase advance of each cell is close to a simple fraction of 2π, like $\mu_{\text{cell}} = \pi/2 = 90^\circ$ or $\mu_{\text{cell}} = \pi/3 = 60^\circ$. This results in rather periodic structures, which are easier to correct for aberrations.

4 Sources and Pre-injectors

A basic type of electron source is the thermionic electron gun. The principle is that of a cathode-ray tube. A cathode is placed in a vacuum tube and the electrons are extracted from the heated cathode and accelerated using DC HV acceleration. Other types of electron guns employ photocathodes.

Positrons can be produced by passing electrons of 50 MeV or more through a converter plate. In the plate, typically a metal like tungsten of 0.5–3 radiation length (Tsai 1974) thickness, the electrons will radiate hard photons by the process of bremsstrahlung. Many of these photons will generate electron–positron pairs by the process of pair creation. The efficiency to produce positrons depends only weakly on the initial electron energy and converter thickness (Nunan 1965). The converter is typically followed by solenoids and accelerator sections to focus and accelerate the positrons. The principle is sketched in ❯ *Fig. 4*.

For a proton source, as currently in use at CERN, the primary material is hydrogen gas, as commercially available for purposes like welding. A single bottle is sufficient to supply all the protons accelerated at CERN in 1 year. Hydrogen gas is injected in a metallic vessel, which is heated to ionize the hydrogen gas and placed on +90 kV tension to accelerate the protons toward the cathode at ground potential. A hole in the cathode allows to extract the protons that can then be further accelerated.

Further information on sources can be found in Scrivens (2003, 2004).

5 RF Acceleration

High-voltage breakdown limits static electric fields to roughly 1 MV/m. One or two orders higher acceleration gradients can be reached with oscillating fields using frequencies in the radio-frequency (RF) range of MHz to GHz.

RF acceleration is restricted to the acceleration of bunches (packets) of particles in which the bunch length is shorter than the RF wavelength.

Higher frequencies allow higher gradients. This can be understood qualitatively as follows: at a high frequency like 1 GHz, the peak voltage is only maintained for a fraction of 1 ns (nanosecond), which is too short to develop corona discharge and high-voltage breakdown.

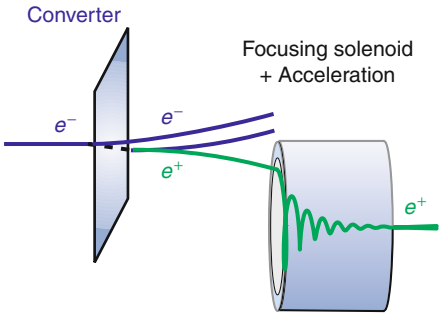


Fig. 4
Principle of a positron source

Table 1
Frequencies and gradients reached

Machine	f_{RF} (GHz)	Gradient (MV/m)
LEP	0.35	8
ILC	1.3	31.5
CLIC	12	100

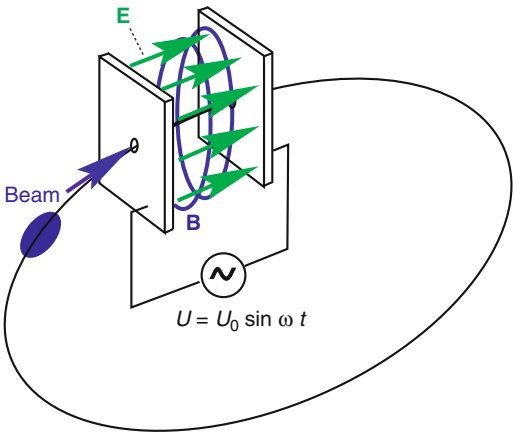


Fig. 5
RF acceleration in a ring

Table 1 shows actual numbers for the RF frequencies with the gradients used in LEP and proposed and tested in prototypes of accelerator structures for future linear colliders – International Linear Collider (ILC) and Compact Linear Collider (CLIC).

Future machines at the high-energy front will work in the tera scale (with collision energies of TeV). With a gradient of 100 MV/m a TeV could be reached with a 10-km long acceleration section.

Another advantage of RF over DC acceleration is that it becomes possible to sum up the energy gain over successive turns in ring accelerators as sketched in *Fig. 5*.

6 Ring Accelerators

In a cyclotron, the RF acceleration is performed at constant frequency and constant magnetic field. As the particle gains energy, the radius of curvature increases. This requires magnets as large as the whole ring, which is only practical for smaller machines.

In a synchrotron, the magnetic field is ramped up together with the energy of the particles, or more precisely, proportional to the increase in particle momentum, such that the radius of curvature is kept constant. All larger ring accelerators are of this type.

We take as example the Super Proton Synchrotron (SPS) at CERN. It can accelerate protons from $p = 14\,\mathrm{GeV}/c$ to $p = 450\,\mathrm{GeV}/c$ in a few seconds. The SPS has a circumference of

6,911.56 m and uses normal-conducting magnets. The RF system used to accelerate the protons is operated at a frequency of about 200 MHz at 10 MV RF voltage. The circumference is kept constant by increasing the RF frequency with the velocity. A change of 2×10^{-3} is sufficient, as the protons are already rather relativistic. See *Table 2* for numerical values.

7 Phase Stability

The particles in a bunch have a spread in energy and velocity.

To accelerate all particles and keep them bunched, it is important to provide more voltage than required on average for the energy gain. Slower, less energetic particles lagging behind then see a higher voltage and receive more acceleration as sketched in *Fig. 6*, which brings them closer to the bunch center. Particle A in *Fig. 6* corresponds to a particle at the center of a bunch. The principle is known as phase stability and was independently discovered by Veksler and McMillan (McMillan 1945; Veksler 1945).

At very high energies, all particles will travel practically at the same speed $v = c$; we can get in a situation, referred to as "above transition" in which the lower-energy particles travel on a shorter circumference and arrive before particle A at the RF. Phase stability is then achieved on the falling side of the sine wave.

To be more quantitative, we look how the traveling time changes with velocity, path length, and momentum. The time T needed to travel a fixed length L depends on the velocity v, where

Table 2

Proton parameters for the CERN SPS at the beginning and end of the acceleration used in fixed-target mode

p (GeV/c)	14	450
E (GeV)	14.0314	450.001
β	0.9977617	0.9999978
γ	14.9545	479.606
f_{RF} (MHz)	200.265	200.395
B (T)	0.063	2.025

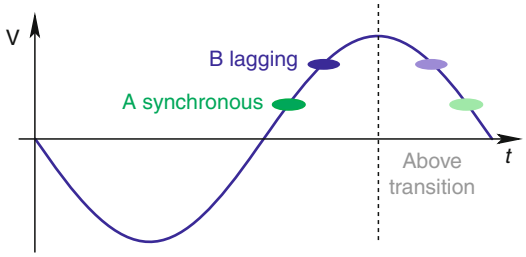


Fig. 6

Phase stability

$T = L/v$. An increase in velocity by Δv decreases the traveling time T, so that

$$\frac{\Delta T}{T} = -\frac{\Delta v}{v} . \tag{13}$$

The relativistic change of velocity v with momentum p is

$$\frac{\mathrm{d}v}{\mathrm{d}p} = \frac{1}{\gamma^2}\frac{v}{p} . \tag{14}$$

We also have to consider in a ring that particles with an energy offset will travel on a different path.

The relative change in path length L with momentum in a ring is called the momentum-compaction factor α_c, defined as

$$\alpha_c = \frac{\Delta L}{L} \Big/ \frac{\Delta p}{p} . \tag{15}$$

The momentum compaction depends on the focusing properties of the magnetic lattice. For a simple (FODO) lattice $\alpha_c \approx 1/Q^2$, where Q is the tune or number of betatron oscillations over the length considered. For the SPS ring we have $Q \approx 26.2$ and $\alpha_c = 1.92 \times 10^{-3}$.

We can now calculate the relative change in time required for one revolution in a ring by considering both the changes in path length and velocity. We get

$$\frac{\Delta T}{T} = \frac{\Delta L}{L} - \frac{\Delta v}{v} = \underbrace{\left(\alpha_c - \frac{1}{\gamma^2}\right)}_{\eta} \frac{\Delta p}{p} . \tag{16}$$

The expression in the bracket is known as phase-slip factor $\eta = \alpha_c - \gamma^{-2}$. During acceleration, both effects exactly cancel when $\gamma^{-2} = \alpha_c$. This is called transition and the corresponding γ called $\gamma_{\mathrm{tr}} = 1/\sqrt{\alpha_c}$. In the CERN SPS, the momentum-compaction factor is $\alpha_c = 1.92 \times 10^{-3}$, which corresponds to $\gamma_{\mathrm{tr}} = 22.8$. Looking at ❯ *Table 2*, we can see that this factor lies between the minimum and maximum γ, so that the transition is crossed during the acceleration. At transition, the phase stability is lost and particles will start to slowly de-bunch. The blowup can be minimized by fast transition crossing and by programming a phase jump in the RF at transition.

7.1 Applications of Accelerators

The concepts discussed so far were rather general and also apply to accelerators used in applications other than particle physics.

Worldwide, there are more than 20,000 accelerators in use. Compared to the high-energy particle accelerators most of these are very small machines used for industrial applications and medicine (Amaldi 2000). More information on accelerators for applications can be found in Chao and Chou (2010) and Greene and Williams (1997).

The remainder of this text is on the concepts that are more specifically of interest for applications in particle physics and in particular relevant for reaching high energies and rates in particle collisions.

8 Fixed-Target Accelerators and Colliders

We now distinguish between two types of accelerators depending on the use of the accelerated particles for high-energy physics.

The first type is the fixed-target accelerator, in which a beam of particles is extracted at the end of the acceleration to hit a target. The second type is the collider, in which two beams of high-energy particles are brought into collisions. Both types are illustrated in ❯ *Fig. 7* for ring accelerators. The same distinction also applies to linear accelerators.

The energy available in particle collisions to produce new particles is the center-of-mass energy $E_{\mathrm{CM}} = \sqrt{s}$, where s is the total four-momentum squared. It can conveniently be calculated using the 4-vector notation of high-energy physics (with units of $c = 1$). The energy/momentum 4-vector of beam 1 is $p_1 = (E_b, \mathbf{p})$. In case of a symmetric collider, the second beam has $p_2 = (E_b, -\mathbf{p})$. In the fixed-target case instead, the second (target) particle is at rest, $p_2 = (m_T, \mathbf{0})$ The four-momentum relations for the two cases are

$$\begin{aligned} \text{Collider}: \quad & p_1 = (E_b, \mathbf{p}_b), \quad p_2 = (E_b, -\mathbf{p}_b), \quad s = (p_1 + p_2)^2 = (2E_b)^2 \\ \text{Fixed target}: \quad & p_1 = (E_b, \mathbf{p}_b), \quad p_2 = (m_T, \mathbf{0}), \quad s = m_b^2 + m_T^2 + 2m_T E_b \end{aligned}$$

In the case of a symmetric collider, the sum of the two beam energies, $2E_b$, is available for new particle production. In the fixed-target case instead, the center-of-mass energy only increases with the square root of the beam energy ($E_{\mathrm{CM}} = \sqrt{2E_b\, m_T\, c^2}$, for "$E \gg m$"), while the rest is "lost" in kinetic energy of the secondary particles. ❯ *Figure 8* shows a comparison of the two cases for proton machines.

At the same beam energy, colliders allow for $\sqrt{2E_b/m_T}$ higher collision energies. For the LHC with protons at $E_b = 7$ TeV, the gain is a factor of 122. The difference is even more marked for collisions with the light electrons. All recent lepton particle accelerators were in fact built as colliders.

Fixed-target accelerators cannot compete with colliders at the energy front. They have instead other advantages which make them complementary to colliders at lower and medium energies. One important area of application for fixed-target proton accelerators is the production of secondary beams.

In symmetric e^+e^- or $p\overline{p}$ particle/antiparticle colliders, it is possible to keep both beams oppositely circulating in a single ring. Examples are the e^+e^- collider LEP that was operated at

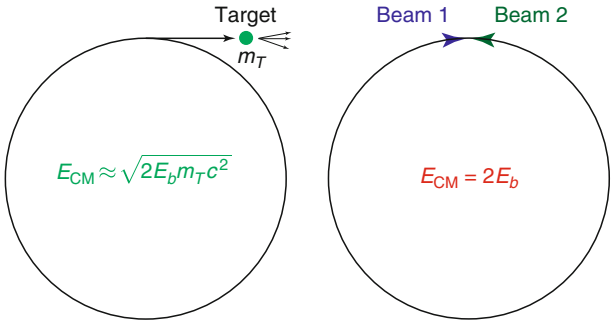


Fig. 7
Fixed-target and collider rings

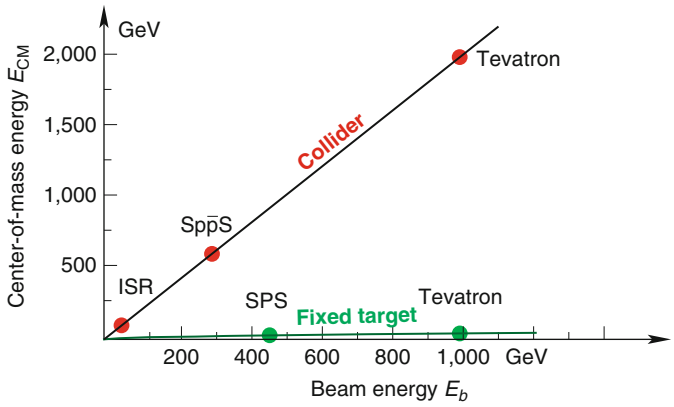


Fig. 8
Comparison of the center-of-mass energies available for collisions, as a function of the proton beam energy

CERN from 1989 to 2000 with up to E_b = 104 GeV and the Sp$\overline{\text{p}}$S (operated at CERN from 1981 to 1991, mostly at E_b = 270 GeV) and the TEVATRON proton–antiproton collider (operated at Fermilab since 1992 with up to E_b = 980 GeV).

To reach the center-of-mass energy of the LHC proton–proton collisions of $2E_b$ = 14 TeV in fixed-target mode would require a beam energy of 10^{17} eV, far beyond the reach of present accelerator technology and dimensions in the kilometer range. Energies up to 10^{20} eV have actually been observed in cosmic rays reaching the earth. This can be used to demonstrate that the collision energies that become available with the LHC are perfectly safe (Ellis et al. 2008).

Examples of symmetric colliders for identical particles using two neighboring rings with crossings in several interaction regions are the ISR operated from 1971 to 1984 and the LHC. Both were built at CERN, primarily as pp colliders. In the LHC, it is also possible to accelerate and collide heavy ions. With a moderate upgrade of the RF system, it would also be possible to collide different particle species like heavy ions and protons in the LHC.

It can be possible to use the same machine in both fixed-target and collider mode and to use several types of particles in the acceleration. An example is the CERN SPS. The SPS started operation in 1976 as fixed-target proton accelerator and was a few years later upgraded with an injection line for antiprotons and collision regions, to become the first proton–antiproton collider in the world. A few years later, it was upgraded with extra 352-MHz RF cavities to allow to accelerate electrons and positrons as injector for LEP. An upgrade of the 200-MHz RF system allowed to accept a larger range of revolution frequencies to also accelerate heavy ions in the SPS.

Once the required hardware is installed, the switching from one mode to the other is mainly a question of controlling and adjusting the RF and magnetic parameters. This can in principle be automated. Switching the SPS from proton to ion operation is done manually, typically within a couple of days. Switching from protons to e^+, e^- acceleration was automated and possible within seconds.

9 Energy and Luminosity

Energy and luminosity are the most important performance parameters of an accelerator for particle physics.

Higher energy is required to allow to produce new, heavier particles.

High luminosity is needed to observe rare processes and for precision measurements.

The demand for higher energies and luminosities has triggered technological developments, both for lepton as well as for proton colliders; new technologies have found their way into accelerators, such as superconductivity both for high-current magnets and radio-frequency accelerating systems. This and a steady advance in the understanding of beam dynamics, computing techniques, beam diagnostics, and beam control have made this possible.

Figure 9 shows the increase in energy over the years. This graph shows both lepton and hadron machines and also the electron–proton collider HERA. To make proton and electron machines more comparable in their discovery potential, the energy per proton (consisting of three quarks and several gluons) was divided by 3. The dashed lines show the impressive progress in maximum energy over the years: an exponential growth with a factor of 4 every 10 years over 4 decades!

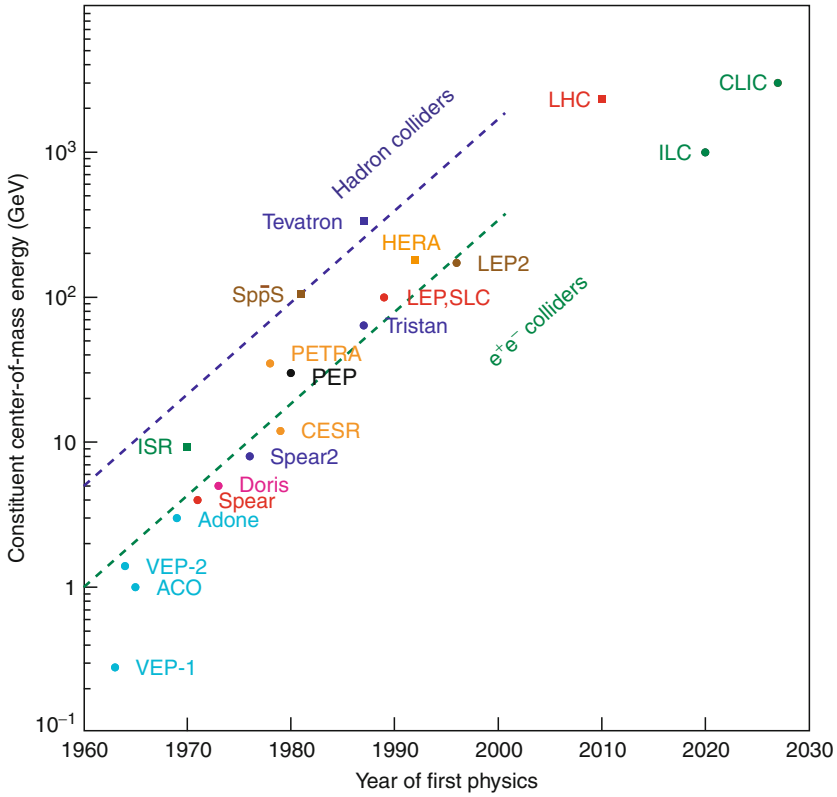


Fig. 9
Growth in collider energies with time

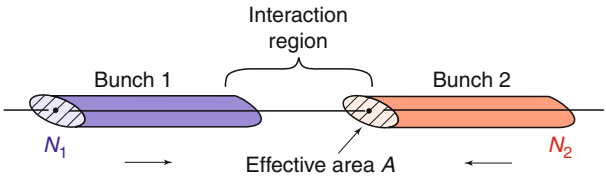


Fig. 10
Luminosity from particles flux and geometry

Proton and lepton accelerators complement each other. At comparable size and cost, protons allow for higher energies. They can be considered as "discovery machines." Lepton colliders are the precision instruments to study the details of the interactions between particles.

Accelerating few particles to very high energy is not all what is required. What is about equally important is a high luminosity, i.e., to allow for a high flux of particles resulting in a sufficiently high number of collisions.

The collision rate $\dot{n}$ for a process of cross section σ is the product of the luminosity $\mathcal{L}$ and the cross section

$$\dot{n} = \mathcal{L}\sigma. \tag{17}$$

The luminosity of a collider is determined by the particle flux and geometry (Herr and Muratori 2003). For head-on collisions as illustrated in Fig. 10, we have that

$$\mathcal{L} = \frac{N_1 N_2 k_b f}{A}. \tag{18}$$

N_1, N_2 are the numbers of particles per bunch, k_b is the number of bunches (per train in case of a linear collider), f the revolution frequency in case of a ring, and the bunch-train crossing frequency in case of a linear collider, and A the effective beam overlap cross section at the interaction point. For beams with Gaussian shape of horizontal and vertical r.m.s. beams sizes σ_x, σ_y colliding head on, we have $A = 4\pi\,\sigma_x\sigma_y$, where $\sigma = \sqrt{\epsilon\beta}$ in both x, y, according to Eq. 10. Strong quadrupole magnets are used around the interaction regions to focus beams down to small values of the β functions at the interaction point (called β^*) to get small beam sizes and high luminosity.

Cross sections are usually given in units of barn (symbol b), where $1\,\mathrm{b} = 10^{-24}\,\mathrm{cm}^2 = 10^{-28}\,\mathrm{m}^2$.

Typical luminosities of LEP, the highest-energy lepton collider built, were $10^{31}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$. The cross section for $e^+e^- \to Z$ is $\sigma_Z \approx 30\,\mathrm{nb}$. Over four million Z events were produced for each of the four experiments installed at LEP.

The cross section of protons in proton–proton collisions is $\sigma_{pp} \approx 50\,\mathrm{mb}$ (at $\sqrt{s} = 10\,\mathrm{GeV}$ and about 2× higher at $\sqrt{s} = 10\,\mathrm{TeV}$). Cross sections for the "most interesting processes" like new particle production in e^+e^- or quark–antiquark annihilation can be very small and generally decrease with the center-of-mass energy squared. For Higgs particle production at the LHC, the relevant order of magnitude for cross sections is femtobarn ($1\,\mathrm{fb} = 10^{-39}\,\mathrm{cm}^2$). The LHC is designed for a very high luminosity of $\mathcal{L} = 10^{34}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$, so that the production rate for rare processes with $\sigma = 1\,\mathrm{fb}$ would still be $\dot{n} = 10^{-5}\,\mathrm{s}^{-1}$ or one in 28 h.

A number of second- and third-generation high-luminosity e^+e^- colliders have been built over the last 2 decades in the medium-energy range (1–10 GeV in the center of mass). They were typically built to operate at a fixed energy that corresponds to the mass of one of the ϕ, ψ, or Υ

resonances to allow to produce mesons with s, c, or b quarks in large quantities. Such machines are also called e^+e^- factories.

The last generation of b-factories, PEP2 at SLAC in the US and KEKB at KEK in Japan both reached peak luminosities exceeding $\mathcal{L} = 10^{34}\ \mathrm{cm}^{-2}\,\mathrm{s}^{-1}$. This was achieved by focusing the beams to micrometer beam sizes at the interaction points and by colliding many ($k_b >$ 1,000) bunches. Details about the past, present, and future e^+e^- colliders at the high-luminosity frontier can be found in Biagini (2009).

10 Vacuum and Beam Lifetime

The particles in accelerators travel in evacuated beam pipes. Beam pipes are typically made of stainless steel or aluminum and have an elliptical cross section of some cm^2. Collisions of the beam with the rest gas are sketched in *Fig. 11* and result in unwanted effects like blowup of the beam size, generation of a beam halo, and loss of particles from the beam pipe causing radiation and backgrounds to the experiments.

Good vacuum conditions in the beam pipes of accelerators are important to minimize these unwanted effects.

Typical numbers are: a good vacuum is in the range of nanoTorr ($p = 1\,\mathrm{nTorr} = 1.33 \times 10^{-7}$ Pa), which at room temperature corresponds to a rest-gas density of

$$\rho_m = \frac{p}{kT} = 3.26 \times 10^{13}\ \mathrm{molecules/m}^3. \tag{19}$$

A typical cross section in beam–gas scattering for electron beams is $\sigma = 6$ barn and corresponds to a cross section for bremsstrahlung with an energy loss of >1% in a rest gas of CO or N_2 molecules. The collision probability is $P_{\mathrm{coll}} = \sigma\,\rho_m = 1.96 \times 10^{-14}/\mathrm{m}$. We can multiply this with the particle velocity $v \approx c$ (for high-energy accelerators) to obtain the loss rate with time. The inverse of this is the electron-beam lifetime from beam–gas scattering at a rest-gas pressure of 1 nTorr,

$$\tau = \frac{1}{P_{\mathrm{coll}}\,c} = 1.7 \times 10^5\ \mathrm{s} = 47\,\mathrm{h}\,. \tag{20}$$

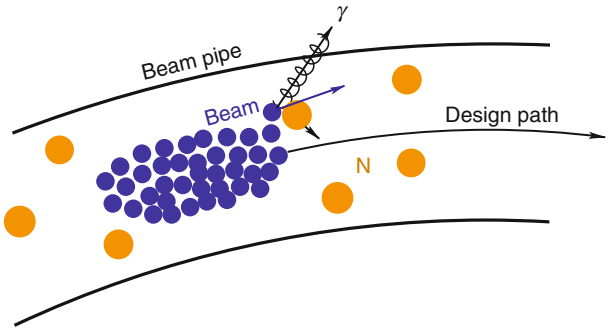


Fig. 11
Beam collisions with the nuclei (N) of the rest gas

A long beam lifetime is particularly important for colliders. Only a very small fraction of the beam particles will actually collide at each beam crossing and leave the beam pipe. The beams can often be kept circulating for several hours before the intensity and luminosity has reduced by a significant amount. Ring accelerators with a long beam lifetime are also called storage rings.

11 Synchrotron Radiation

Generally, radiation is emitted by any accelerated charge. For highly relativistic particles in accelerators this is referred to as synchrotron radiation. As shown below, the synchrotron radiation remains negligible in linear acceleration but becomes very significant when high-energy electron beams are deflected by magnetic fields. For accelerators for particle physics, synchrotron radiation can be considered mostly as an unwanted effect. The energy loss has to be compensated by the acceleration system and the radiation results in heating of accelerator components and backgrounds to the particle detectors. The combination of the energy loss by synchrotron radiation and the acceleration by the RF system, however, also has a positive effect, which is that it results in a damping of the transverse motion that limits the transverse beam size. This can be qualitatively understood as follows: the synchrotron radiation is emitted in the direction of the motion, which reduces the particle momentum in all three x, y, z components. The acceleration by the electrical field in z direction only compensates the energy loss of the particle in z direction. If the energy loss would be continuous, the transverse beam sizes would shrink to zero. The quantum fluctuations of the synchrotron radiation emitted in discrete steps results in a small, finite transverse beam emittance.

The high achievable power, and the rather unique properties of synchrotron radiation, which allow for very short energetic pulses and the possibility to cover a broad range in the ultraviolet and X-ray spectrum, make synchrotron radiation very attractive for applied research. Many electron machines have been built as dedicated synchrotron light sources.

The properties of synchrotron radiation are well understood and described in textbooks (Sokolov 1986; Jackson 1998 and Hofmann 2004).

As discussed by Schwinger (1949) in his classical paper, the power radiated by an accelerated particle of charge e is described by the relativistic version of Lamor's formula,

$$P = \frac{e^2\gamma^2}{6\pi\epsilon_0 m^2 c^3}\left(\dot{\mathbf{p}}^2 - \beta^2\dot{p}^2\right), \tag{21}$$

where $\dot{\mathbf{p}}$ and $\dot{p}$ are the time derivatives of the particle's momentum vector and absolute value, m the mass of the particle, and $\beta = v/c$ and $\gamma = (1-\beta^2)^{-1/2}$ the usual Lorentz quantities.

We will now consider the two opposite cases of

- Acceleration in the direction of motion as relevant for linear accelerators
- Acceleration perpendicular to the motion as relevant in rings

In the first case of linear acceleration with $\mathbf{v} \| \dot{\mathbf{v}}$ we have no change in direction, only in magnitude,

$$\left(\frac{\mathrm{d}\mathbf{p}}{\mathrm{d}t}\right)^2 = \left(\frac{\mathrm{d}p}{\mathrm{d}t}\right)^2,$$

so that

$$\left(\frac{\mathrm{d}\mathbf{p}}{\mathrm{d}t}\right)^2 - \beta^2\left(\frac{\mathrm{d}p}{\mathrm{d}t}\right)^2 = \dot{p}^2(1-\beta^2) = \frac{\dot{p}^2}{\gamma^2} \quad \text{and} \quad P = \frac{e^2}{6\pi\epsilon_0 m^2 c^3}\,\dot{p}^2. \tag{22}$$

The two terms nearly cancel, resulting in a suppression by a factor of γ^2. As numerical example, we take the highest acceleration gradient of 100 MV/m from ❯ *Table 1* and find, that the power loss is only 11 keV/s or 0.4 eV loss for a 1 TeV, 10 km long CLIC like machine. Synchrotron radiation in linear acceleration is negligible.

Now the second case, of motion on a circular path in a ring. At constant energy, the motion on a circular path in a ring implies that we have an acceleration perpendicular to the velocity, $\mathbf{v} \perp \dot{\mathbf{v}}$. The second term is zero (at constant energy the magnitude of the momentum is also constant). There is no cancellation and we get a significant effect:

$$\left(\frac{\mathrm{d}\mathbf{p}}{\mathrm{d}t}\right)^2 - \beta^2\underbrace{\left(\frac{\mathrm{d}p}{\mathrm{d}t}\right)^2}_{0} = \dot{\mathbf{p}}^2 \quad \text{and} \quad P = \frac{e^2}{6\pi\epsilon_0 m^2 c^3}\,\gamma^2\,\dot{\mathbf{p}}^2. \tag{23}$$

For the circular motion in the uniform magnetic field we have from the Lorentz force and Newton's law that

$$F = |\dot{\mathbf{p}}| = evB = \frac{vp}{\rho} = \frac{m\gamma v^2}{\rho}. \tag{24}$$

We find that the power radiated by circular motion in a uniform magnetic field increases with the fourth power of γ,

$$P = \frac{e^2 v^4}{6\pi\epsilon_0 c^3\,\rho^2}\,\gamma^4. \tag{25}$$

We multiply this with the time it takes to complete one turn, $T = 2\pi\rho/v$, and find that the energy loss of a particle by synchrotron radiation over one turn is

$$U_0 = \frac{e^2}{3\epsilon_0}\frac{\beta^3\gamma^4}{\rho} \propto \frac{1}{\rho}\frac{E^4}{m^4}, \qquad \text{where} \quad \frac{e^2}{3\epsilon_0} = 6.032\times10^{-9}\ \text{eV m}. \tag{26}$$

A practical limit was reached with electrons at LEP at beam energies around 100 GeV, corresponding to a Lorentz factor $\gamma \approx 2\times10^5$ when 3% of the particle energy was lost on a single turn. More details and further references can be found in review articles on LEP (Assmann et al. 2002; Bailey et al. 2002; Häubner 2004; and Burkhardt and Jowett 2009). Accelerator physics aspects are summarized in Brandt et al. (2000) and the RF system in Butterworth et al. (2008).

Colliders for the TeV range require either the use of heavier particles like protons in a ring or the use of linear colliders.

12 The Highest Energies

The world's largest and highest-energy particle accelerator today is the LHC at CERN. It is installed in the 26.7 km long tunnel used previously by LEP. The LHC is built with two beam pipes that cross at four interaction regions. This allows to accelerate and collide particles of the same charge, *pp* (proton on proton) and heavy ions. The mass of these particles is much higher (1,836 times in case of protons) than that of electrons. Synchrotron radiation from protons at LHC energies becomes noticeable but is not yet a limitation. The maximum beam energy in the

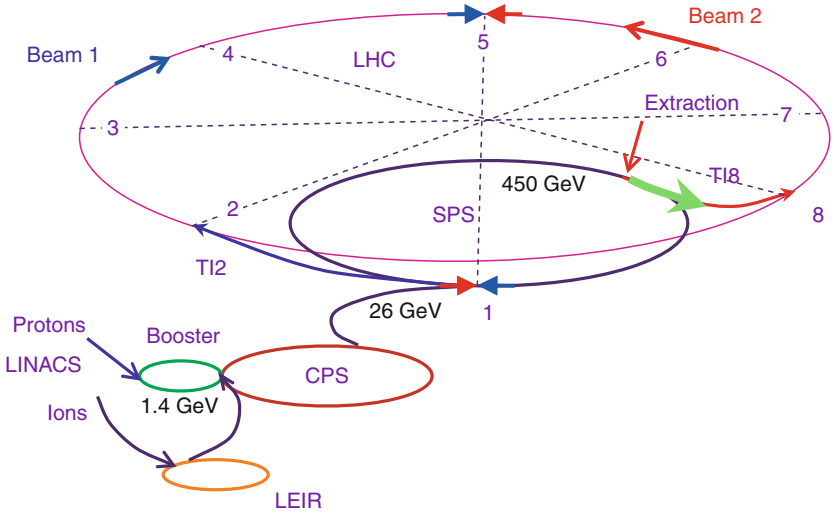


Fig. 12
Schematic view of the LHC with its injectors

LHC, or more precisely the beam momentum p, is given by the maximum bending field strength B according to Eq. 6 and the bending radius $\rho = 2{,}804$ m, which is constrained by the tunnel geometry. The LHC utilizes superconducting NbTi magnets operated at superfluid-helium temperature of 1.9 K, which allows for fields up to $B = 8.33$ T and $p = 7\,\text{TeV}/c$.

Filling the LHC with protons or ions requires a chain of pre-accelerators. The first stage is a linear accelerator. It is followed by several synchrotrons, the Booster, PS and SPS rings, see *Fig. 12*.

To reach the design peak luminosity of the LHC of $10^{34}\,\text{cm}^{-2}\,\text{s}^{-1}$ is very challenging and requires collisions of many bunches, stored in each of the two rings (by design 2,808 bunches). Another major challenge for the LHC is that the total energy stored in the beams and the magnets is very high. An uncontrolled loss of the beam or the superconductivity ("quench") could result in damage of parts of the accelerator.

A current of 12 kA is required to reach the design field of 8.33 T in the LHC dipoles. The inductance of an LHC dipole is $L = 100$ mH. The energy stored in a single dipole at full current is $I^2\,L/2 = 7.2$ MJ, which adds up to 9 GJ for the 1,232 dipole magnets in the LHC. This is comparable to the kinetic energy of a large passenger plane at full speed.

The total energy in the beam reaches 360 MJ. This is several orders of magnitude higher than in other machines and well above the damage level for uncontrolled beam loss. The LHC is equipped with a fast beam protection system. Beam losses are monitored by several thousand monitors all around the ring. In case of magnet trips or abnormal beam losses, the LHC beams will be dumped within a few turns.

The LHC is commissioned in steps. First beams were injected in the LHC in 2008. First collisions at the injection energy of 2×450 GeV were obtained in 2009. Operation at a physics energy of 2×3.5 TeV started in spring 2010.

Further information on the LHC can be found in Evans (2009) and Evans and Bryant (2008).

Plans for future linear e^+e^- colliders are being studied in detail along two main paths, the International Linear Collider (ILC) and Compact Linear Collider (CLIC).

The ILC aims for a beam collision energy of 0.5 TeV, upgradeable to 1 TeV in the center of mass, whereas CLIC extends the linear-collider energy reach into the multi-TeV range, nominally 3 TeV, which leads to different technologies. ILC is based on superconducting RF acceleration technology with high RF-to-beam efficiency, CLIC takes advantage of a novel scheme of two-beam acceleration with normal-conducting copper cavities at high frequency and high accelerating field.

More information about the ILC can be found on the ILC web page (http://www.linearcollider.org/cms/) and the ILC Reference Design Report (ILC 2007). Information about CLIC and CLIC design and parameters can be found on the CLIC web pages (http://clic-study.web.cern.ch/CLIC-Study/ and http://clic-study.web.cern.ch/CLIC-Study/Design.htm).

13 Conclusion

The availability of particle beams with well-defined properties from particle accelerators is crucial for most of the methods and techniques described in this handbook. Even detectors built to observe particles from natural sources for astrophysics and space instrumentation often rely on accelerators for testing and calibration of their detectors.

The basic concepts and types of particle accelerators are described in this chapter.

The development of particle accelerators has been to a large extent driven by requirements of particle physics research for higher energies and intensities. Smaller, lower-energy particle accelerators are used in many scientific, industrial, and medical applications.

14 Cross-References

In this chapter, we describe the principles of accelerators which are used to provide particles and particle collisions with well-defined properties.

The methods and techniques used to detect, study, and visualize particles are described in detail the following chapters: ❯ Chaps. 4, "Data Analysis," ❯ 6, "Particle Identifiaction," from ❯ 11, "Gaseous Detectors" to ❯ 20, "Calorimeters," and ❯ 31, "Neutron Detection."

Some of these chapters deal with specific applications of particle accelerators: Neutron physics with a neutron spallation source (❯ Chaps. 30, "Spallation – Neutrons Beyond Nuclear Fission" and ❯ 31, "Neutron Detection") requires high-intensity proton beams from medium-energy proton accelerators. Many of the more applied studies employing particle beams and detectors for art (❯ Chap. 34, "Radiation Detectors and Art"), archaeology (❯ Chap. 26, "Accelerator Mass Spectrometry and its Applications in Archaeology, Geology and Environmental Research"), and materials science (❯ Chap. 29, "Particle Detectors in Materials Science") can already be done with smaller, more compact lower-energy accelerators. The techniques used for studies of synchrotron radiation and FEL instrumentation are described in a specific chapter (❯ Chap. 8, "Synchrotron Radiation and FEL Instrumentation").

References

Amaldi U (2000) The importance of particle accelerators. Europhys News 31N6:5–9

Assmann R, Lamont M, Myers S (2002) A brief history of the lep collider. Nucl Phys Proc Suppl 109B:17–31. CERN-SL-2002-009

Bailey R et al (2002) The LEP collider. C R Acad Sci (Paris) 9:1107–1120

Biagini M (ed) (2009) e^+e^- colliders: past and present experiences and future frontiers. ICFA Beam Dyn Newsl 48:23–278

Brandt D, Burkhardt H, Lamont M, Myers S, Wenninger J (2000) Accelerator physics at LEP. Rep Prog Phys 63:939

Burkhardt H, Jowett JM (2009) A retrospective on LEP. ICFA Beam Dyn Newsl 48:143–152

Butterworth A et al (2008) The LEP2 superconducting RF system. Nucl Instrum Method A587: 151–177

Chao AW, Chou W (2010) Reviews of accelerator science and technology: medical applications of accelerators, vol 2. World Scientific, Singapore

Conte M, MacKay WW (2008) An introduction of particle accelerators. World Scientific, Singapore

Courant E, Snyder H (1958) Theory of the alternating-gradient synchrotron. Ann Phys 3:1

Ellis JR, Giudice G, Mangano ML, Tkachev I, Wiedemann U (2008) Review of the Safety of LHC collisions. J Phys G35:115004

Evans L (2009) The large hadron collider: a marvel of technology. EPFL Press, Lausanne

Evans L, Bryant P (eds) (2008) LHC machine. J Instrum 3:S08001

Greene D, Williams PC (1997) Linear accelerators for radiation therapy (medical science). Taylor & Francis, New York, NY

Häubner K (2004) Designing and building LEP. Phys Rep 403–404:177–188

Herr W, Muratori B (15–26 Sep 2003) Concept of luminosity, CAS – CERN accelerator school: Intermediate course on accelerator Physics. Zeuthen, Germany, pp 361–378. http://cdsweb.cern.ch/record/941318

Hill GW (1886) On the part of the motion of lunar perigee which is a function of the mean motions of the Sun and Moon. Acta Math 8:1–36

Hofmann A (2004) The physics of synchrotron radiation. Cambridge University Press, Cambridge, UK

ILC (2007) ILC reference design report. http://www.linearcollider.org/cms/?pid=1000437

Jackson JD (1998) Classical electrodynamics, 3rd edn. Wiley, New York

Lee SY (2004) Accelerator physics. World Scientific, Singapore

Mathieu E (1868) Mémoire sur le mouvement vibratoire d'une membrane de forme elliptique. Journal des Mathématiques Pures et Appliquées 13:137–203

McMillan EM (1945) The synchrotron—A proposed high energy particle accelerator. Phys Rev 68(5–6):143–144

Nunan CS (1965) A positron linear accelerator design. Proc Pac IEEE Trans Nucl Sci 12(3):465

Schwinger J (1949) On the classical radiation of accelerated electrons. Phys Rev 75:1912. http://link.aps.org/doi/10.1103/PhysRev.75.1912

Scrivens R (2003) Electron and ion sources for particle accelerators. CAS 2003, CERN-2006-002

Scrivens R (2004) Proton and ion sources for high intensity accelerators. Proc EPAC 2004 and CERN-AB-2004-075, http://cdsweb.cern.ch/record/793626

Sokolov AA, Ternov IM (1986) Radiation from relativistic electrons. American Institute of Physics, New York

Tsai Y-S (1974) Pair production and Bremsstrahlung of charged leptons. Rev Mod Phys 46:815–851

Veksler V (1945) Concerning some new methods of acceleration of relativistic particles. J Phys USSR 9:153

8 Synchrotron Radiation and FEL Instrumentation

Shaukat Khan · Klaus Wille
Technische Universität Dortmund, Dortmund, Germany

1 A Brief History of Radiation Sources . . . 161

2 Radiation from Accelerated Electrons . . . 162

3 Generation of Synchrotron Radiation . . . 163
3.1 Acceleration of Electrons to Ultrarelativistic Energies . . . 164
3.1.1 Conventional Electron Linacs . . . 165
3.1.2 Superconducting Linacs and Energy Recovery . . . 165
3.1.3 Synchrotrons . . . 167
3.1.4 Electron Storage Rings . . . 168
3.1.5 Electron Beam Optics . . . 170
3.1.6 Radiation Effects . . . 172
3.2 Insertion Devices . . . 174
3.2.1 Wavelength Shifters and Superbends . . . 174
3.2.2 Wigglers and Undulators . . . 175
3.3 Synchrotron Radiation Sources Worldwide . . . 176

4 Applications of Synchrotron Radiation . . . 176
4.1 Diffraction . . . 177
4.2 Spectroscopy . . . 178
4.3 Imaging . . . 179
4.4 Other Applications . . . 180
4.4.1 Time-Resolved Studies . . . 180
4.4.2 Far-Infrared Radiation . . . 181
4.4.3 X-Ray Holography . . . 181
4.4.4 Metrology . . . 181
4.4.5 X-Ray Lithography . . . 181

5 The Next Generation . . . 182
5.1 Storage Rings . . . 182
5.2 Linac-Based Free-Electron Lasers . . . 182
5.3 Energy-Recovery Linacs . . . 183

C. Grupen, I. Buvat (eds.), *Handbook of Particle Detection and Imaging*,
DOI 10.1007/978-3-642-13271-1_8, © Springer-Verlag Berlin Heidelberg 2012

6 Conclusions *184*

References *184*

Further Reading *185*

Abstract: For almost one century, x-rays have been the primary tool to probe the atomic structure of matter. With the advent of synchrotron radiation sources in the 1960s and more recently free-electron lasers, the photon flux, coherence, spectral brightness, and tunability of short-wavelength radiation has improved dramatically. After briefly reviewing the history of x-ray sources, the generation of radiation by accelerating electrons will be addressed. Synchrotron radiation is produced by circular acceleration of relativistic electrons in magnetic fields. Therefore, the discussion focuses on linear and circular particle accelerators, on the principles of particle optics as well as on magnetic devices called wigglers and undulators. After giving a brief overview of the applications of synchrotron radiation, newly emerging radiation sources, in particular free-electron lasers, will be discussed. It will become clear that x-ray science is far from settling into a routine but is presently undergoing a more rapid development than ever.

1 A Brief History of Radiation Sources

On the evening of the 8th of November 1895, W. C. Röntgen discovered that a discharge tube which he had wrapped in black cardboard caused a faint glow on a nearby fluorescence screen, and he recognized this to be a new kind of radiation, which penetrated opaque material like cardboard or wood, and he could even "see" the bones inside his hand (Röntgen 1895). These x-rays, as Röntgen named them (also called Röntgen rays in German), were not bent by magnetic fields (unlike cathode rays, i.e., electrons) but were also not noticeably deflected by a prism (unlike visible light). Röntgen did speculate that they might be ultraviolet light or even "aether waves," which were believed to exist at that time. The fact that x-rays were indeed electromagnetic radiation with short wavelength became only clear after 1912, when other scientists like M. von Laue, W. H. Bragg and his son W. L. Bragg, P. Debye, P. Scherrer, and others started to exploit another stunning property of x-rays, namely, their ability to reveal the structure of crystalline matter on the Ångström scale (1 Å = 10^{-10} m) by diffraction.

Much more efficient than Röntgen's discharge tube is the x-ray tube as we know it today, which was invented in 1913 by W. D. Coolidge at General Electric in New York (Coolidge 1917). Here, electrons are generated by a heated cathode and accelerated toward an anode, where their sudden deceleration produces *bremsstrahlung* with a continuous spectrum and, in addition, a line spectrum caused by fluorescence. With rotating anodes to provide better cooling, x-ray tubes reached a brightness six orders of magnitude higher than that of Röntgen's original apparatus.

While x-rays became an indispensable tool in biology and condensed matter physics, the need for higher intensity but also for monochromatic radiation with tunable wavelength and small divergence arose. With the advent of particle accelerators using radiofrequency (closely linked to the progress in radar technology in World War II), electrons could be accelerated to ultrarelativistic energies. When forced on a circular trajectory by a magnetic field, electrons produce radiation tangentially to their trajectory within a narrow opening angle and over a broad spectral range. This type of radiation is called synchrotron radiation. Forty years after its prediction, synchrotron radiation was for the first time directly observed and photographed in 1947 at the General Electric synchrotron in Schenectady/NY, which happened to have a vacuum vessel made of glass (Elder et al. 1947).

Even though it gave synchrotron radiation its name, a synchrotron would be a poor radiation source since it ramps the electron energy E up on a macroscopic timescale, while the radiation power increases with E^4 and becomes only significant toward the end of the ramp cycle. Therefore, synchrotron radiation facilities are not really synchrotrons but always electron (or positron) storage rings in which the electron energy is kept constant.

As a first generation of synchrotron radiation sources, scientists started in the 1960s to use radiation from e^+e^- colliders parasitically. The second generation emerged in the 1970s when electron storage rings dedicated to this purpose were built. Without counter-propagating positrons, the electron beam emittance (as defined below) could be reduced, leading to radiation with smaller source size and divergence. Third-generation radiation sources, built since the 1990s, are larger storage rings with even smaller beam emittance and a large number of so-called insertion devices, particularly undulators. The purpose of these devices is to produce more intense radiation compared to radiation from a simple circular path. In undulators, electrons move on a sinusoidal trajectory emitting radiation with a reduced opening angle and a line spectrum consisting of a fundamental wavelength and higher harmonics thereof.

The label "fourth-generation light source" is nowadays claimed by several types of facilities. The ultimate x-ray source based on storage rings should have a high beam energy to reach shorter wavelength and a large circumference, allowing for smaller emittance. In 2009, PETRA III at DESY in Hamburg/Germany was commissioned and is currently the largest storage-ring-based light source with a circumference of 2,304 m (Balewski 2010). Free-electron lasers (FELs) based on linear accelerators have now reached subvisible wavelengths, from 108 nm in 2000 at DESY in Hamburg down to 1.5 Å in 2009 at SLAC in Menlo Park/USA (Ding et al. 2009). FELs using conventional linear accelerators provide ultrashort and extremely intense radiation pulses, but are limited to low repetition rate. Superconducting energy-recovery linear accelerators (ERLs), on the other hand, are envisaged as complementary sources with a large repetition rate and ultrashort pulse duration but with lower intensity than FELs. Existing ERLs produce radiation in the infrared regime but several x-ray facilities have been proposed (Bilderback et al. 2009).

In addition to synchrotron radiation sources based on conventional accelerator technology, "table-top" sources driven by laser-plasma accelerators have been proposed. Even though these accelerators have not yet reached the required level of stability, remarkable progress has been made in this field (Fuchs et al. 2009).

The history of x-ray sources is best illustrated by the increase of their radiation intensity over the years. In ❯ *Fig. 1*, the peak brilliance (defined below in ❯ Eq. 5) is shown from the year 1895 until today. The peak brilliance has not only increased by more than 25 orders of magnitude but the slope has also become steeper with the advent of synchrotron light sources and is steepening again with the first x-ray FELs.

2 Radiation from Accelerated Electrons

When an electron is at rest or moves with constant speed in free space, the electric field lines point to the electron. Now, let the electron be accelerated for a short time. If the field lines would instantaneously follow that acceleration, there would be no radiation. Given the finite speed of light, however, a distant observer will see the field lines still pointing to a spot where the electron would be had it not been accelerated. A nearby observer, on the other hand, will see the field lines pointing toward the electron. The transition between the two regimes takes

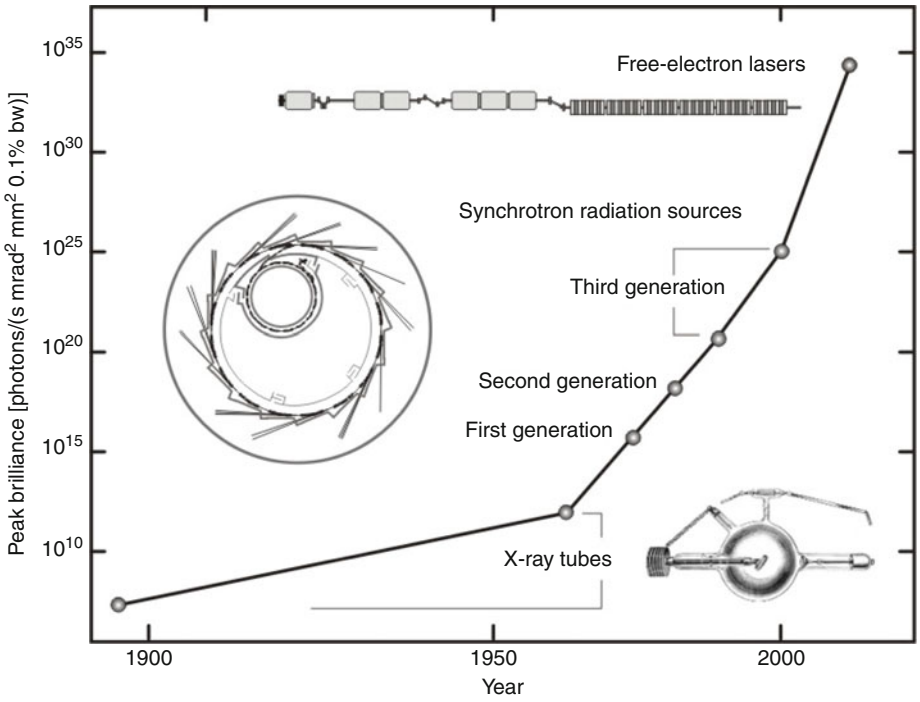


Fig. 1

Peak brilliance of radiation sources (photon rate per source size in mm², solid angle in mrad², and 0.1% bandwidth) from 1895 until today, covering more than 25 orders of magnitude. Schematically shown is an early x-ray tube, a third-generation synchrotron light source, and a linac-based free-electron laser

place in a spherical zone which expands at the speed of light and in which the field lines are distorted. The distortion of the electric field gives also rise to a magnetic field, and observers detecting such a distortion moving at the speed of light will call it "electromagnetic radiation" (*Fig. 2*).

The magnitude of the field distortion depends on the acceleration, and since it is much easier to accelerate electrons (or positrons), radiation from heavier particles is usually negligible.

Examples of electron acceleration are the oscillatory motion of electrons in a radio transmitter antenna, the abrupt deceleration of electrons in an x-ray tube, and the circular acceleration of electrons in a storage ring which gives rise to synchrotron radiation.

3 Generation of Synchrotron Radiation

An electron beam of energy E and current I perpendicular to a magnetic field B radiates the power

$$P = A\,I\,B \quad \text{with} \quad A = \frac{c e^2 E^3}{3\varepsilon_0 \left(m_e c^2\right)^4}. \tag{1}$$

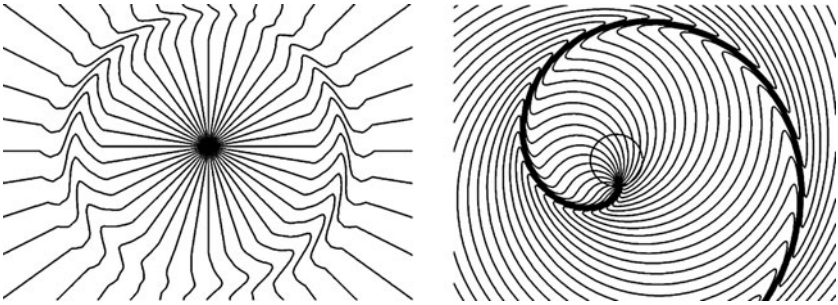


Fig. 2
Screenshots from running the simulation program `Radiation2D` (Shintake 2003). In the *left figure*, a charge has been briefly accelerated upward, causing a wave to travel spherically outward at the speed of light. In the *right figure*, a charge is on a circular orbit, emitting a spiral-shaped wavefront

Here, e is the elementary charge, c is the velocity of light, ε_0 is the dielectric constant, and m_e is the electron mass. The radiation spectrum, i.e., the power per time unit and spectral energy width as function of the photon energy E_p is given by the relation

$$\frac{dP}{dE_p} = \frac{P}{E_c} S\left(\frac{E_p}{E_c}\right) \quad \text{with} \quad S(\xi) = \frac{9\sqrt{3}}{8\pi} \xi \int_{\xi}^{\infty} K_{5/3}(u)\, du\,, \tag{2}$$

where $K_{5/3}(u)$ is a modified Bessel function. The critical photon energy

$$E_c = \frac{3eh}{4\pi m_e^3 c^4} E^2 B \quad \text{or} \quad E_c[\text{keV}] = 0.665 \cdot E^2[\text{GeV}^2] \cdot B[\text{T}] \tag{3}$$

with h being Planck's constant divides the spectrum into two parts of equal power and thus characterizes the typical energy range of photons emitted by electrons in a magnetic field. On a double-logarithmic scale, as shown in *Fig. 3*, the shape of the radiation spectrum given by $S(\xi)$ is always the same. For a normal-conducting electromagnet with an iron yoke, saturation limits the magnetic field to about 1.5 T. Consequently, the generation of photons in the x-ray regime (several keV) requires electron beam energies in the GeV range. As an example, the European Synchrotron Radiation Facility (ESRF) in Grenoble/France with a beam energy of 6.04 GeV produces photons with a critical energy of 20.6 keV in bending magnets with B = 0.85 T. Synchrotron light sources have beam energies ranging from below 1 GeV (e.g., the Metrology Light Source in Berlin/Germany with 630 MeV) to 8 GeV (SPring-8 in Hyogo/Japan). As described in the next paragraph, conventional synchrotron light sources comprise an electron accelerator (usually a synchrotron) and a storage ring in which the electrons circulate for hours at constant beam energy.

3.1 Acceleration of Electrons to Ultrarelativistic Energies

Electrostatic accelerators are limited to energies of some MeV. For a synchrotron light source with electrons in the GeV range, accelerators driven by radiofrequency (rf) such as linear accelerators or synchrotrons are required.

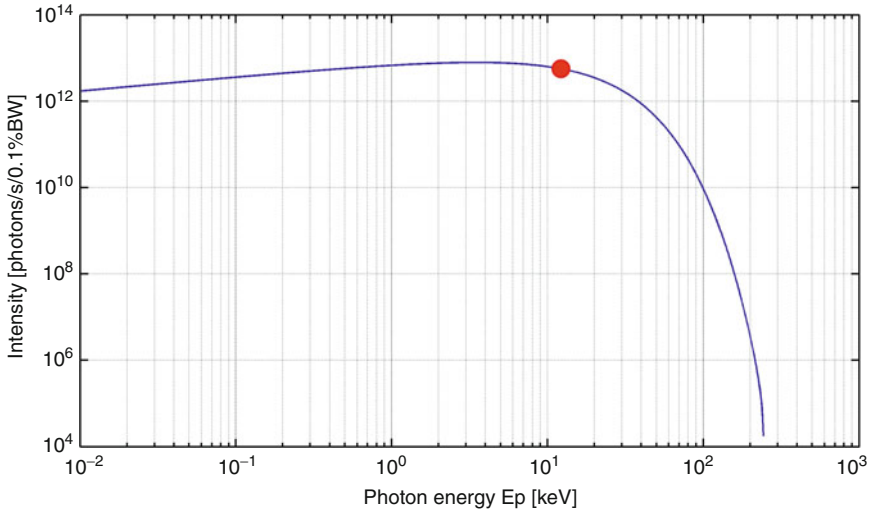


Fig. 3

Spectrum of synchrotron radiation from ultrarelativistic electrons. In this example, the electron energy is 3.5 GeV, the beam current is 100 mA, and the magnetic field is 1.5 T. The *dot* at 12.2 keV indicates the critical energy

3.1.1 Conventional Electron Linacs

Most linear accelerators (linacs) for high-energy electron beams consist of cylindrical waveguide structures operating in the TM_{01} mode and providing a strong longitudinal electric field at their center. However, the phase velocity of an electromagnetic wave in a cylindrical waveguide is higher than the velocity of light c. In order to accelerate electrons traveling at a speed slightly below c, it is necessary to reduce the phase velocity of the co-propagating wave by adding disks with holes to the waveguide (*Fig. 4*) at which the wave is partially reflected. The superposition of the reflected part with the main wave results in a phase velocity matching the electron speed.

The largest electron linac until now is the Stanford Linear Accelerator (SLAC) in Menlo Park/USA with a length of 3 km. Here, the electromagnetic wave has a frequency of 3 GHz (S band) and is produced by pulsed power klystrons of several tens of megawatt. The resulting energy gain is about 15 MeV per meter. In the past, the SLAC linac had reached a maximum energy of 50 GeV and was later employed at lower energy as injector of a B-meson factory. Today, its final third is used to deliver electrons of 13.6 GeV for LCLS, an x-ray free-electron laser. For conventional synchrotron radiation sources, however, much smaller linacs are used as injectors.

3.1.2 Superconducting Linacs and Energy Recovery

In order to improve the electric field gradient and the duty factor of linacs, superconducting structures have been developed. The duty factor is the fraction of time, at which the linac contains rf power. Resistive power losses in the walls limit normal-conducting copper structures to duty factors around 1%.



Fig. 4
Conventional normal-conducting linac structure for the acceleration of electrons. The disks reduce the phase velocity of an electromagnetic wave in a cylindrical waveguide to slightly below the velocity of light

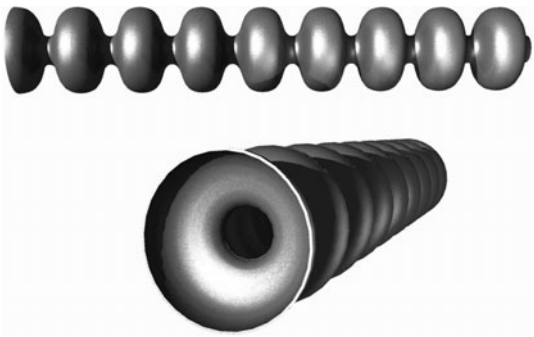


Fig. 5
Superconducting linac structure, consisting of an array of bell-shaped cavities made of highly purified niobium

Superconducting linacs consist of an array of bell-shaped cavities (*Fig. 5*) made of niobium. Compared to cylindrical structures, the bell shape eliminates the problem of multipacting (building up of an electron avalanche) which would limit the gradient to a very low value. A typical frequency is 1.3 GHz (L band) and gradients exceeding 40 MV/m have been achieved for individual cavities. For mass production, however, typical values are around 25 MV/m.

In a linac, each electron passes the structure only once and is either injected into a storage ring or directly used, for example, to produce radiation, and then absorbed by a beam dump, where the whole energy is lost. As a consequence for linac-driven radiation sources, a rather high average rf power is required and most of it is not used to produce radiation. The idea of an energy-recovery linac (ERL) (Tigner 1965) is to guide the bunched beam through the linac structure a second time (*Fig. 6*). If the phase of the particles is shifted by 180° with respect to the accelerating phase, the returning beam excites an electromagnetic wave in the linac structure and transfers its energy back to the rf field. In this case, the external

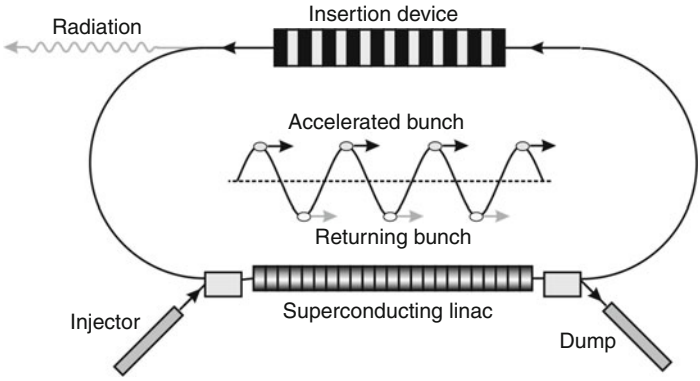


Fig. 6
Principle of energy recovery. A bunched particle beam is accelerated and subsequently decelerated in the same superconducting linac returning most of its energy to the rf field

power transmitter only has to compensate the power losses in the system. Since these losses are rather large in normal-conducting structures, the energy-recovery principle requires the use of superconducting cavities.

3.1.3 Synchrotrons

The first machines producing synchrotron radiation were – as the name suggests – synchrotrons, cyclic accelerators with a constant electron beam orbit (*Fig. 7*). Here, bending magnets define a circular trajectory and quadrupole magnets focus the beam. During acceleration by one or several rf cavities operating in the TM_{010} mode, the field of the magnets increases proportional to the beam energy. A cavity provides an accelerating voltage of several 100 kV, and at every turn the beam particles increase their energy accordingly.

At very low beam energy, the remnant field of magnets with iron yokes varies rather strongly and a stable beam orbit is not possible. Therefore, a pre-accelerator (typically a linac or a microtron) is required to inject electrons with an energy of at least some 10 MeV into the synchrotron. In order to periodically ramp the fields of the magnets up and down, they are in many synchrotrons driven by a sinusoidal current through the coils. The frequency of this current may be several 10 Hz, using a resonant circuit formed by the magnet coils and a bank of capacitors. This way, the stored energy alternates between the magnetic field of the coils and the electric field of the capacitors and only resistive losses have to be compensated. The beam energy achieved by electron synchrotrons ranges from some 100 MeV to several GeV.

The very first experiments with synchrotron radiation were performed at electron synchrotrons with photon beamlines installed tangentially to the electron orbit in bending magnets. Apart from the small beam current, the periodically changing beam energy E strongly limited the photon flux, since the radiation power is proportional to E^4. Consequently, the radiation intensity during the ramp cycle of a synchrotron is most of the time negligible.

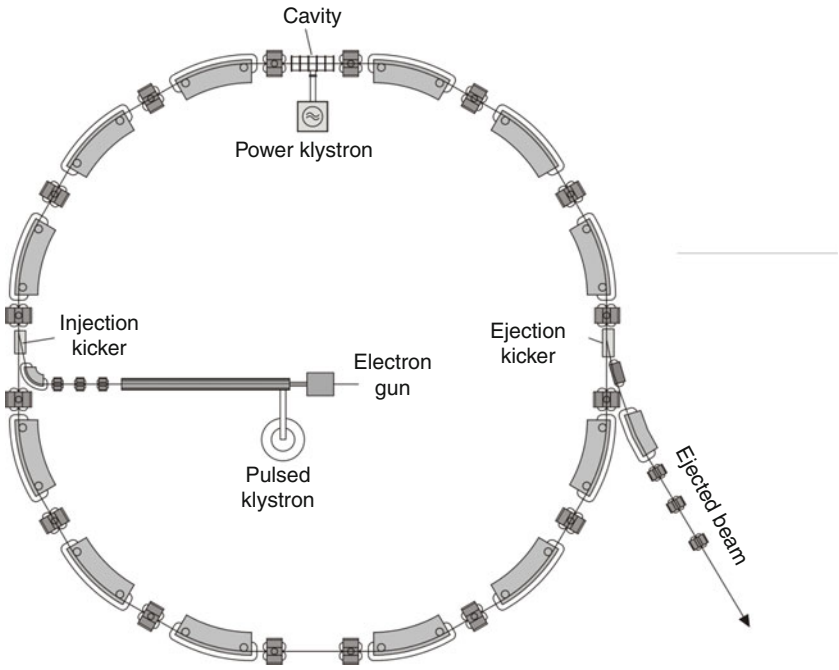


Fig. 7

A synchrotron with beam injected from a linac. The beam particles gain energy from an rf cavity at each turn. When reaching their final energy after many turns, the beam is ejected, e.g., to be accumulated in a storage ring

3.1.4 Electron Storage Rings

The photon flux produced by storage rings at constant beam energy is much higher than that of synchrotrons. Originally designed for high-energy physics with colliding beams, storage rings are very similar to synchrotrons. They are also circular machines with bending magnets and quadrupoles to guide and focus the beam. However, synchrotrons are optimized for varying magnetic fields, which requires, for example, to minimize eddy currents, while storage rings are optimized for beam quality and lifetime. Among other issues, this implies a more elaborate vacuum system providing a residual-gas pressure of the order of 10^{-7} Pa. For a beam lifetime of 10 h, for example, the average electron has to avoid fatal collisions over a distance of 10^{10} km, roughly the circumference of Saturn's orbit around the Sun. Furthermore, rf cavities are required to keep the beam energy constant by compensating the energy loss due to synchrotron radiation at each turn. For a long beam lifetime, additional rf voltage is needed to retain electrons which have lost or gained energy in interactions with residual-gas atoms or among themselves. Most storage rings are refilled when the beam current has dropped to, for example, half the initial value but over the last decade some synchrotron light sources have adopted a "top-up" mode of operation, injecting electrons every few minutes to keep the beam current constant on a sub-percent level (Ohkuma 2008). Top-up operation increases the time-averaged photon flux and

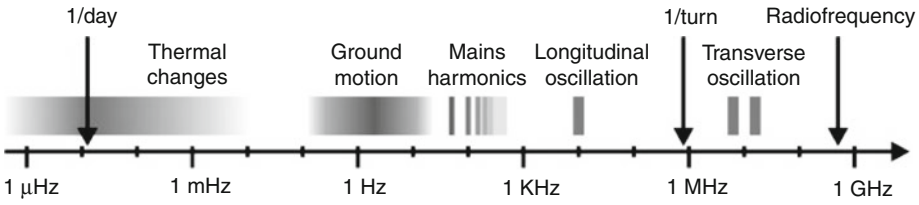


Fig. 8

Order-of-magnitude frequencies in electron storage rings. A typical frequency for rf systems is 500 MHz. Revolution frequencies (1/turn) range from 0.13 MHz to several MHz. The frequencies for horizontal and vertical oscillations are slightly different, usually around 10 MHz, while the longitudinal motion is three orders of magnitude slower. Not shown are harmonics and sidebands occurring in the beam spectrum

greatly improves the thermal stability of the storage ring and its x-ray beamlines. On the other hand, the demands on the injection system and on radiation safety are high.

A linac or synchrotron is used to accelerate electrons and inject them into the storage ring. The injection process is repeated in order to accumulate a stored beam current of several 100 mA. While a high current yields a high flux of photons, it also gives rise to collective effects, particularly coupled-bunch instabilities (Khan 2006). Here, each bunch induces electromagnetic fields by interacting with the surrounding vacuum chamber. These so-called wake fields act on subsequent bunches and thus excite longitudinal and transverse oscillations of the whole beam. Coupled-bunch instabilities limit the storable beam current and increase the average beam size, divergence, and energy spread, which directly influences the quality of emitted synchrotron radiation. In the worst case, sudden beam loss can occur. In order to keep beam instabilities at bay, the vacuum chamber should be large and highly conductive, sudden changes of its cross section should be avoided, and higher-order modes of the rf cavities should be reduced. In addition, bunch oscillations can be damped by active feedback systems measuring and correcting the position of each bunch at each turn.

Apart from beam instabilities, unwanted beam motion over a wide frequency range may arise from power supply ripple at harmonics of the mains frequency (50 or 60 Hz), from mechanical vibrations (around 10 Hz), from ground motion due to traffic (a few Hz), ocean waves (typically 0.2 Hz), diurnal and seasonal temperature changes, and other effects. Slow drifts are compensated by adjusting the settings of corrector dipole magnets every few seconds, while vibrations and power supply ripple are counteracted by sophisticated feedback systems with bandwidths in the kHz range (Hubert et al. 2009).

Sudden beam loss may also be induced by positive residual-gas ions which are attracted by the negative charge of the electron beam. A common countermeasure is to leave a gap in the fill pattern of the storage ring, allowing the ions to drift away. If the facility has a positron source (such as DORIS III in Hamburg), ion effects can be avoided by using a positron beam.

In summary, electron storage rings are optimized to deliver a high photon flux which either decays smoothly or is kept nearly constant. A lot of effort is put into stabilizing the beam over a frequency range from hundreds of MHz to sub-Hz (see *Fig. 8*). Furthermore, electron storage rings are optimized for small beam emittance, which in turn defines the brilliance of the photon beam. These quantities are described in the next paragraph.

3.1.5 Electron Beam Optics

In most applications of synchrotron radiation, a high photon flux is preferred, i.e., a large number of photons per unit time, beam current, and spectral width,

$$S\left[\frac{\text{photons}}{s \cdot A \cdot 0.1\%\,\text{bw}}\right], \tag{4}$$

where the spectral bandwidth (bw) is usually defined as 0.1% of the photon energy. In addition, a small transverse size of the radiation source and a small divergence is often required, which is expressed by the definition of the "spectral brightness" or "brilliance"

$$B = \frac{S}{4\pi^2 \cdot \sigma_x \sigma_y \sigma_{x'} \sigma_{y'}}\left[\frac{\text{photons}}{s \cdot A \cdot 0.1\%\,\text{bw} \cdot \text{mm}^2\,\text{mrad}^2}\right]. \tag{5}$$

Here, σ_i is one standard deviation of the horizontal and vertical electron position with $i = x$ and y, respectively, and of the horizontal and vertical trajectory angles $x' \equiv \mathrm{d}x/\mathrm{d}s$ and $y' \equiv \mathrm{d}y/\mathrm{d}s$. It is common practice to express angles as derivatives of the spatial coordinates with respect to the position s along the orbit, denoted by a prime. The rms (root mean square) horizontal and vertical beam size is a function of s, given by

$$\sigma_{x,y}(s) = \sqrt{\varepsilon_{x,y} \cdot \beta_{x,y}(s)}, \tag{6}$$

where $\varepsilon_{x,y}$ is the beam emittance and $\beta_{x,y}(s)$ is the amplitude function, also known as "beta function". The emittance $\varepsilon_{x,y}$ is proportional to the area occupied by the beam in the respective phase space (x, x') or (y, y') spanned by spatial and angular coordinates. Liouville's theorem states that this area is constant under the influence of conservative forces. Due to energy dissipation via synchrotron radiation, this is not exactly true for electron beams, but the emittance can nevertheless be regarded as constant. The smaller the horizontal and vertical emittance, the larger the brightness of the photon beam.

Besides bending magnets with homogeneous fields, strong quadrupole magnets with constant transverse field gradients are employed to keep the beam small. Quadrupole magnets comprising four hyperbolic pole faces (*Fig. 9*) focus the beam in one coordinate and act like defocusing lenses in the other direction. An overall focusing structure is always composed of several quadrupole magnets of alternating polarity.

The beam optics is mainly defined by quadrupole magnets with their respective strength and polarity chosen such that the beam is transversely confined at each position along the circumference of the storage ring. Since the beta function is uniquely defined for a given position s, the periodicity conditions

$$\beta_{x,y}(s + C) = \beta_{x,y}(s) \quad \text{and} \quad \beta'_{x,y}(s + C) = \beta'_{x,y}(s) \tag{7}$$

hold, in which C is the circumference of the storage ring and the prime again denotes the derivative with respect to s. The quadrupole magnets give rise to an attractive potential in which particles oscillate transversely around the ideal trajectory. This motion is known as betatron oscillation. The number of oscillations per revolution around the ring, $Q_{x,y}$, is called betatron tune and is typically of the order of 10. At particular tune values (integer, half integer, third integer etc.), magnetic field errors act repeatedly in the same direction and tend to increase the betatron oscillation. Such a resonance can lead to beam loss or at least reduce the beam lifetime, and therefore tune values being ratios of small integers must be avoided.

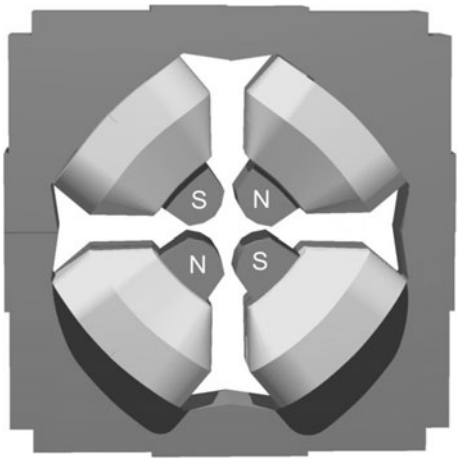


Fig. 9
Quadrupole magnet with four coils on an iron yoke with hyperbolic pole faces. In this example, an electron beam pointing into the image plane is focused vertically and defocused horizontally

A particle deviating from the reference energy E of the storage ring travels along a trajectory, which is horizontally displaced from the ideal orbit by

$$\Delta x = D_x(s) \cdot \frac{\Delta E}{E}, \tag{8}$$

where $D_x(s)$ is the dispersion function. Since dispersion is caused by horizontally deflecting dipole magnets, the vertical dispersion is usually negligible and the index x can be omitted. Like the beta functions, the dispersion and its derivative also satisfy the periodicity conditions

$$D(s + C) = D(s) \quad \text{and} \quad D'(s + C) = D'(s). \tag{9}$$

The energy distribution of a circulating electron beam is in good approximation given by a Gaussian distribution with a standard deviation of the order of 0.1%, and the maximum dispersion is typically 1 m or less. Since dispersion from dipole magnets increases the horizontal beam size significantly, most synchrotron light sources are based on achromatic lattices with dispersion-free straight sections to accommodate wigglers and undulators (see Sect. 3.2).

Due to the strong focusing in synchrotron light sources, even a small energy deviation causes a significant variation of the tune. A large chromaticity, defined as tune change per energy offset $\Delta Q_{x,y}/(\Delta E/E)$, causes off-energy particles to hit a resonance and thus reduces the beam lifetime. The chromaticity can be reduced by employing sextupole magnets in which the field depends quadratically on the transverse coordinates. However, in the presence of nonlinear fields, the motion of a particle with large transverse deviation becomes chaotic, and there is a distinct boundary beyond which particles are lost. To keep this so-called dynamic aperture large, many relatively weak sextupoles should be distributed around the ring rather than a few strong ones.

3.1.6 Radiation Effects

The synchrotron radiation power P emitted by an electron beam with energy E and current I in a magnetic field B is given by ❯ Eq. 1. With $B = E/(ecR)$, the energy loss of a single electron per turn is

$$W = \frac{eP}{I} = \frac{e^2}{3\varepsilon_0 (m_e c^2)^4} \frac{E^4}{R} \quad \text{or} \quad W[\text{keV}] = 88.5 \cdot \frac{E^4[\text{GeV}^4]}{R[\text{m}]} \tag{10}$$

ranging from below 100 keV to several MeV. The radiated power has to be compensated by the accelerating rf system, and due to the strong dependence on E, it imposes economic limits on the beam energy for electron or positron storage rings.

A given energy loss per turn, W, is restored by a sinusoidal rf voltage with amplitude U_{rf}, if the condition

$$W = eU_{\text{rf}} \sin \Psi_s, \tag{11}$$

is fulfilled, where Ψ_s is the so-called synchronous phase. Electrons deviating from Ψ_s undergo oscillations in longitudinal direction and in energy. This motion is known as synchrotron oscillation and the synchrotron tune Q_s (defined in analogy to $Q_{x,y}$ as number of oscillations per turn) is of the order of 10^{-2}, i.e., one longitudinal oscillation takes about 100 turns.

Synchrotron radiation has a damping effect on betatron and synchrotron oscillations. The reason for transverse damping is that photon emission from an oscillating electron causes a loss of longitudinal and transverse momentum, while only the longitudinal component is restored by the rf system. The longitudinal damping effect is a consequence of the dependence of the radiated power on electron energy. Radiation damping follows an exponential law with damping constants (i.e., inverse damping times) of

$$a_x = \frac{W}{2ET}(1 - \mathcal{D}), \quad a_y = \frac{W}{2ET}, \quad a_s = \frac{W}{2ET}(2 + \mathcal{D}) \tag{12}$$

for horizontal, vertical, and longitudinal oscillations, where T is the revolution time and $\mathcal{D}$ is a parameter which depends on the magnetic structure and can be neglected for many storage rings. An example with $W = 100$ keV, $E = 1$ GeV, and $T = 1$ μs shows that damping constants are of the order of 100 s^{-1}.

Synchrotron radiation not only damps particle oscillations but, due to its stochastic nature, also excites them. The equilibrium between the two effects determines the horizontal emittance (and thus the beam size and divergence) as well as the energy spread of the beam (which, in turn, determines the bunch length). The horizontal emittance is given by

$$\varepsilon_x = \frac{55h}{64\pi\sqrt{3} m_e^3 c^5} E^2 \frac{\langle \mathcal{H}(s)/R^3 \rangle}{(1 - \mathcal{D})\langle 1/R^2 \rangle} \tag{13}$$

with the same symbols as before and the function

$$\mathcal{H}(s) = \gamma_x(s) D^2(s) + 2\alpha_x(s) D(s) D'(s) + \beta_x D'^2(s) \tag{14}$$

using the conventional notation

$$\alpha_x(s) \equiv -\frac{1}{2}\beta_x'(s) \quad \text{and} \quad \gamma_x(s) \equiv \frac{1 + \alpha_x^2(s)}{\beta_x(s)}. \tag{15}$$

Taking the average $\langle \ldots \rangle$ is only required inside the bending magnets where $1/R$ is nonzero. The horizontal emittance of third-generation light sources is in the range of a few 10^{-9} rad m.

Assuming only horizontally deflecting dipole magnets, the vertical emittance is given by field errors of the magnetic structure, typically leading to $\varepsilon_y \approx 0.01\varepsilon_x$.

In order to obtain a low emittance, the horizontal beta function as well as the dispersion should be small inside the bending magnets. The magnetic structure of most synchrotron light sources is derived from the Chasman–Green lattice (Chasman et al. 1975), also known as double-bend achromat (DBA). It comprises a symmetric arrangement of two bending magnets with nonzero dispersion between them and no dispersion elsewhere (Fig. 10). Deriving the horizontal emittance according to Eq. 13 for a DBA lattice is beyond the scope of this introductory text (but straightforward). Starting with a given beta function β_0 and its derivative β_0' at the zero-dispersion end of the bending magnets, a remarkably simple result is obtained: the lowest possible emittance is achieved for

$$\beta_0 = 2\sqrt{3/5}\,L = 1.549\,L \quad \text{and} \quad \beta_0' = -2\sqrt{15} = -7.746\,, \tag{16}$$

where L is the length of each bending magnet, and the emittance scales approximately as

$$\varepsilon_x \propto \left(\frac{L}{R}\right)^3, \tag{17}$$

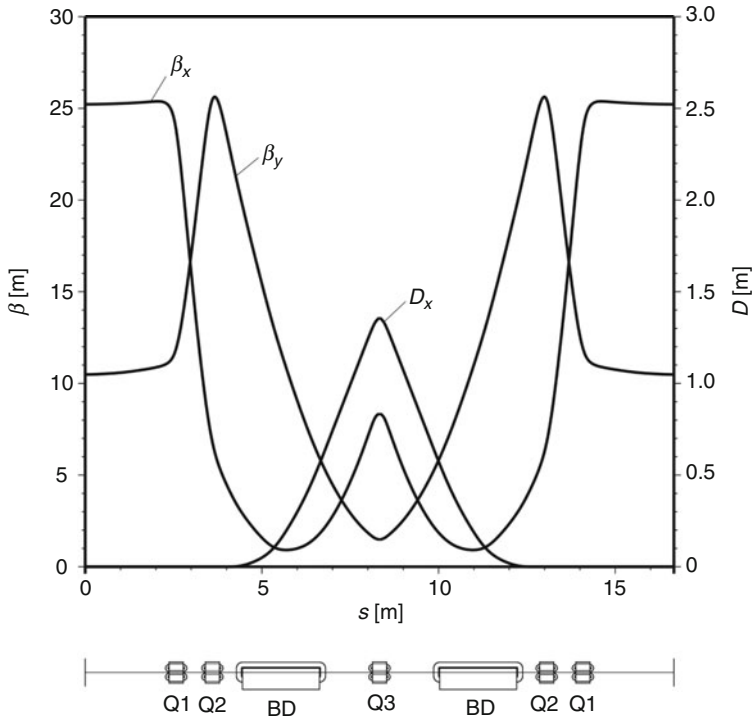


Fig. 10

Beta functions $\beta_{x,y}(s)$ and dispersion $D_x(s)$ in one unit cell of a Chasman–Green lattice, comprising two bending magnets and five quadrupoles arranged in three "families": Q1 and Q3 focus the beam horizontally, Q2 vertically

that is, a low emittance is achieved by using many magnets with small bending angle rather than few strong ones. In practice, the values of Eq. 16 are not used because they result in an extremely large chromaticity, but the emittance achieved with other values is only slightly larger. In some facilities (e.g., PETRA III in Hamburg), the emittance is further reduced by producing additional radiation in so-called damping wigglers. Wigglers and other insertion devices are described in the next section.

3.2 Insertion Devices

Synchrotron radiation from bending magnets is limited in several ways. Since the radiation is emitted in a wide fan tangentially to the electron orbit, only a small fraction of the photon flux hits the sample several ten meters away from the source point. Furthermore, the radiation spectrum is broad and most of its intensity is not used in experiments requiring monochromatic light. The accessible wavelength range is limited by the beam energy and the maximum magnetic field of ~ 1.5 T in conventional magnets (see Eq. 3). In order to overcome these limitations, special magnetic devices have been developed and an important feature of modern synchrotron light sources is to include a large number of free straight sections to accommodate these so-called insertion devices.

3.2.1 Wavelength Shifters and Superbends

Since the critical photon energy is proportional to the magnetic field of a bending magnet (see Eq. 3), superconducting magnets can be employed to shift the photon spectrum to shorter wavelengths. Superconducting insertion devices as shown in *Fig. 11* with adjacent magnets to cancel the bending angle are known as wavelength shifters. Superconducting magnets replacing conventional bending magnets of a storage ring are called superbends. At the position of such a device, the beta function should be small to minimize its influence on the beam emittance.

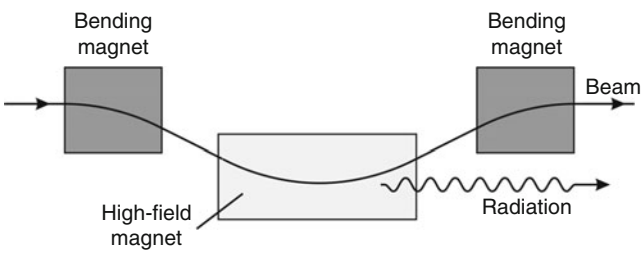


Fig. 11
Wavelength shifter comprising a superconducting high-field magnet and two adjacent magnets to cancel the bending angle

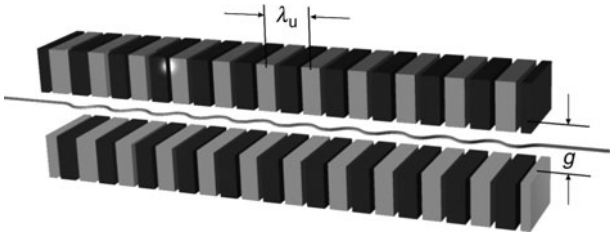


Fig. 12
Wigglers and undulators with alternating magnetic poles. The period length λ_u is the longitudinal distance between two like-sign poles. The field on the midplane axis is determined by the period length and the gap *g*

3.2.2 Wigglers and Undulators

Wigglers and undulators consist of a large number of short bending magnets with alternating polarity arranged along a straight line (*Fig. 12*). Characteristic parameters are the period length λ_u (the longitudinal distance between two like-sign magnets) and the peak field $\hat{B}$ on the midplane axis of the device. For electromagnetic poles, the period length of undulators is usually above 10 cm. In order to reach shorter period lengths, permanent magnets (CoSm or NdFeB) are commonly used, often combined with steel poles to achieve higher fields. For such a hybrid design, the maximum field can be approximated by (Brown et al. 1983)

$$\hat{B}\,[\mathrm{T}] = 3.33 \exp\left[-\frac{g}{\lambda_u}\left(5.47 - 1.8\frac{g}{\lambda_u}\right)\right]. \tag{18}$$

Evidently, a small gap g between opposite poles allows to reach high field values. Therefore, undulator vacuum chambers often have a very small vertical aperture and thin walls, or the magnets are even placed inside the vacuum vessel. Assuming a horizontally deflecting device, the vertical field component oscillates periodically along the beam axis according to

$$B_y(s) = \hat{B} \sin\left(\frac{2\pi}{\lambda_u} s\right) \tag{19}$$

leading to a sinusoidal beam trajectory with

$$x(s) = \frac{\lambda_u}{2\pi}\frac{K}{\gamma} \sin\left(\frac{2\pi}{\lambda_u} s\right) \quad \text{and} \quad x'(s) = \frac{K}{\gamma} \cos\left(\frac{2\pi}{\lambda_u} s\right), \tag{20}$$

where $\gamma = E/(m_e c^2)$ is the Lorentz factor and

$$K \equiv \frac{\lambda_u e \hat{B}}{2\pi m_e c} \quad \text{or} \quad K = 93.4 \cdot \lambda_u\,[\mathrm{m}] \cdot \hat{B}\,[\mathrm{T}] \tag{21}$$

is a dimensionless strength parameter. For $K \gg 1$, the radiation spectrum is similar to that of a bending magnet multiplied by the number of magnetic poles. In this case, the device is called a "wiggler." For K of the order of 1 or smaller, the maximum angle of the trajectory in Eq. 20 does not exceed $1/\gamma$, the typical opening angle of synchrotron radiation, and interference of radiation from different poles occurs. Such a device, called "undulator," has a line spectrum

with a fundamental wavelength of

$$\lambda = \frac{\lambda_u}{2\gamma^2}\left(1 + \frac{K^2}{2} + \gamma^2\Theta^2\right) \tag{22}$$

and harmonics thereof. In forward direction at $\Theta = 0$, only odd harmonics appear. Away from the beam axis, even harmonics show up and the radiation is redshifted, that is, the wavelength is increased by $\lambda_u\Theta^2/2$. For an undulator with N periods, the linewidth is approximately given by

$$\frac{\Delta\lambda}{\lambda} \approx \frac{1}{N}\,. \tag{23}$$

Therefore and since the opening angle of undulator radiation is proportional to $\sqrt{N}$ in each dimension, a factor of N^2 is gained in brilliance over radiation from a bending magnet. Undulator design is still an active research topic. Recent developments include superconducting undulators with period lengths of 10 mm or less (Casalbuoni et al. 2006) and cryogenic undulators (Hara et al. 2004) with permanent magnets cooled to 100 K to increase their field. Undulators for synchrotron light sources are several meters long with typically 100 periods. Despite the large forces between opposite magnets, the wavelength is tuned by changing the gap g mechanically with a precision on the μm level. For free-electron lasers (see below), the total undulator length can exceed 100 m.

So far, "planar" insertion devices were discussed with electron trajectories in their midplane producing linearly polarized light. Radiation from bending magnets contains a circular component above or below their midplane. For planar wigglers, the out-of-plane helicities cancel unless the field strength of opposite poles is different (asymmetric wigglers). A much larger degree of circular polarization is achieved in so-called elliptical undulators with helical beam trajectories.

3.3 Synchrotron Radiation Sources Worldwide

Presently, about 60 dedicated synchrotron radiation sources in more than 20 countries are in operation, and their number is still growing. Examples of third-generation light sources commissioned in the last two decades are given in *Table 1*. Ever since 1997, SPring8 in Japan has the largest beam energy with 8 GeV, and some facilities have reached a beam current of 0.5 A. The facility with the largest circumference and lowest emittance is PETRA III at DESY in Germany, where one octant of a former e^+e^- collider has been remodeled, and other large storage rings may follow this example. A more complete list of synchrotron radiation sources can be found, for example, under www.lightsources.org.

4 Applications of Synchrotron Radiation

Soon after their discovery, x-rays became an invaluable tool to probe the structure of matter. Creating images of the shadow of an object (like the human body) is only one of many applications. The major classes of techniques are diffraction, spectroscopy, and imaging. In addition to these classical x-ray applications, the unique properties of synchrotron radiation have given rise to a number of other techniques described at the end of this section.

Table 1
Examples of third-generation synchrotron radiation sources

Facility (first beam) location/country	Circumference [m]	Beam energy [GeV]	Current [mA]	Hor. emittance [nm rad]
ESRF (1992) Grenoble/France	844	6	200	4
ALS (1993) Berkeley/USA	196.8	1.9	400	4.2
ELETTRA (1993) Trieste/Italy	259.2	2.0	330	7.0
APS (1995) Argonne/USA	1104	7	100	3
MAX II (1996) Lund/Sweden	90	1.5	280	9
DELTA (1996) Dortmund/Germany	115.2	1.5	130	18
SPring8 (1997) Hyogo/Japan	1436	8.0	100	3
BESSY (1998) Berlin/Germany	240	1.7	300	5.2
SLS (2000) Villigen/Switzerland	288	2.4	400	5
CLS (2003) Saskatoon/Canada	170.9	2.9	250	18
SPEAR3 (2003) Stanford/USA	234.1	3.0	500	16
SOLEIL (2006) Gif-sur-Yvette/France	354.1	2.75	500	3.7
Diamond (2006) Didcot/UK	561.6	3.0	300	2.7
SSRF (2008) Shanghai/China	432	3.5	300	3.9
ALBA (2009) Cerdanyola/Spain	268.8	3.0	400	4.3
PETRA III (2009) Hamburg/Germany	2304	6.0	100	1

4.1 Diffraction

Elastic scattering of x-rays from molecules or crystal lattices generates a diffraction pattern, which can be recorded by position-sensitive devices like photographic films, drift chambers, or CCD chips (*Fig. 13*). Since the sample atoms are not excited and recoil is negligible, the wavelengths of incident and scattered radiation are the same. Neglecting multiple scattering effects, the diffraction pattern is related to the spatial structure of the scattering object. Usually,

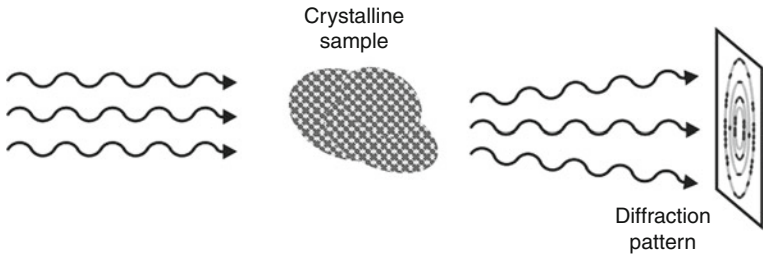


Fig. 13
X-rays scattered elastically from a crystalline sample form a diffraction pattern which contains information on its microscopic spatial structure

the image is enhanced by a lattice factor due to samples with translational symmetry. Atoms in crystals form a unit cell which repeats itself, and the study of other objects (proteins, in particular) requires to arrange them in a crystal-like fashion. However, when using extremely intense radiation pulses like those generated by x-ray free-electron lasers, diffraction images from a single protein molecule will be possible.

The diffraction pattern, that is, the radiation intensity as function of scattering angles, is given by the product of the lattice factor and a complex structure factor, which is directly related to the spatial electron density within a unit cell. Since only the square of this quantity is measured, the phase information of the scattered light waves is lost, but techniques have been devised to get around the phase problem and recover the spatial structure from the diffraction pattern. Spectacular successes of x-ray diffraction were, among others, the studies of deoxyribonucleic acid (Watson and Crick 1953) and the ribosome, see, e.g., Ban et al. (2000). These two examples are almost 50 years apart and both were awarded a Nobel Prize (1962 and 2009).

4.2 Spectroscopy

Another important class of applications involves photoelectric absorption of synchrotron radiation (*Fig. 14*). Here, an electron (photoelectron) is ejected from the atom with a certain kinetic energy, while the atom is left in an excited state. The hole created in an inner atomic shell is then filled by an outer-shell electron liberating energy by emitting either a photon (fluorescence) or another electron (Auger effect).

Spectroscopic information can thus be gained by studying the absorption of photons as function of wavelength or by observing the spectrum of fluorescent radiation, by measuring the kinetic energies of photoelectrons, or by measuring the Auger electron spectrum. In the case of gas-phase samples, additional information can be gained from measuring the velocities of the ionized atoms.

The goal of a spectroscopic study may be to either study the energy levels of a particular system, to analyze the chemical content of a sample by known energy levels, or to gain spatial information. As an example of the latter case, extended x-ray absorption fine structure (EXAFS) is a well-established method to determine distances between the absorbing atom and its surroundings, see, e.g., Koningsberger and Prins (1988). When a photoelectron is liberated, the interference between the spherical wave of the outgoing electron and waves reflected by the

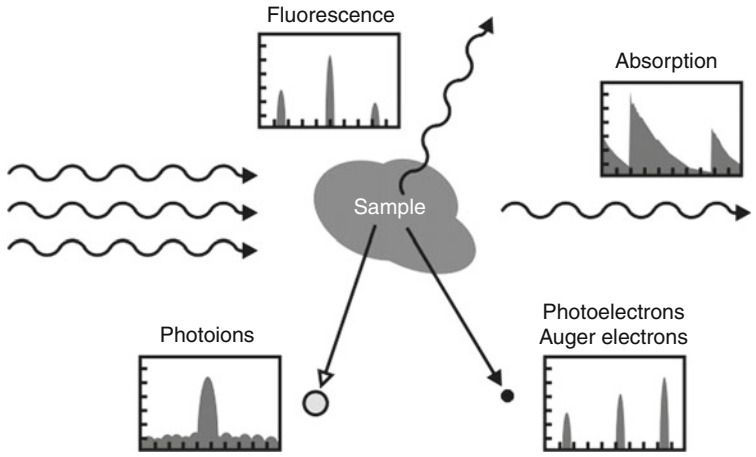


Fig. 14

Spectroscopic information is obtained by measuring the energy of photons (transmitted or emitted by excited atoms), electrons (photoelectrons or Auger electrons), or ionized atoms of the sample

neighboring atoms leads to wiggles in the absorption spectrum slightly above the absorption edge. These wiggles are visible in the absorption spectrum of an incident monochromatized photon beam as well as in the spectrum of photoelectrons. The distances of neighboring atoms can then be calculated from the frequency content of the respective spectra.

Fluorescent radiation spectroscopy offers a destruction-free way to analyze the chemical content of a sample which is nowadays also employed by historians and archaeologists, for example, to reveal hidden layers of a painting (Dik et al. 2008).

4.3 Imaging

In order to directly image an object, one advantage of x-rays is their large penetration depth, another is their short wavelength which translates into spatial resolution (x-ray microscopy) and yet another advantage is the strong dependence of x-ray absorption on properties like atomic number and magnetic state. The quality of images can be greatly enhanced by subtracting data taken above and below an absorption edge of particular chemical elements. One example is coronary angiography (imaging blood vessels in the vicinity of the heart) by taking pictures above and below the K edge of iodine as contrast material (Dill et al. 1998). The advantage of intense synchrotron radiation is that intravenous injection (rather than intraarterial catheterization) of the contrast material is sufficient. Another imaging technique is tomography, measuring the transmission through an object from several angles and reconstructing the 3-dimensional structure (Bonse and Bush 1996). In most applications, the image contrast is provided by absorption, but phase contrast has also been successfully employed. Microscopy at short wavelengths is possible using Fresnel zone plates (Yun et al. 1999) to focus the x-rays, for example, to study the structure of wet biological samples in the "water window" (300–500 eV).

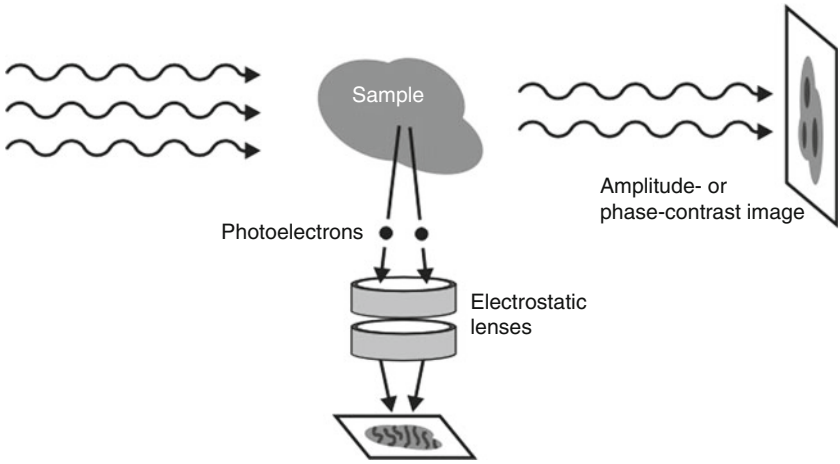


Fig. 15
A sample is visualized by recording x-rays passing through the sample or by passing photoelectrons emitted from its surface through an electron microscope

Apart from x-rays, imaging can also be performed with photoelectrons, as sketched in *Fig. 15*. In PEEM (photoemission electron microscopy), a sample is illuminated with x-rays and photoelectrons are passed through the electrostatic lenses of an electron microscope to study, for example, magnetic domains (Stöhr et al. 1993), or the chemical composition of nanostructures.

4.4 Other Applications

Apart from the applications outlined so far, which constitute the bulk of experiments with x-rays and synchrotron radiation, there is a growing number of additional techniques. The following – certainly incomplete – description is meant to give a flavor of the wide range of scientific opportunities with synchrotron radiation.

4.4.1 Time-Resolved Studies

Since synchrotron radiation is pulsed, much of what has been discussed before can be done in a time-resolved fashion. As in laser spectroscopy, time-resolved studies are usually carried out as pump-probe experiments, where the sample is "pumped", that is, excited in some way by an intense radiation pulse (e.g., from a laser) and the subsequent x-ray pulse is used to probe the state of the sample as function of the delay between the two pulses (Chergui and Zewail 2009). By moving mirrors on the micrometer level, the arrival time of the pump pulse is conveniently controlled on the femtosecond scale. Typically, the pulse duration of synchrotron radiation is several 10 ps, but methods exist to shorten electron bunches (and thus the emitted radiation pulses) to a few picoseconds and even to generate pulses on the 100 fs scale. Pulses from free-electron lasers can be as short as 10 fs.

4.4.2 Far-Infrared Radiation

While the previous discussion was focused on short-wavelength radiation, there are several synchrotron radiation sources offering coherent radiation in the far-infrared (THz) regime. This radiation penetrates nonconducting materials and its absorption coefficient is very sensitive to, for example, the water content of a sample, to different dielectric constants, and to the electric conductivity. Spectroscopy in the THz regime (e.g., Schmidt et al. 2009) addresses collective molecular states (rotation or vibration) and other "fingerprint" properties, making it relevant, for example, in security applications like the detection of explosives.

There used to be a "THz gap" due to the lack of bright radiation sources which has been partially filled by laser-based techniques. Nowadays, electron storage rings with reduced bunch length can also create coherent and intense THz pulses with large bandwidth and short duration (Abo-Bakr et al. 2003). In addition, new storage rings and linear accelerators dedicated to the purpose of generating THz radiation have been proposed.

4.4.3 X-Ray Holography

Holographic images are created by the interference of a reference wave and coherent laser light passing a sample. As usual, higher spatial resolution requires shorter wavelength. In addition to that, the idea of holography with short-wavelength radiation is attractive because it works without lenses (which are not available for x-rays) and preserves phase information. Holographic experiments with synchrotron radiation (Eisebitt et al. 2004) were successfully conducted employing the coherence of synchrotron light restricted in space by a pinhole and in wavelength by a monochromator – at the expense of intensity. The coherence of synchrotron light may also be improved by creating a periodic density modulation of the electron beam at the radiation wavelength. Such a modulation is an intrinsic feature of free-electron lasers.

4.4.4 Metrology

It is an outstanding property of synchrotron radiation that its intensity and spectrum, for example, from a bending magnet, can be calculated with high accuracy in the framework of classical and quantum electrodynamics once the electron energy, the beam current, and the magnetic field is known. Therefore, synchrotron radiation is ideally suited to serve as a primary standard for the calibration of radiation sources and detectors. There is a great demand for detector calibration in science (e.g., satellite-based astronomy) and industry which the national metrology agencies (e.g., NIST in the USA or PTB in Germany) meet by employing synchrotron radiation sources. For very sensitive detectors, the radiation may even come from a single electron circulating in a large storage ring (Klein et al. 2010).

4.4.5 X-Ray Lithography

Micromechanical devices such as sub-millimeter gears, nozzles, waveguides etc. are fabricated by x-ray lithography (Heuberger 1985), followed by other process steps like electroplating and

molding. In the lithography process, an x-ray sensitive photoresist is illuminated by synchrotron radiation through a patterned x-ray absorbing mask, and the exposed photoresist is subsequently etched away to form a 3-dimensional structure. The short wavelength of synchrotron radiation yields a high spatial resolution, its low angular divergence creates structures with high aspect ratios and smooth walls with a high degree of parallelism.

5 The Next Generation

5.1 Storage Rings

Third-generation light sources have reached limits that are not easily surpassed. In storage rings, the beam is subject to the equilibrium between radiation excitation and damping. The horizontal emittance can only be reduced by increasing the circumference or by adding damping wigglers. Both has been done at PETRA III in Hamburg/Germany, reaching an emittance of 1 nm rad.

The bunch length in storage rings is dictated by radiation effects as well, and results in radiation pulses of several 10 ps duration. If the magnetic lattice is flexible enough, the momentum-compaction factor $\alpha = (\Delta L/L)/(\Delta p/p)$ can be reduced to achieve pulses as short as 1 ps (rms) at the expense of bunch charge and emittance. Laser-based methods allow to extract radiation from a 100-fs long "slice" of a bunch (Zholents and Zolotorev 1996).

5.2 Linac-Based Free-Electron Lasers

Since 2009, the Linear Coherent Light Source (LCLS) (Ding et al. 2009) is operational at SLAC in Menlo Park/USA, producing radiation at a wavelength of 1.5 Å with a peak brilliance exceeding that of conventional synchrotron light sources by nine orders of magnitude. This large factor is partly accomplished by the coherent emission of photons, and partly due to a reduction in bunch length and beam emittance. LCLS is a free-electron laser (FEL), a light amplifier based on stimulated emission from a beam of ultrarelativistic electrons which – in contrast to conventional lasers based on bound electrons – can be tuned continuously in wavelength.

In linear accelerators (linacs), the emittance is given by the electron source and – if care is taken to avoid space-charge effects – decreases during acceleration as $1/\gamma$, where γ is the Lorentz factor. In this context, the "normalized" emittance $\varepsilon_n = \varepsilon\,\gamma$ is usually quoted. With $\varepsilon_n = 0.5\,\mu$m rad as an example and the beam energy of LCLS (13.6 GeV or $\gamma = 2.7 \cdot 10^4$), the absolute emittance would be $\varepsilon = \varepsilon_n/\gamma = 18$ pm rad, which is significantly smaller than in storage rings.

The bunch length in linacs can be strongly reduced by a combination of energy "chirp" (energy variation along the bunch) and so-called bunch compressors, that is, dispersive sections with energy-dependent path lengths. This way, radiation pulses of the order of 10 fs were achieved at FLASH (DESY/Hamburg) and LCLS.

The intensity of coherently emitted radiation increases quadratically with the number of participating electrons, which leads to a large enhancement over conventional synchrotron radiation, where the intensity scales linearly. Coherent emission is accomplished by micro-bunching, that is, modulating the electron density with the periodicity of the radiation

wavelength. Present-day FELs are based on the SASE principle (self-amplified spontaneous emission) (Kondratenko and Saldin 1980) where spontaneous radiation in the first part of a long undulator acts back on the electrons by periodically modulating their kinetic energy. As these off-energy electrons proceed in the undulator, electrons with lower energy follow a longer trajectory and fall behind, while those with higher energy catch up. This process initiates a density modulation which gives rise to coherent radiation, which in turn increases the modulation. The result of this positive feedback loop is an exponential rise of radiation intensity along the undulator. The distance over which the radiation power increases e-fold is called gain length,

$$L_{\mathrm{g}} \propto \frac{\gamma\, \sigma_r^{2/3}}{n_{\mathrm{e}}^{1/3}}\,, \tag{24}$$

and depends on γ, on the electron density n_{e}, and on the transverse beam size σ_r. The process saturates after typically 20 gain lengths because the maximum achievable density modulation is reached. Presently, several FELs based on the SASE process are under construction or proposed. However, since the amplified radiation starts from noise, the pulses exhibit unwanted fluctuations in intensity as well as their temporal and spectral shape.

An alternative route is to "seed" a short-wavelength FEL with well-defined radiation pulses. Monochromatized synchrotron radiation has been discussed as source of seed pulses ("self seeding") but the present effort concentrates on seeding with ultrashort pulses from external lasers with near-visible wavelength, and two different strategies are pursued. When intense laser pulses are focused into a gas, the interaction with atomic electrons gives rise to high harmonics of the laser wavelength, which can be used to directly seed an FEL. Alternatively, when an electron beam is seeded with a longer wavelength, its subsequent nonharmonic density modulation leads to the emission of higher-harmonic radiation. Worldwide, pioneering work in both directions is underway (Lambert et al. 2008; Xiang et al. 2010) to improve the quality of radiation pulses with seeded FEL sources.

5.3 Energy-Recovery Linacs

With electrons circulating in a storage ring, a beam current of several 100 mA can be obtained which is out of the question for linacs: generating a beam with a current of 100 mA at 5 GeV, for example, would require a dc power of 500 MW. Another advantage of storage rings is that they are multiuser facilities, delivering radiation to many beamlines simultaneously. On the other hand, the equilibrium emittance and bunch length in a storage ring is much larger than what can be nowadays obtained with a low-emittance electron gun and a linac.

Using a superconducting accelerating structure, the principle of energy recovery as outlined in Sect. 3.1.2 allows to generate bunches with small emittance and short duration and with a high repetition rate for multiple users. Energy recovery linacs (ERLs) to drive infrared FELs are already employed, for example, at the Jefferson Laboratory in Newport News/USA, accelerating 10 mA of beam current to 150 MeV. ERL-driven x-ray sources do not yet exist, but R&D to this end is underway, for example, at KEK in Tsukuba/Japan and the Cornell University in Ithaca/USA (Bilderback et al. 2009). According to their conceptual design, these ERLs will be similar to storage rings, but with two orders of magnitude higher brightness, better transverse coherence, and much shorter pulse duration.

6 Conclusions

At the time of writing, 115 years after Röntgen's discovery, for which he was awarded the very first Nobel Prize in physics (1901), research with x-rays is still undergoing a rapid and exciting development. The technology of storage rings is mature and many ring-based synchrotron light sources exist worldwide as workhorses for a large number of applications in physics, chemistry, biology, and materials sciences with steadily improving experimental techniques.

Advances in accelerator technology have led to new radiation sources. Free-electron lasers (FELs) can deliver pulses of extreme brilliance and ultrashort duration, opening up new scientific opportunities such as snapshots of the structure of single proteins with sub-picosecond resolution. As a first hard-x-ray FEL, LCLS at SLAC was commissioned in 2009, and others are already under construction. Another challenging and not yet demonstrated technology is that of ERL-based x-ray sources. Their properties will be somewhere between those of storage rings and FELs, addressing scientific questions for which the peak power of FELs is too extreme or the pulse rate too low. It can be expected that these different types of accelerator-based radiation sources (storage rings, FELs, ERLs, and maybe others yet to be invented) will coexist, serving different classes of experiments.

Finally, first attempts to generate synchrotron radiation with electrons from "table-top" laser-plasma accelerators have been successful (Fuchs et al. 2009). While these novel accelerators are still in their infancy with strong shot-to-shot variation in energy and bunch charge, there has been significant progress over the last decade and one of the long-term goals is to build table-top (or at least laboratory-size) synchrotron light sources and FELs.

References

Abo-Bakr M et al (2003) Brilliant, coherent far-infrared (THz) synchrotron radiation. Phys Rev Lett 90:094801

Balewski K (2010) Commissioning of PETRA III. In: Proceedings of the international particle accelerator conference, Kyoto/Japan, p 1280

Ban N et al (2000) The complete atomic structure of the large ribosomal subunit at 2.4 angstrom resolution. Science 289:905

Bilderback DH et al (2009) Energy recovery linac (ERL) coherent hard x-ray sources. New J Phys 12:035011

Bonse U, Bush F (1996) X-ray computed microtomography using synchrotron radiation. Prog Biophys Mol Biol 66:133

Brown G et al (1983) Wiggler and undulator magnets – a review. Nucl Inst Meth 208:65

Casalbuoni S et al (2006) Generation of x-ray radiation in a storage ring by a superconductive cold-bore in-vacuum undulator. Phys Rev ST Accel Beams 9:010702

Chasman R, Green GK, Rowe EM (1975) Preliminary design of a dedicated synchrotron radiation facility. IEEE Trans Nucl Sci 22:1765

Chergui M, Zewail AH (2009) Electron and X-ray methods of ultrafast structural dynamics: advances and applications. Chem Phys Chem 10(10):28

Coolidge WD (1917) U.S. Patent 1,211,092

Dik J et al (2008) Visualization of a lost painting by Vincent van Gogh using synchrotron radiation based x-ray fluorescence elemental mapping. Anal Chem 80:6436

Dill T et al (1998) Intravenous coronary angiography with synchrotron radiation. Eur J Phys 19:499

Ding Y et al (2009) Measurements and simulations of ultralow emittance and ultrashort electron beams in the linear coherent light source. Phys Rev Lett 102:254801

Eisebitt S et al (2004) Lensless imaging of magnetic nanostructures by x-ray spectro-holography. Nature 432:885

Elder FR, Gurewitsch AM, Langmuir RV, Pollock HC (1947) Radiation from electrons in a synchrotron. Phys Rev 71:829

Fuchs M et al (2009) Laser-driven soft-X-ray undulator source. Nature Phys 5:826

Hara T et al (2004) Cryogenic permanent magnet undulators. Phys Rev ST Accel Beams 7:050702
Heuberger A (1985) X-ray lithography with synchrotron radiation. Z Phys B – Condens Matter 61:473
Hubert N et al (2009) Global orbit feedback systems down to DC using fast and slow correctors. In: Proceedings of the DIPAC 2009, Basel, Switzerland, www.jacow.org
Khan S (2006) Collective phenomena in synchrotron radiation sources. Springer, Berlin
Klein R, Thornagel R, Ulm G (2010) From single photons to milliwatt radiant power – electron storage rings as radiation sources with high dynamic range. Metrologia 47:R33
Kondratenko AM, Saldin EL (1980) Generation of coherent radiation by a relativistic electron beam in an ondulator. Part Accel 10:207
Koningsberger DC, Prins R (1988) X-ray absorption: principles, applications, techniques of EXAFS, SEXAFS and XANES. Wiley, New York
Lambert G et al (2008) Injection of harmonics generated in gas in a free-electron laser providing intense and coherent extreme-ultraviolet light. Nat Phys 4:296
Ohkuma H (2008) Top-up operation in light sources. In: Proceedings of the 2008 European particle acceleration conference, Genova/Italy, p 36, www.jacow.org
Röntgen WC (1895) Ueber eine neue Art von Strahlen (Vorläu_ge Mittheilung). In: Sitzungsberichte der Würzburger Physik.-Medic.-Gesellschaft
Schmidt DA et al (2009) Rattling in the cage: ions as probes of sub-picosecond water network dynamics. J Am Chem Soc 131:18512
Shintake T (2003) Real-time animation of synchrotron radiation, Nucl Intstr Meth A 507:89; the program Radiation2D 2.0 can be downloaded from http://www-xfel.spring8.or.jp
Stöhr J et al (1993) Element-specific magnetic microscopy with circularly polarized light. Science 259:658
Tigner M (1965) A possible apparatus for electron clashing-beam experiments. Nuovo Cimento 37:1228
Watson JD, Crick FHC (1953) A structure for deoxyribose nucleic acid. Nature 171:737
Xiang D et al (2010) Demonstration of the echo-enabled harmonic generation technique for short-wavelength seeded free electron lasers. Phys Rev Lett 105:114801
Yun W et al (1999) Nanometer focusing of hard x-rays by phase zone plates. Rew Sci Instrum 70: 2238
Zholents AA, Zolotorev MS (1996) Femtosecond x-ray pulses of synchrotron radiation. Phys Rev Lett 76:912

Further Reading

Als-Nielsen J, McMorrow D (2001) Elements of modern x-ray physics. Wiley, New York
Attwood D (1999) Soft x-rays and extreme ultraviolet radiation. Cambridge University Press, Cambridge
Duke PJ (2000) Synchrotron radiation. Oxford University Press, Oxford
Falta J, Möller T (eds) (2010) Forschung mit Synchrotronstrahlung (in German). Vieweg+Teubner, Wiesbaden
Pietsch U, Holy V, Baumbach T (2004) High-resolution x-ray scattering: from thin films to lateral nanostructures. Springer, Berlin
Schmüser P, Dohlus M, Roßbach J (2008) Ultraviolet and soft-x-ray free-electron lasers. Springer, Berlin
Wiedemann H (2003) Synchrotron radiation. Springer, Berlin
Wiedemann H (2007) Particle accelerator physics. Springer, Berlin
Wille K (2001) The physics of particle accelerators. An introduction. Oxford University Press, Oxford
Winick H (ed) (1994) Synchrotron radiation sources – a primer. World Scientific, Singapore

9 Calibration of Radioactive Sources

Dirk Arnold · Herbert Janßen
Physikalisch-Technische Bundesanstalt (PTB), Braunschweig, Germany

1 Introduction ***188***

2 Radioactive Decay ***189***
2.1 Alpha Decay 189
2.2 Beta$^-$ Decay 190
2.3 Beta$^+$ Decay 190
2.4 Electron Capture 190
2.5 Gamma Decay 190
2.6 Internal Conversion 190
2.7 Other Decay Modes 191

3 Activity Standards ***191***

4 Primary Methods for the Calibration of Activity Standards ***193***

5 Secondary Methods for the Calibration of Activity Standards ***196***

6 International Comparability of Activity Standards ***197***

7 Conclusions ***198***

8 Cross-References ***198***

References ***199***

C. Grupen, I. Buvat (eds.), *Handbook of Particle Detection and Imaging*,
DOI 10.1007/978-3-642-13271-1_9, © Springer-Verlag Berlin Heidelberg 2012

Abstract: All detector systems for the measurement of radioactivity in the different fields of applications need to be calibrated in terms of efficiency with radioactive sources of known activities and of known radionuclides. This is true for the measurement of environmental radioactivity, activities of sources for medical applications, or activities in the field of nuclear industry and nuclear research. It is the task of National Metrology Institutes (NMIs) and Calibration Laboratories to calibrate radioactive sources in terms of activity and to provide activity standards appropriate to the special needs of their customers. This chapter describes the methods to calibrate the activity of radioactive materials, the different types of calibration sources, and the way to establish the traceability and international comparability of activity standards.

1 Introduction

Our known world consists of a bit more than 100 different elements. In one of the last issues of the Chart of Nuclides (Magill et al. 2006), more than 2,500 nuclides are listed, most of them are radioactive and only about 10% of them are stable nuclides. Most of the radionuclides are short-lived and do practically not occur in the environment. But about 100–200 radionuclides are of interest for research, for medical or industrial applications, or in the field of radiation protection and the surveillance of the environment. All these radionuclides have their individual decay parameters. Particle and photon emissions with different energies and emission probabilities as well as individual half-lives are characteristic for each radionuclide. Therefore, it is necessary to calibrate and provide radioactive standards for a large number of different radionuclides.

Due to the large variety of different decay schemes, the methods for the measurement of activity are quite different. These methods can be divided into secondary and primary methods. A secondary measurement device needs to be calibrated with an activity standard of a radionuclide in order to measure the activity of another source with the same radionuclide. In other words, the efficiency of the instrument for the detection of gamma rays and alpha and beta particles must be determined. Typical secondary measurement devices are germanium detectors, NaI(Tl) detectors, or ionization chambers. Primary measurement devices do not require activity standards in order to be calibrated. The efficiency of such devices is either known to be 100%, or can be determined by measurements of other physical quantities than activity (e.g., from dimensional quantities for a defined solid-angle detector), or a method is used where the knowledge of the efficiency is not required. In these cases, only some basic decay parameters of the radionuclide under study must be known in order to determine the activity of a radioactive source.

The description of calibration methods in this chapter of the handbook is limited to those methods that measure the photons or particles produced within the radioactive decay. Since the number of decaying atoms is related to the number of existing atoms, also methods measuring the number of atoms, e.g., based on mass spectrometry, could be used to determine the activity. However, these methods are not part of the descriptions here.

Primary methods are typically established at National Metrology Institutes (NMIs) worldwide. A major task of the NMIs is to establish the traceability and international comparability

of standards. Already in 1875, the "Convention of the Meter" was signed by representatives of 17 nations and the International Bureau of Weights and Measures (BIPM) was founded. The BIPM together with the NMIs is responsible for the equivalence between the national measurement standards and leads to the comparability of the calibration of a radioactive source in, e.g., the United States with calibrations in Europe or Japan.

2 Radioactive Decay

The radioactive decay of a nuclide is the phenomenon of the spontaneous conversion of one nuclide into a nuclide of another type. The primary significance of radioactive decay is the disappearance of the original nuclide. The radiations accompanying the decay in the form of photons and particles are characteristic for the decay of the various nuclides of different types and may help to identify special decay paths. Strictly speaking, the activity $A(t)$ of an amount of a nuclide in a specified energy state at a given time t is the expectation value, at that time, of the number $\mathrm{d}N(t)$ of spontaneous nuclear transitions in a time interval $\mathrm{d}t$ from that energy state. By this definition, zero activity would be equivalent to stability of the nuclide (NCRP 1985).

$$A(t) = -\frac{\mathrm{d}N(t)}{\mathrm{d}t} = \lambda N(t),$$

where λ is a constant. The derived SI unit (International System of Units) of the activity is the becquerel (Bq). 1 Bq is one decay per second. It can be shown (Evans 1955) that the decay constant λ is related to the characteristic mean life τ of a radionuclide by $\tau = 1/\lambda$, which in turn can be expressed by the half-life of that radionuclide, $T_{1/2}$, (see Chap. 10, "Radiation Protection") by

$$\lambda = \frac{\ln 2}{T_{1/2}}.$$

There exist several different possibilities to release the excess energy during a radioactive decay depending on the number of protons Z and the atomic number A of the radionuclide X. Due to the conversion, charged particles, neutrinos, neutrons, gamma rays, or X rays could be produced and emitted. The main nuclide conversions are as follows.

2.1 Alpha Decay

$$^{A}_{Z}X \rightarrow {}^{A-4}_{Z-2}Y + {}^{4}_{2}\mathrm{He}$$

With the special name α particle for the emitted helium nucleus. The helium nucleus carries two protons and two neutrons. Consequently, the atomic number A of the remaining nucleus is reduced by four and Z is reduced by two. Typically, there are not only one but several alpha transitions of different energy to the daughter nuclide possible. Therefore, α particles with different but discrete energies could be emitted.

2.2 Beta⁻ Decay

$$^{A}_{Z}X \rightarrow {}_{Z+1}{}^{A}Y + e^{-} + \tilde{\nu}_e$$

With the special name β particle for the emitted electron. In the beta decay, a neutron in the nucleus is converted into a proton, an electron, and an electron antineutrino. Due to the fact that the decay energy is distributed to the electron and the electron antineutrino, the produced electrons in a large number of decays do not have all the same discrete energy but show a continuous energy spectrum which is typical for a specific nuclide.

2.3 Beta⁺ Decay

$$^{A}_{Z}X \rightarrow {}_{Z-1}{}^{A}Y + e^{+} + \nu_e$$

In the beta^{+} decay, a proton in the nucleus is converted into a neutron, a positron, and an electron neutrino.

2.4 Electron Capture

$$^{A}_{Z}X + e^{-} \rightarrow {}_{Z-1}{}^{A}Y + \nu_e$$

The electron capture is another possibility for the conversion of a proton. An electron from the atomic shell (mainly from the inner shell) is captured, and a proton in the nucleus is converted into a neutron and an electron neutrino.

The above-mentioned decay modes change the number of protons and/or the number of neutrons in the nucleus. Typically, the resulting nucleus Y is not in its ground state but in an excited state after such a conversion. There are two possibilities for the transition from the excited state Y^* to the ground state Y. One is the gamma decay.

2.5 Gamma Decay

$$^{A}_{Z}Y^{*} \rightarrow {}^{A}_{Z}Y + \gamma$$

In this decay a gamma ray is emitted. Due to the fact that the exited states Y^* have discrete and nuclide-characteristic energy levels, the resulting gamma rays have also discrete energies. Another possibility for the transition from the excited state Y^* to the ground state Y is the internal conversion.

2.6 Internal Conversion

$$^{A}_{Z}Y^{*} \rightarrow {}^{A}_{Z}Y + e^{-}$$

The internal conversion is characterized by the emission of an electron from the atomic shell.

2.7 Other Decay Modes

The listed decay modes are the main modes relevant for the calibration of activity standards. A minor mode is the spontaneous fission which can occur in the case of nuclei with a high atomic number. In this case, the nucleus decays into two daughter nuclei and two or three neutrons.

3 Activity Standards

One major task of National Metrology Institutes (NMIs) is the realization, conservation, and dissemination of the physical units. As part of this task, the NMIs hold the national measurement standards. The national measurement standard (ISO/IEC 99 2007) stands at the top of a calibration hierarchy followed by reference measurement standards and working measurement standards typically used in a laboratory for the calibration or verification of a measuring instrument. A main feature of such a calibration hierarchy is that the measurement uncertainty necessarily increases along the sequence of calibrations.

For the dissemination of the unit of the physical quantity "activity," the NMIs produce activity standards traceable to their national measurement standards. Depending on the physical and chemical properties of the elements and the requirements of customers, radioactive standards are produced in various forms, as solid, liquid, or gaseous sources in different geometries as point sources, small- and large-area sources, or as small- and large-volume sources. The large variety of different types of radioactive standards is caused by the needs of the customers to calibrate their measurement instruments in the ideal case in the same geometry and with the same source composition as a source under investigation where the activity has to be determined. In many cases, the preparation of activity standards starts with a radionuclide solution produced in a nuclear reactor or at an accelerator facility by activation of inactive material with neutrons or charged particles. Typically, the strong activity of such a solution is quantitatively diluted in one or more steps using dedicated and calibrated balances following a dilution scheme as illustrated in ❯ *Fig. 1*.

The aim of the dilution procedure is to prepare sources suitable for different measurement devices in order to determine the activity (Sibbens and Altzitzoglou 2007) and in parallel also to produce sources which meet the needs of users for their special measurement arrangements. The major advantage is that the activities of all sources produced from the same master solution can be linked due to the quantitative mass measurements. It is a common approach that several different primary and secondary calibration methods are combined to determine the activity of the set of sources that originates from the same master solution.

❯ *Figure 2* shows a selection of typical calibration standards from a National Metrology Institute as volume, area, and point sources. The radioactive solutions in the ampoules or in the Kautex bottle can be quantitatively diluted and transferred to any customer-specific laboratory container, e.g., a Marinelli beaker. The point and area sources are produced by quantitatively dropping the radioactive solutions on a source carrier. The area sources shown in the picture are adapted to the size of filters used, e.g., for the measurement of radioactive releases to the air from nuclear power plants.

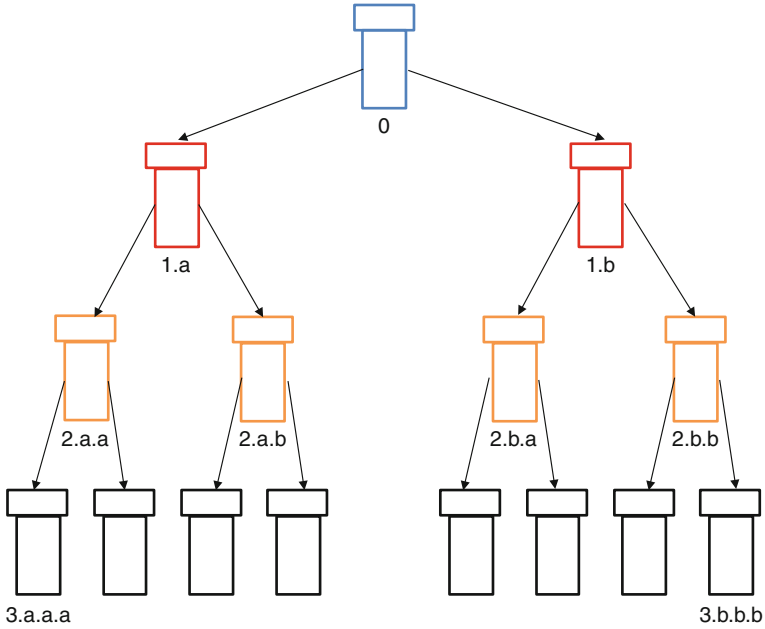


Fig. 1

Dilution scheme starting with the raw solution "0" followed by successively lower activity concentrations in the steps "1," "2," and "3," respectively



Fig. 2

Examples of different types of calibration standards (PTB 2010)

4 Primary Methods for the Calibration of Activity Standards

Primary measurement devices (Pommé 2007) do not require activity standards in order to be calibrated. That means that the efficiency of the system can be determined by measurements of physical quantities other than activity. As an example, the geometrical efficiency of a defined solid-angle spectrometer can be calculated from measurements of its geometrical dimensions. Depending on the radiation emitted during the decay of a specific radionuclide, different high-efficiency counting systems named as 4π *counting systems* exist.

For the measurement of alpha- or pure beta-particle-emitting radionuclides, an often used system is a 4π *proportional counter*. Especially designed thin sources with negligible self-absorption are placed in the center of a gas-filled proportional counter. *Figure 3* shows dropped sources on a thin VYNS foil as source support. A typical counting gas of the proportional counter is P10 (a mixture of 90% Ar and 10% CH_4). With such a system, virtually 100% of the emitted radiation is detected. However, special correction methods are necessary to determine the non-efficiency and the corresponding uncertainty.

For a special group of gamma-ray-emitting radionuclides, 4π γ *counting* can be used for the calibration. In this case, the detector is typically a large well-type NaI(Tl) detector. The method works well for multiphoton-emitting radionuclides, especially in the case that the photons are emitted in coincidence. In this case, the number of photons emitted per decay may be larger than one and the probability for the non-detection diminishes. Problems can occur if delayed transitions or decay branches with only single-photon-emitting radionuclides exist.

In the case that a beta decay is accompanied by a prompt gamma-ray transition, the traditional and widely spread 4π $\beta-\gamma$ *coincidence counting* method (Bobin 2007) can be used for the activity determination. In the classical approach, a 4π proportional counter for the measurement of the beta particle is combined with a NaI(Tl) detector for the measurement of the



Fig. 3
Especially designed thin sources with negligible self-absorption for 4π β and 4π α counting as well as for 4π β–γ coincidence counting with proportional counters

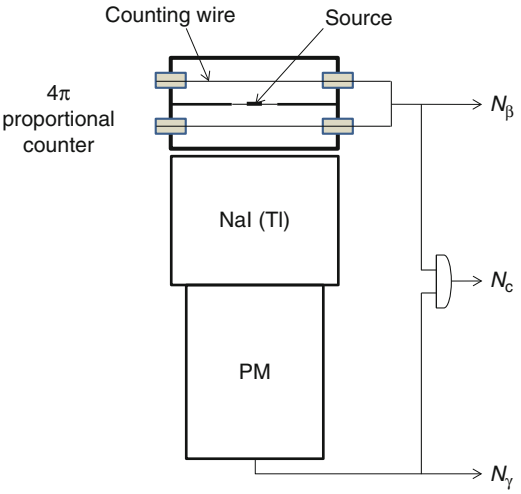


Fig. 4
Schematic drawing of a 4π β–γ coincidence counting system

gamma ray. *Figure 4* shows a schematic drawing of such a coincidence counting system. The counting rates N_β of the 4π proportional counter and N_γ of the NaI detector are measured as well as the coincidence counting rate N_c of events in both detectors in a defined time interval.

Each of the three counting rates N_β, N_γ, and N_c is proportional to the "unknown activity" and the respective counting efficiencies:

$$N_\beta = A(t) \cdot \epsilon_\beta ,$$

$$N_\gamma = A(t) \cdot \epsilon_\gamma ,$$

$$N_c = A(t) \cdot \epsilon_\beta \cdot \epsilon_\gamma .$$

These three equations can be combined to:

$$A(t) = \frac{N_\beta \cdot N_\gamma}{N_c} ,$$

where the unknown activity is defined only by the three measured counting rates.

The coincidence counting system has the advantage that the activity of the radioactive source is determined without any explicit knowledge of the counting efficiencies of the two detectors. However, the described approach works only if the beta detector is not sensitive for gamma radiation. This is obviously not always true. Compton scattering of gamma rays in the beta detector, for example, can produce a signal in the beta detector. The three simple expressions for the counting rates have then to be modified,

$$N_\beta = A(t) \cdot \left[\epsilon_\beta + (1 - \epsilon_\beta) \left[\frac{1}{1 + \alpha_T} \left(\alpha_T \cdot \varepsilon_{ce} + \varepsilon_{\beta\gamma} \right) \right] \right],$$

$$N_\gamma = \frac{A(t) \cdot \epsilon_\gamma}{1 + \alpha_T},$$

$$N_c = A(t) \cdot \left[\frac{\epsilon_\beta \cdot \epsilon_\gamma}{1 + \alpha_T} + (1 - \epsilon_\beta) \cdot \varepsilon_c \right],$$

where α_T is the total internal conversion coefficient, ε_{ce} is the detection efficiency for conversion electrons in the beta detector, $\varepsilon_{\beta\gamma}$ is the detection efficiency for gamma rays in the beta detector, and ε_c is the probability of observing additional coincidences from gamma rays first detected in the beta detector. The above-mentioned ratio of counting rates does not any more give the activity directly. However, this ratio is a smooth function of the efficiency of the beta detector and the ratio N_c/N_γ in many cases is a very good approximation of ε_β. This function may be followed by variation of the efficiency of the beta detector which can be done by successively adding different absorbers around the radioactive source or by variation of the discrimination level in the respective counting chain. From an appropriate extrapolation of the measurement results to unity in N_c/N_γ or, equivalently, to zero in the measured inefficiency $(1 - \varepsilon_\beta)$, the activity can be determined. *Figure 5* shows as an example the result of a measurement series on a ^{177}Lu source. The coincidence method described above also works for α and e^+ emitters and even for electron-capture nuclides. It should be mentioned, however, that this method can be very laborious and time consuming in certain cases.

Other primary methods are based on special liquid scintillation counting (LSC) methods (Broda et al. 2007) in the case that the counting efficiency could be calculated. Physicochemical processes in the light emission and the statistics of the photon emission are taken into account

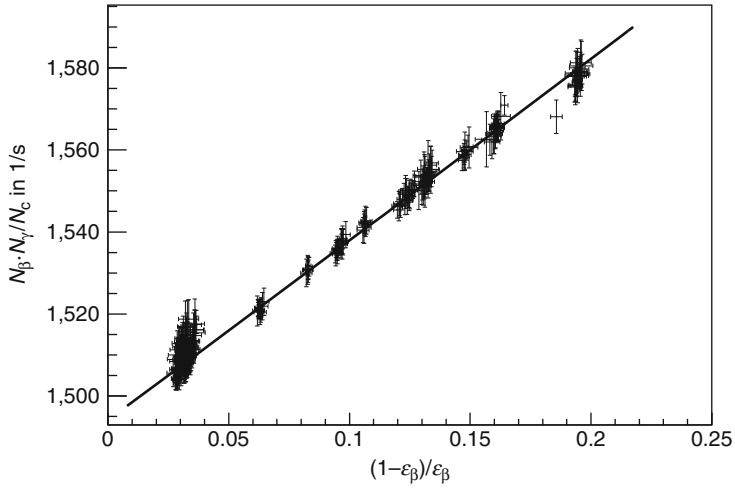


Fig. 5

Result of a 4π β–γ coincidence counting measurement of a ^{177}Lu source with a variation of the efficiency of the beta detector by adding different absorbers. The extrapolation of the fitted curve to $(1 - \varepsilon_\beta)/\varepsilon_\beta = 0$ gives the activity of the measured source

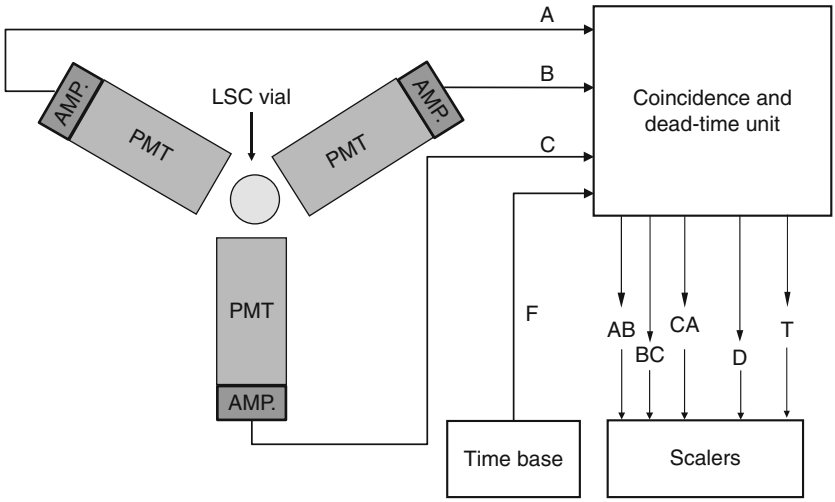


Fig. 6
Schematic drawing of the Triple-to-Double Coincidence Ratio (TDCR) system with three photomultiplier tubes (PMT) and amplifiers around an LSC vial with the radioactive solution and the logical unit to measure triple (T) and double (D) coincidences

in a so-called free-parameter model. The CIEMAT/NIST method uses an activity standard as tracer to determine the free parameters. In most cases, the low-energy β emitter H-3 is used for this purpose. In the so-called Triple-to-Double Coincidence Ratio method (TDCR), these parameters are calculated from ratios of measured counting rates of double and triple coincidences (*Fig. 6*). One advantage of both methods is that the source preparation is much easier compared to other methods. The radioactive solution is directly added to a liquid scintillation cocktail, and the self-absorption in the source is negligible. The achievable measurement uncertainties are comparable, in some cases even superior to those obtained by the classical coincidence counting method.

Another method where the efficiency can be calculated is the Defined Solid Angle (DSA) counting for alpha-particle-emitting radionuclides (Pommé 2007). The alpha-particle source with a typical diameter of 1–2 cm is placed in a defined distance to a diaphragm of known shape in front of a solid-state detector. The solid angle is therefore well defined by dimensional quantities and the detection efficiency of the alpha detector is one for all alpha particles passing through the diaphragm.

5 Secondary Methods for the Calibration of Activity Standards

Most of the primary standardization methods are time consuming and need sophisticated source preparation. Therefore, the primary methods are often supplemented by secondary methods. The "workhorses" in many metrology institutes are the ionization chambers (Schrader 2007). The re-entrant pressurized ionization chambers (*Fig. 7*) are very stable and reproducible measuring devices over many years. In addition, the ionization chambers show a

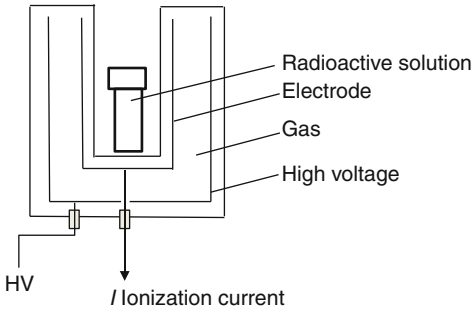


Fig. 7
Schematic drawing of a 4π ionization chamber (IC)

very good linearity as a function of the activity. The easy-to-use chambers are calibrated with radionuclide solutions in ampoules. The determined calibration factors can be used on a long time scale for the determination of the activity for the same radionuclide under the same measurement conditions. With other words, ionization chambers once calibrated with a specific radionuclide are used to preserve the activity unit for that radionuclide for many years.

High-purity germanium spectrometers are also widely used as secondary measurement devices for the calibration of activity standards. A set of gamma-ray-emitting radionuclides is used to calibrate a germanium detector over a whole energy range. With such a calibrated detector, the activity for an unknown source of a radionuclide used for the calibration can be determined as well as the activity of a gamma-ray-emitting radionuclide with photon energies that fall in the calibrated energy interval. In addition, the germanium detector has a good energy resolution and can be used to measure not only one but several radionuclides at a time in one source. Gammaspectrometric measurements with germanium detectors are therefore a common tool to measure the "main" radionuclide and potential impurities at the same time.

NaI(Tl) detectors are also commonly used as secondary measurement device. The cheap and easy-to-use detector systems are also used to preserve the activity unit for gamma-ray-emitting radionuclides for many years.

It should be noted that all secondary instruments are suited for activity measurements only if the same measurement conditions are met as for the calibration. Otherwise corrections factors have to be established which allow to calculate the activity of a source from measurements under non-calibrated conditions.

6 International Comparability of Activity Standards

A major task of the NMIs is to establish the traceability and international comparability of standards. Already in 1875, the "Convention of the Meter" was signed by representatives of 17 nations and the International Bureau of Weights and Measures (BIPM) was founded. Based on this convention, the "Système International de Référence (SIR)" (Ratel 2007) was established at the BIPM to compare activity standards from different national metrology laboratories. The system comprises two pressurized ionization chambers. The advantage of the system is that the comparison of the result of two or more laboratories need not be done at the same time. The system

measures the ionization current of a standard provided by an NMI together with the activity. The ratio of the measured ionization current to the activity provided by the NMI is a measure of the efficiency for the specific radionuclide of interest. These figures can be compared for a variety of radionuclides among the participating laboratories in order to define the degree of equivalence of measurements in the different countries. In 1999, the directors of the national metrology institutes of 38 Member States of the BIPM and representatives of two international organizations signed a Mutual Recognition Arrangement (CIPM MRA 1999) for national measurement standards and for calibration and measurement certificates issued by NMIs. A number of other institutes have signed since then. This arrangement demands a sophisticated quality management system at the national metrology institutes and a larger number of comparisons in order to validate the claims of the NMIs for their calibration and measurement capabilities (CMC 2010). In principle, the NMIs are forced to perform comparison for each radionuclide. However, due to the large number of different radionuclides, a system was established that groups radionuclides together with specific measurement techniques (Karam 2007) in order to limit the number of comparisons to a smaller number. The SIR system together with the regulations of the MRA and the performed key comparisons (KCDB 2010) guarantees the comparability and equivalence of activity measurements between the national metrology laboratories worldwide. In addition, the accreditation of calibration laboratories in the different countries guarantees the traceability of activity measurements to the standards of NMIs. As one result of this complex system, a customer who receives an activity standard from a calibration laboratory can retrace how different or equal the activity measurement is done for the same radionuclide in another country.

7 Conclusions

The calibration of radioactive sources is a service offered by national metrology institutes as well as from accreted calibration laboratories. The costumer receives standards with traceable activity, reliable uncertainties, and with the confirmation that the measurement methods and results are comparable to those obtained in other countries. The usage of such radioactive standards is therefore an essential part of the quality system in research institutes. It is one main contribution in order to compare the results of research groups working in different institutes.

8 Cross-References

- Chapter 1, “Interactions of Particles and Radiation with Matter”
- Chapter 6, “Particle Identification”
- Chapter 8, “Synchrotron Radiation and FEL Instrumentation”
- Chapter 10, “Radiation Protection”
- Chapter 11, “Gaseous Detectors”
- Chapter 12, “Tracking Detectors”
- Chapter 13, “Photon Detectors”
- Chapter 15, “Scintillation Counters”
- Chapter 16, “Semiconductor Counters”
- Chapter 17, “Gamma-Ray Detectors”
- Chapter 21, “New Solid State Detectors”

- Chapter 22, "Radiation Damage Effects"
- Chapter 25, "Technology for Border Security"
- Chapter 27, "Geoscientific Applications of Particle Detection and Imaging Techniques with Special Focus on the Monitoring Clay Mineral Reactions"
- Chapter 28, "Particle Detectors Used in Isotope Ratio Mass Spectrometry, with Applications in Geology, Environmental Science and Nuclear Forensics"
- Chapter 29, "Particle Detectors in Materials Science"
- Chapter 32, "Instrumentation for Nuclear Fusion"
- Chapter 33, "The Use of Neutron Technology in Archaeological and Cultural Heritage Research"
- Chapter 34, "Radiation Detectors and Art"
- Chapter 35, "Radiation-Based Medical Imaging Techniques: An Overview"
- Chapter 37, "SPECT Imaging: Basics and New Trends"
- Chapter 38, "PET Imaging: Basics and New Trends"
- Chapter 41, "Quantitative Image Analysis in Tomography"
- Chapter 43, "Evaluation and Image Quality in Radiation-Based Medical Imaging"

References

Bobin C (2007) Primary standardization of activity using the coincidence method based on analogue instrumentation. Metrologia 44:S27–S31

Broda R, Cassette P, Kossert K (2007) Radionuclide metrology using liquid scintillation counting. Metrologia 44:S36–S52

CIPM MRA (1999) CIPM Mutual Recognition Arrangement. http://www.bipm.org/pdf/mra.pdf. Accessed 24 Feb 2011

CMC (2010) The BIPM database of the calibration and measurement capabilities http://kcdb.bipm.org/appendixC/. Accessed 24 Feb 2011

Evans RD (1955) The Atomic Nucleus. McGraw-Hill, New York

ISO/IEC 99 (2007) Guide 99: international vocabulary of metrology – Basic and general concepts and associated terms (VIM). ISO, Geneva

Karam L (2007) Application of the CIPM MRA to radionuclide metrology. Metrologia 44:S1–S6

KCDB (2010) The BIPM key comparison database http://kcdb.bipm.org/AppendixB/KCDB_ApB_search.asp. Accessed 24 Feb 2011

Magill J, Pfennig G, Galey J (2006) Karlsruher Nuklidkarte, 7th edn. Haberbeck, Lage/Lippe. European Commission, report EUR 22276 EN

NCRP (1985) National council on radiation protection and measurements; NCRP report No. 58, A handbook of radioactivity measurements procedures. National Council on Radiation Protection and Measurements, Bethesda

Pommé S (2007) Methods for primary standardization of activity. Metrologia 44:S17–S26

PTB (2010) Catalogue of activity standards. http://www.ptb.de/en/org/6/61/611/_index.htm. Accessed 24 Feb 2011

Ratel G (2007) The Système International de Référence and its application in key comparisons. Metrologia 44:S7–S16

Schrader H (2007) Ionization chambers. Metrologia 44:S53–S66

Sibbens G, Altzitzoglou T (2007) Preparation of radioactive sources for radionuclide metrology. Metrologia 44:S71–S78

10 Radiation Protection

Claus Grupen
Siegen University, Siegen, Germany

1 Introduction ***203***

2 Units of Radiation Protection ***204***

3 Basic Nuclear Physics ***207***

4 Basic Interactions ***210***

5 Radiation Sources ***213***
5.1 Particle Radiation 213
5.2 Photon Sources 215
5.3 Neutron Sources 216
5.4 Cosmic-Ray Sources 217

6 Radiation Detectors ***218***
6.1 Ionization Chambers 218
6.2 Proportional Counters and Geiger–Müller Counters 218
6.3 Scintillation Counters 220
6.4 Semiconductor Counters 222
6.5 Neutron Dosimeters 223
6.6 Personal Dosimeters 223

7 Safety Standards ***227***

8 Organization of Radiation Protection ***227***

9 Environmental Radiation ***229***

10 Biological Effects of Radiation ***232***

11 Conclusions ***234***

Acknowledgment ***234***

References ***234***

C. Grupen, I. Buvat (eds.), *Handbook of Particle Detection and Imaging*,
DOI 10.1007/978-3-642-13271-1_10, © Springer-Verlag Berlin Heidelberg 2012

Further Reading 234

Suppliers of Radiation-Protection Equipment 235

Abstract: Radiation protection is a very important aspect for the application of particle detectors in many different fields, like high energy physics, medicine, materials science, oil and mineral exploration, and arts, to name a few. The knowledge of radiation units, the experience with shielding, and information on biological effects of radiation are vital for scientists handling radioactive sources or operating accelerators or X-ray equipment. This article describes the modern radiation units and their conversions to older units which are still in use in many countries. Typical radiation sources and detectors used in the field of radiation protection are presented. The legal regulations in nearly all countries follow closely the recommendations of the International Commission on Radiological Protection (ICRP). Tables and diagrams with relevant information on the handling of radiation sources provide useful data for the researcher working in this field.

1 Introduction

Radiation is everywhere. Radiation emerges from the soil, it is in the air, and our planet is constantly bombarded by energetic cosmic radiation. Even the human body is radioactive: About 9,000 decays of unstable nuclei occur per second in the human body.

Since the beginning of the twentieth century, mankind has been able to transform nuclei (Ernest Rutherford) and to artificially create new radioactive nuclei, in particular, since the discovery of nuclear fission in the late 1930s. Since one cannot "see" or "smell" ionizing radiation, one needs measurement devices that can detect it, and also a scale by which to judge on its possible dangerous effects. This leads to the definition of units for the activity of radioactive nuclei and quantifications for the effect on humans in terms of absorbed energy and biological effectiveness of different types of radiation. Radiation protection is a truely interdisciplinary field. It concerns, among others, physicists, engineers, lawyers, and health-care professionals.

Radioactivity was discovered by Henri Antoine Becquerel in 1896, when he realized that radiation emerging from uranium ores could blacken photosensitive paper. Originally it was believed that this was due to some fluorescence radiation from uranium salts. However, the photosensitive film was also blackened without previous exposure of the uranium ore to light. The radiation spontaneously emerging from uranium was not visible to the human eye. Therefore, it was clear that one was dealing with a new phenomenon. In the context of radiation protection also the discovery of X rays by Wilhelm Conrad Röntgen has to be mentioned. This radiation emerged from materials after bombardment with energetic electrons. Actually the discovery by Röntgen in December 1895 had been a factor of stimulating Becquerel to investigate fluorescence radiation from uranium salts.

This radioactivity was a phenomenon of the natural environment. Nobody was able to turn inactive materials into radioactive sources by chemical techniques. Not until 1934 did Frederic Joliot and Irène Curie manage to produce new radioactive materials using nuclear physics methods. Only a few years later Otto Hahn and Fritz Straßmann (1938/1939) succeeded in inducing fission of uranium nuclei.

The importance of radioactivity and of radiation protection for mankind and the environment is quite substantial.

2 Units of Radiation Protection

The unit of activity is becquerel (Bq). 1 Bq is one decay per second. The old unit curie (Ci) corresponds to the activity of 1 g of radium-226:

$$1\,\mathrm{Ci} = 3.7 \times 10^{10}\,\mathrm{Bq}\,,$$
$$1\,\mathrm{Bq} = 27 \times 10^{-12}\,\mathrm{Ci} = 27\,\mathrm{pCi}\,. \tag{1}$$

The radioactive decay law

$$N = N_0\,\mathrm{e}^{-\lambda t} \tag{2}$$

describes the decrease of nuclei in time. The decay constant λ is related to the *lifetime* of the radioactive source as

$$\lambda = \frac{1}{\tau}\,. \tag{3}$$

One has to distinguish the half-life $T_{1/2}$ from the lifetime. The half-life is the time after which a half of the initially existing atomic nuclei has decayed:

$$N(t = T_{1/2}) = \frac{N_0}{2} = N_0\,\mathrm{e}^{-T_{1/2}/\tau}. \tag{4}$$

Therefore, we have

$$T_{1/2} = \tau\,\ln 2\,. \tag{5}$$

Correspondingly, the decay constant is related to the half-life by

$$\lambda = \frac{1}{\tau} = \frac{\ln 2}{T_{1/2}}\,. \tag{6}$$

The activity A of a radioactive source characterizes the number of decays per second:

$$A = -\frac{\mathrm{d}}{\mathrm{d}t}(N_0\,\mathrm{e}^{-\lambda t}) = \lambda\,N_0\,\mathrm{e}^{-\lambda t} = \lambda\,N = \frac{1}{\tau}N\,. \tag{7}$$

The activity in Bq does not say very much about possible biological effects. These are related to the deposited energy by the radioactive source in matter. The energy dose (absorbed energy ΔW per mass unit Δm),

$$D = \frac{\Delta W}{\Delta m} = \frac{1}{\rho}\frac{\Delta W}{\Delta V} \tag{8}$$

(ρ – density, ΔV – volume element), is measured in gray:

$$1\ \text{gray (Gy)} = 1\ \text{joule (J)}\ /\ 1\ \text{kilogram (kg)}. \tag{9}$$

Gray is related to the old unit rad (radiation absorbed dose, 1 rad = 100 erg/g; still in use in the USA) according to (1 joule (J) = 1 watt second (W s) = $1\,\frac{\mathrm{kg\,m^2}}{\mathrm{s^2}} = 10^7\,\frac{\mathrm{g\,cm^2}}{\mathrm{s^2}} = 10^7$ erg):

$$1\,\mathrm{Gy} = 100\,\mathrm{rad} = 10^4\,\mathrm{erg/g}\,. \tag{10}$$

In terms of deposited energy in units popular in particle physics and medicine one has

$$1\,\mathrm{Gy} = 6.24 \times 10^{12}\,\mathrm{MeV/kg}\,. \tag{11}$$

For indirectly ionizing radiation (i.e., photons and neutrons, but not electrons and other charged particles) a further quantity characterizing the energy dose, the "kerma," is defined. Kerma is an abbreviation for Kinetic Energy Released per unit Mass. (Occasionally one also finds Kinetic Energy Released in Matter (or: in Material)). The kerma k is defined as the sum of the initial energies of all charged particles, ΔE, liberated in a volume element ΔV by indirectly ionizing radiation divided by the mass Δm of this volume element:

$$k = \frac{\Delta E}{\Delta m} = \frac{1}{\rho}\frac{\Delta E}{\Delta V}, \tag{12}$$

where ρ is the density of the absorbing material.

Kerma relates only to the energy transferred to the charged particles: It does not depend on which fraction of the energies of the charged particles is transported out of the volume by particle motion or by bremsstrahlung. Therefore, kerma is sometimes also called dose unit of the first interaction step. The unit of kerma is also gray (Gy).

Gray and rad describe the pure physical energy absorption. These units cannot easily be translated into the biological effect of radiation. Electrons, for example, ionize relatively weakly while, in contrast, α rays are characterized by a high ionization density. Therefore, biological repair mechanisms cannot be very effective in the latter case. The relative biological effectiveness depends on the type of radiation, the radiation energy, the temporal distribution of the dose, and other quantities.

For the radiation field R one obtains the dose equivalent H_R from the energy dose D_R according to

$$H_R = w_R\, D_R, \tag{13}$$

where w_R is the radiation weighting factor. This dose equivalent is measured in sievert (Sv). In the same way the old energy dose unit *rad* is converted to roentgen equivalent man "rem" by

$$H_R[\text{rem}] = w_R\, D_R[\text{rad}]. \tag{14}$$

Correspondingly, the relation

$$1\,\text{Sv} = 100\,\text{rem} \tag{15}$$

holds. The radiation weighting factors w_R, as recommended by the International Commission on Radiological Protection ICRP, are shown in ❯ *Table 1*. In the USA slightly different radiation weighting factors are used.

Furthermore, dose units for penetrating external radiation (depositing most of their energy in the first 10 mm of tissue) and for radiation of low penetration depth (70 µm skin depth) have been introduced in many national radiation-protection regulations. In personal dosimetry these operative units are denoted with $H_\text{P}(10)$, $H_\text{P}(0.07)$.

Apart from these units another quantity is used for the amount of created charge, the roentgen (R). One roentgen is that radiation dose of X rays and γ rays which liberates one electrostatic charge unit of electrons and one of ions in 1 cm^3 of air (at standard temperature and pressure). If the unit roentgen is expressed by the ion dose I in coulomb/kilogram, one obtains

$$1\,\text{R} = 2.58 \times 10^{-4}\,\text{C/kg} \tag{16}$$

or, equivalently,

$$1\,\text{C/kg} = 3{,}880\,\text{R}. \tag{17}$$

Table 1
Radiation weighting factors w_R

Type of radiation and energy range	Radiation weighting factor w_R
Photons, all energies	1
Electrons and muons,[a] all energies	1
Neutrons $E_n < 10$ keV	5
Neutrons 10 keV $\leq E_n \leq$ 100 keV	10
Neutrons 100 keV $< E_n \leq$ 2 MeV	20
Neutrons 2 MeV $< E_n \leq$ 20 MeV	10
Neutrons with $E_n >$ 20 MeV	5
Protons, except recoil protons, $E >$ 2 MeV	5
α particles, fission fragments, heavy nuclei	20

[a] Muons are short-lived elementary particles that are produced predominantly in cosmic radiation.

The tissue equivalent of roentgen is given by

$$1\,\mathrm{R} = 0.88\,\mathrm{rad} = 8.8\,\mathrm{mGy}. \tag{18}$$

In many cases it is necessary to convert a partial-body dose into a whole-body dose. Therefore, a weighting factor w_T has to be attributed to the irradiated organs of the body. This effective dose equivalent E is defined as

$$E = H_{\mathrm{eff}} = \sum_{T=1}^{n} w_T H_T\,, \tag{19}$$

where H_T is the average dose equivalent in the irradiated organ or tissue and w_T is the weighting factor for the Tth organ or tissue.

The tissue weighting factors, according to the recommendation of the ICRP (from 2006), are given in *Table 2*.

It is assumed that the inhomogeneous irradiation of the body with an effective dose equivalent H_{eff} bears the same radiation risk as a homogeneous whole-body irradiation with $H = H_{\mathrm{eff}}$.

For the general case of partial-body irradiation in a complex radiation field one has

$$E = H_{\mathrm{eff}} = \sum_{T} w_T\, H_T = \sum_{T} w_T \sum_{R} w_R\, D_{T,R}. \tag{20}$$

The determination of the dose-equivalent rate by a pointlike radiation source of activity A can be accomplished using the following formula:

$$\dot{H} = \Gamma \frac{A}{r^2}\,. \tag{21}$$

In this equation, r is the distance from the radiation source (in meters) and Γ is a specific radiation constant that depends on the type and energy of the radiation. The dose-rate radiation constants are specific for each radioisotope, and they are different for β and γ radiation. *Table 3* lists the dose-rate constants for some commonly used radiation sources. (The values given in the literature show a somewhat spread on the order of 10% or sometimes even more. Some authors quote the dose-rate constants to four digits. In view of the obvious uncertainties,

Table 2
Tissue weighting factors w_T

Organ or tissue	Tissue weighting factor w_T
Gonads	0.08
Red bone marrow	0.12
Colon	0.12
Lung	0.12
Stomach	0.12
Breast	0.12
Bladder	0.04
Liver	0.04
Esophagus	0.04
Thyroid gland	0.04
Skin	0.01
Periosteum (bone surface)	0.01
Brain	0.01
Salivary glands	0.01
Other organs or tissue	0.12

only two digits are given here. The numbers for the γ dose-rate constants in this table are taken from the ORNL/RSIC-45/R1 report (Unger and Trubey 1982) from the Oak Ridge National Laboratory, except the one for ^{133}Xe. The constants can also be calculated from Eq. 21 by using the known γ energies of the different isotopes along with their branching ratios, and the known mass absorption coefficients for a source activity of 1 Bq at a distance of 1 m. In this way the γ dose-rate constant for ^{133}Xe was determined.)

3 Basic Nuclear Physics

An atomic nucleus of mass number A consists of Z protons and N neutrons: $A = Z + N$. The atomic number Z of stable nuclei is related to the atomic mass approximately according to

$$Z_{\text{stable}} = \frac{A}{1.98 + 0.0155\, A^{2/3}}\,. \tag{22}$$

For light nuclei ($Z \leq 20$, calcium) $Z = A/2$ holds; for heavy nuclei one has approximately $Z = A/2.5$. The atomic number Z characterizes the chemical properties of an atom. Nuclei with fixed Z and variable N are called isotopes. If the isotopes are radioactive they are called radioisotopes. Nuclei with a fixed sum of protons and neutrons, i.e., constant mass number A, are called isobars. Nuclei with fixed neutron but varying proton number are called isotones. Protons and neutrons are approximately of the same mass, $m_{\text{neutron}}/m_{\text{proton}} = 1.001\,38$.

Nuclei with excess neutrons are normally β^- emitters. In this nuclear process a neutron (n) transforms into a proton (p) under emission of an electron (e^-) and an electron antineutrino ($\overline{\nu}_e$),

$$n \longrightarrow p + e^- + \overline{\nu}_e. \tag{23}$$

Table 3

Dose-rate constants Γ for some β- and γ-ray emitters (Grupen 2010; Krieger 2002; Sauter 1982; Unger and Trubey 1982)

Radioisotope	β dose-rate constant $\left(\frac{\text{Sv m}^2}{\text{Bq h}}\right)$
$^{32}_{15}$P	9.1×10^{-12}
$^{60}_{27}$Co	2.6×10^{-11}
$^{90}_{38}$Sr	2.0×10^{-11}
$^{131}_{53}$I	1.7×10^{-11}
$^{204}_{81}$Tl	1.3×10^{-11}
Radioisotope	**γ dose-rate constant $\left(\frac{\text{Sv m}^2}{\text{Bq h}}\right)$**
$^{11}_{6}$C	1.9×10^{-13}
$^{22}_{11}$Na	3.6×10^{-13}
$^{24}_{11}$Na	5.1×10^{-13}
$^{41}_{18}$Ar	1.9×10^{-13}
$^{40}_{19}$K	2.2×10^{-14}
$^{51}_{24}$Cr	6.3×10^{-15}
$^{54}_{25}$Mn	1.4×10^{-13}
$^{56}_{25}$Mn	2.5×10^{-13}
$^{59}_{26}$Fe	1.8×10^{-13}
$^{57}_{27}$Co	4.1×10^{-14}
$^{60}_{27}$Co	3.7×10^{-13}
$^{65}_{28}$Ni	8.0×10^{-14}
$^{65}_{30}$Zn	8.9×10^{-14}
$^{85}_{36}$Kr	4.2×10^{-16}
$^{95}_{40}$Zr	1.3×10^{-13}
$^{99m}_{43}$Tc	3.3×10^{-14}
$^{110m}_{47}$Ag	4.5×10^{-13}
$^{124}_{51}$Sb	2.9×10^{-13}
$^{125}_{53}$I	7.4×10^{-14}
$^{131}_{53}$I	7.6×10^{-14}
$^{133}_{54}$Xe	1.4×10^{-14}
$^{133m}_{54}$Xe	3.0×10^{-14}
$^{137}_{55}$Cs	1.0×10^{-13}
$^{133}_{56}$Ba	1.2×10^{-13}
$^{152}_{63}$Eu	2.0×10^{-13}
$^{226}_{88}$Ra	3.3×10^{-15}
$^{226}_{88}$Ra in equilibrium with its decay products	2.2×10^{-13}

Free neutrons have a lifetime of 886 s (Amsler et al. 2008). Light nuclei with excess protons are mostly β^+ emitters. In this case a proton decays into a neutron under emission of a positron (e^+) and a neutrino (ν_e),

$$p \longrightarrow n + e^+ + \nu_e. \tag{24}$$

Free protons are stable ($\tau_p > 10^{31}$ years) since there are no lighter baryons into which they could decay. In β^+ emitters also electron capture can occur:

$$p + e^- \longrightarrow n + \nu_e. \tag{25}$$

β decays frequently lead to excited nuclear levels of the final-state nucleus ("daughter nucleus"). The excited daughter nucleus de-excites into the ground state under the emission of γ rays. Since the energy difference between the excited nucleus and the ground state is fixed, γ rays, in contrast to decay electrons from nuclear β decay, have a discrete energy.

An example for β decay is

$$\begin{array}{l} {}^{137}_{55}\mathrm{Cs} \longrightarrow {}^{137}_{56}\mathrm{Ba}^* + e^- + \overline{\nu}_e \\ \qquad\quad \hookrightarrow\ {}^{137}_{56}\mathrm{Ba} + \gamma \end{array}. \tag{26}$$

The electron can receive a maximum energy of 0.51 MeV (1.17 MeV if the decay proceeds directly into the ground state). The photon from the decay of the excited Ba^* has an energy of 662 keV. All this information is best summarized in a decay-level diagram (see ❯ *Fig. 1*).

Heavy, high-mass nuclei tend to decay under the emission of an α particle, i.e., a helium nucleus. This decay mode is frequently in competition to β^+ decay, but the proton excess for most heavy nuclei can more easily be reduced by the emission of α particles. Also, there are theoretical reasons (nuclear shell model, strong binding of helium nuclei) why α decay is usually favored. Thus, for example, ${}^{238}_{92}\mathrm{U}$ decays into excited states of the ${}^{234}_{90}\mathrm{Th}$ isotope. Since the

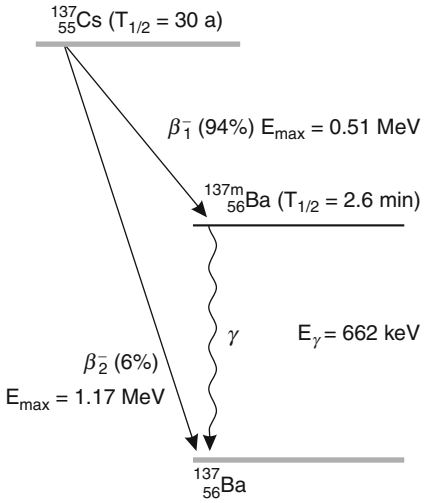


Fig. 1
Decay-level diagram of ${}^{137}_{55}\mathrm{Cs}$ (Grupen 2010)

nuclear levels are characterized by fixed energies, the emitted α particles in this two-body decay are monoenergetic. Apart from α emission, heavy nuclei (for $Z \geq 90$) can also decay by spontaneous fission.

As a result of a nuclear transmutation atomic electrons can also be emitted: An atomic nucleus will normally release its excess energy E_{ex} by γ emission. However, it is also possible that its excitation energy is transferred directly onto an electron in the atomic shell. Such electrons are called *conversion electrons*. If a vacancy in the atomic shell is produced by electron capture or conversion, the electrons in the atomic shell will try to reach a more favorable energetic state. In this way a vacancy in the K shell can be filled up by an electron from the L shell. The energy difference $E_K - E_L$ is liberated and can either be emitted as characteristic X rays with $E_X = E_K - E_L$ or, if $E_K - E_L > E_L$, it can be transferred directly to another L electron that will leave the atom with the energy $E_K - 2E_L$. Such an electron is called *Auger electron*.

4 Basic Interactions

Charged particles (electrons, positrons, protons, helium nuclei, etc.) will ionize matter in a direct way, in contrast to neutral particles (neutrons, neutrinos, etc.) or short-wavelength electromagnetic radiation (X rays and γ rays), which are ionizing only indirectly. Strictly speaking, radiation is never directly measured, rather it can only be detected via its interaction with matter. A large number of specific interaction processes exists. These interactions are characteristic for charged particles, neutrons, neutrinos, X rays, and γ radiation.

Charged particles lose their energy essentially by excitation and ionization and by bremsstrahlung.

In the field of practical radiation protection it is sometimes desirable to consider only the local energy deposit, i.e., collisions with relatively low energy transfer. The idea is that in collisions with large energy transfers long-range δ electrons are created that are weakly ionizing and therefore have only little biological effect, in contrast to the high ionzation density generated by low energy transfers. (The energy spectrum of ionization electrons falls off like $1/\epsilon^2$, where ϵ is the kinetic energy of the electrons knocked out from the atom. These electrons are also called "knock-on electrons" or δ rays.) This led to the introduction of the concept of the linear energy transfer (LET). The LET of charged particles is the ratio of the average energy loss ΔE, where only collisions with energy transfers smaller than a given cut-off parameter E_{cut} are considered, and the traversed distance Δx,

$$\mathrm{LET} = L_{E_{cut}} = \left(\frac{\Delta E}{\Delta x}\right)_{E_{cut}} . \tag{27}$$

The energy cut parameter is usually given in eV. A value of LET_{100} indicates that only collisions with energy transfers below 100 eV are considered. High linear-energy-transfer radiation corresponds to high biological effectiveness, and this type of radiation is very important for cancer treatment. At the end of their range, charged particles produce a strong ionization peak ("Bragg peak") which is a consequence of the $1/v^2$ dependence of the energy loss (v – velocity of the charged particle); see *Fig. 2*.

Neutrons are of particular importance for radiation protection because of their high relative biological effectiveness. For neutrons with energies that are typical in the field of radiation

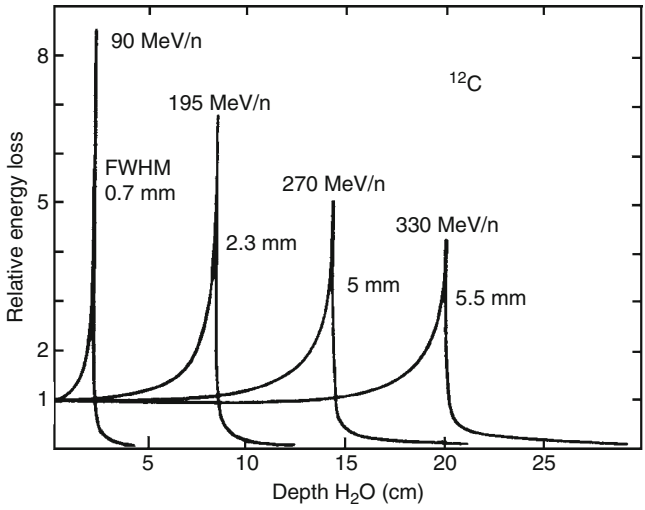


Fig. 2
Energy loss of carbon ions ^{12}C in water as a functions of depth (Grupen and Shwartz 2008; Kraft 1996, 2000)

protection ($E_{\text{kin}} \leq 10$ MeV), the following detection reactions can be considered:

$$n + {}^{6}_{3}\text{Li} \longrightarrow \alpha + {}^{3}_{1}\text{H}, \tag{28}$$

$$n + {}^{10}_{5}\text{B} \longrightarrow \alpha + {}^{7}_{3}\text{Li}, \tag{29}$$

$$n + {}^{3}_{2}\text{He} \longrightarrow p + {}^{3}_{1}\text{H}, \tag{30}$$

$$n + p \longrightarrow n + p. \tag{31}$$

Just as with neutrons, photons must first produce charged particles in an interaction process, which are then normally detected via processes of ionization, excitation, and scintillation. The interactions of photons are fundamentally different from those of charged particles since in a photon interaction process the photon is either completely absorbed (photoelectric effect, pair production) or scattered through relatively large angles (Compton effect). Photons are absorbed in matter according to

$$I = I_0 \, \mathrm{e}^{-\mu x}, \tag{32}$$

where μ is the mass attenuation coefficient. For the Compton effect the photon survives the interaction, and consequently one has to distinguish the mass absorption coefficient from the mass attenuation coefficient. For that purpose one defines the Compton scattering coefficient:

$$\mu_{\text{cs}} = \frac{E'_\gamma}{E_\gamma} \, \mu_{\text{Compton effect}}. \tag{33}$$

In this relation, E_γ and E'_γ are the energies of the photons before and after scattering in the Compton process. The Compton absorption coefficient is then the difference between the total

probability for the Compton effect and the Compton scattering coefficient:

$$\mu_{ca} = \mu_{\text{Compton effect}} - \mu_{cs} . \tag{34}$$

The attenuation or absorption coefficients are frequently normalized to the area density:

$$\mu(\text{cm}^{-1}) = \mu\left((\text{g/cm}^2)^{-1}\right) \rho \tag{35}$$

(ρ – density of the absorber in g/cm^3).

Figure 3 shows the mass attenuation and mass absorption coefficients for photons in lead.

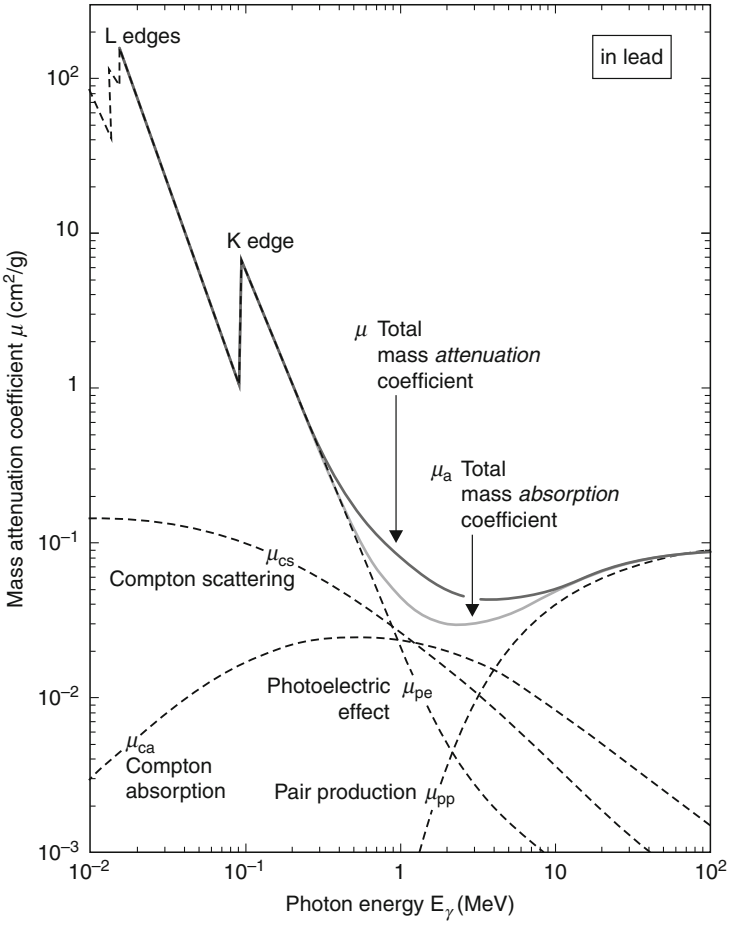


Fig. 3

Energy dependence of the mass attenuation coefficient μ and mass absorption coefficient μ_a for photons in lead. μ_{pe} describes the photoelectric effect, μ_{pp} pair production, μ_{cs} Compton scattering, and μ_{ca} Compton absorption. μ_a is the total mass absorption coefficient ($\mu_a = \mu_{pe} + \mu_{pp} + \mu_{ca}$) and μ the total mass attenuation coefficient ($\mu = \mu_{pe} + \mu_{pp} + \mu_c$ with $\mu_c = \mu_{cs} + \mu_{ca}$) (Grupen 2010)

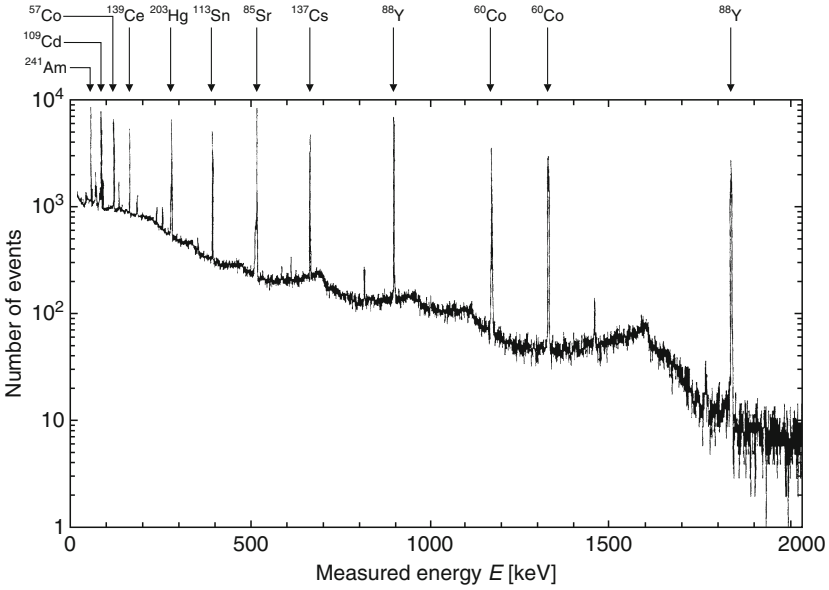


Fig. 4
Photopeak identification in a mixture of radioisotopes based on the pulse-height spectrum recorded with a high-purity germanium detector (Kalthoff 1996)

Radioactive isotopes are best identified by the full-absorption peak ("photopeak") which represents a fingerprint of the decaying nucleus (see *Fig. 4*).

5 Radiation Sources

One might assume that only radioisotopes are significant sources of radiation. The rapid development in basic physics research and its technical applications have created a variety of possibilities for producing nearly all sufficiently long-lived elementary particles and photons in the form of radiation sources. The energy range from ultracold particles ($\ll 25$ meV) up to energies of 1 TeV can be covered. If, in addition, cosmic rays are considered, also particles and photons with energies in excess of 1 TeV are available albeit at low intensity. In the following subsections the main methods of production of ionizing radiation are described.

5.1 Particle Radiation

All charged particles can be accelerated and stored up to very high energies in accelerators. Most accelerators are circular installations in which the particles to be accelerated are kept in a vacuum beam pipe by magnetic dipole fields. The particles are then accelerated by high-frequency alternating electromagnetic fields in so-called cavities which are fed by klystrons. (A klystron is an electron tube that generates microwaves by velocity modulation. Usually a beam of electrons

is passed through a resonant cavity where it interacts with high-frequency radio waves, in the course of which a bunched beam is produced. At accelerators klystrons are used as amplifiers at microwave and radio frequencies to produce high-power driving forces for the particle beams.) For beam focusing quadrupole magnets and magnetic correction lenses are used.

Typically electrons, positrons, protons, and antiprotons are accelerated, e^+ and e^- to, say, 100 GeV in electron synchrotrons and p and $\bar{p}$ to a maximum of 7 TeV (p) in proton synchrotrons. By extracting the accelerated particles a large number of secondary particles can be produced by collisions with an external target, which then can also be stored themselves. In the field of experimental high energy physics mainly pion, kaon, and muon beams are produced for this purpose and also neutrino beams can be made available. Beams of negative pions have also been used in the field of medicine for tumor treatment. Heavy ions, which can also be accelerated in synchrotrons up to high energies, are also an excellent tool for tumor treatment in the framework of hadron and heavy-ion therapy.

Even though charged pions, kaons, and muons are relatively short-lived $\left(\tau_\pi^0 = 26\,\text{ns},\ \tau_K^0 = 12\,\text{ns},\ \tau_\mu^0 = 2.2\,\mu\text{s}\right)$, they can still be used at high energies as secondary radiation because of the relativistic time dilatation. Muons can even be stored in circular accelerators. For energies of about 10 GeV the lifetimes of charged pions, kaons, and muons are given by (γ is the Lorentz factor):

$$\tau_\pi = \gamma\tau_\pi^0 = \frac{E}{m_\pi c^2}\tau_\pi^0 \approx 1.9\,\mu\text{s}\,, \tag{36}$$

$$\tau_K = \gamma\tau_K^0 \approx 240\,\text{ns}\,, \tag{37}$$

$$\tau_\mu = \gamma\tau_\mu^0 \approx 210\,\mu\text{s}\,. \tag{38}$$

Muons of this energy can travel a distance of about 60 km before they decay. A linear accelerator usually serves as injector for a circular accelerator. This type of accelerator is also frequently used in medicine. In linear accelerators particles (mostly electrons and positrons) can be accelerated up to energies of 1 TeV.

At high energies it makes sense – at least for electrons – to use linear accelerators. A circular movement represents an accelerated motion and accelerated charged particles suffer an energy loss by synchrotron radiation. The time-dependent emission of radiation of a charged particle of energy E in a synchrotron with bending radius r is

$$\frac{\mathrm{d}W}{\mathrm{d}t} \propto \left(\frac{E}{m_0 c^2}\right)^4 \frac{1}{r^2}\,, \tag{39}$$

where m_0 is the rest mass of the accelerated particle. Because of the low electron mass compared to that of a proton ($m_e/m_p \approx 1/2{,}000$) the energy loss of electrons in circular accelerators at high energies can be considerable. For example, the energy loss of 100-GeV electrons in the Large Electron–Positron collider LEP at CERN (circumference 27 km, bending radius $r = 3{,}100$ m) per revolution was

$$W \propto \frac{\mathrm{d}W}{\mathrm{d}t}\,2\pi r, \quad W = C\,\frac{E^4}{r}\,, \tag{40}$$

$$W = 8.85 \times 10^{-5}\,\text{GeV}^{-3}\,\text{m}\,\frac{(100\,\text{GeV})^4}{3{,}100\,\text{m}} \tag{41}$$

$$= 2.85\,\text{GeV}\,, \tag{42}$$

so that electron accelerators at high energies can practically only be operated as linear accelerators. Because of the high proton mass the synchrotron-radiation loss for these particles does not currently limit their energies in synchrotrons.

The production of neutrino beams is also worth mentioning. Neutrino beams play a dominant role in particle physics when the validity of the standard model of particle physics is to be tested. Since the interaction cross section of neutrinos (ν) is exceedingly small, neutrino beams of high intensity have to be generated so that statistically significant interaction rates are achieved. Normally muon neutrinos are used for this purpose, where these neutrinos are produced in pion or muon decays ($\pi^+ \rightarrow \mu^+ + \nu_\mu$, $\pi^- \rightarrow \mu^- + \overline{\nu}_\mu$, $\mu^+ \rightarrow e^+ + \nu_e + \overline{\nu}_\mu$, $\mu^- \rightarrow e^- + \overline{\nu}_e + \nu_\mu$). In planned so-called neutrino factories the neutrino fluxes will be very high: high enough for aspects of radiation protection to play an important role.

5.2 Photon Sources

X-ray tubes represent a classical photon source. This energetic electromagnetic radiation was discovered by Wilhelm Conrad Röntgen in 1895. In typical X-ray tubes photons with energies up to several hundred keV can be produced, even the MeV range is accessible. The X-ray spectrum is continuous since it is created by electron bremsstrahlung. It is, however, superimposed by discrete X-ray lines, which are characteristic for the used anode material. The energies of characteristic X rays can be determined using Moseley's law:

$$E(\mathrm{K}_\alpha) = Ry\,(Z-1)^2 \left(\frac{1}{n^2} - \frac{1}{m^2}\right). \tag{43}$$

In this equation n and m are the principal quantum numbers and Ry is the Rydberg constant (13.6 eV). For example, for K_α radiation on lead ($n = 1$, $m = 2$) one obtains $E(\mathrm{K}_\alpha^{\mathrm{Pb}}) = 66.9\,\mathrm{keV}$.

The photon energy range which is classically covered by X-ray tubes is also obtained by synchrotron-radiation sources. However, the photon flux of synchrotron-radiation sources is higher by many orders of magnitude. The synchrotron-photon spectrum is continuous just like the bremsstrahlung spectrum of X-ray tubes. Because of the high intensity of the synchrotron-photon spectrum intense monoenergetic photon beams can be produced by monochromators. (A device that produces monochromatic radiation from a polychromatic source. For X rays normally diffraction by crystals, which act like an optical grating, is used.) Such monochromatic X rays from electron synchrotrons are, for example, used in diagnostics in the field of noninvasive coronary angiography. The coronary arteries are marked with stable iodine as contrast agent. Two digital images, one below and one above the K absorption edge, are taken, and subtracted in a computer. This allows to suppress the absorption in the surrounding tissue thereby yielding an image of the blood vessels of high contrast ("dual-energy technique" or "K-edge subtraction technique").

Decays of neutral pions ($\pi^0 \rightarrow \gamma + \gamma$) or annihilations of electrons and positrons ($e^+ + e^- \rightarrow \gamma + \gamma$, used in positron emission tomography (PET)) are sources of energetic photons. For completeness also decays of radioisotopes have to be mentioned, which – after α or β activity – decay frequently by γ emission into the ground state. In these decays photons with energies up to several MeV can be produced. As a consequence of nuclear transformation also excitations in the electron shell can occur leading to characteristic X rays.

In electron–photon interactions energetic electrons can produce high-energy photons by the inverse Compton effect. This process plays a dominant role in X-ray and γ-ray astronomy.

5.3 Neutron Sources

Neutrons are predominantly produced in strong interactions. In spallation neutron sources, energetic hadrons (mostly protons) produce a large number of neutrons in reactions with heavy nuclei. It is possible to create up to 30 neutrons per reaction by the bombardment of nuclei with hadrons. The spallation neutrons are created over a wide energy range and could ideally be used for the purpose of transmutation of nuclear waste. In such transmutations long-lived fission products from nuclear-fission reactors can be transformed by neutron interactions into short-lived or even stable isotopes. This technique, if being used on a large scale, could represent an attractive alternative to the disposal of radioactive waste in underground cavities or other deposits. In dedicated neutron generators single neutrons can be produced by the bombardment of special targets with protons, deuterons, or alpha particles. In this way neutrons in the MeV range are obtained.

A classical technique is the neutron production in (α, n) reactions. α-ray-emitting radioisotopes are mixed with a beryllium isotope. The α rays interact with ^{9}Be to produce neutrons of around 5 MeV according to the reaction

$$\alpha + {}^{9}\mathrm{Be} \longrightarrow {}^{12}\mathrm{C} + n\,. \tag{44}$$

As α emitter one can use radium (^{226}Ra), americium (^{241}Am), plutonium (^{239}Pu), polonium (^{210}Po), or curium (^{242}Cm, ^{244}Cm). The neutron yield for these (α, n) reactions is of the order 10^{-4} per α particle.

In fission reactors highly radioactive fission products are generated. Since the fission materials, for example, ^{235}U, are relatively neutron rich, the fission products contain too many neutrons. The neutron excess can be decreased by the emission of prompt or delayed neutrons. Nuclear-fission reactors are therefore a rich source of neutrons. The fission products can also try to reach a stable final state by successive β^- decays.

In the Sun, which is a fusion reactor, globally four protons are fused into helium. In a fusion power plant deuterium and tritium will be fused to helium:

$$d + t \longrightarrow {}^{4}\mathrm{He} + n\,. \tag{45}$$

In this reaction an energetic neutron is generated (E_n = 14.1 MeV). The original hope that fusion power plants would be completely clean from the point of view of radiation protection is not fully tenable. These energetic neutrons are very difficult to shield. They will activate the reactor materials and produce radioisotopes, so that a fusion reactor also represents a potential danger from the point of view of radiation protection. On the other hand, it is desirable that energetic neutrons are produced, despite their disadvantageous effects on the materials, as they are the source of power of fusion reactors. It has to be mentioned, however, that the big advantage of fusion reactors is that uncontrolled chain reactions cannot occur, which is definitely a possible problem with nuclear-fission reactors.

Finally, it should be mentioned that under extreme conditions – as can be found, for example, in supernova explosions – the pressure in the hydrogen plasma can be so high that electrons

are merged with protons. In such a deleptonization process

$$p + e^- \longrightarrow n + \nu_e \tag{46}$$

neutrons are produced. Because of the average lifetime of neutrons of 886 s, these neutrons – even at very high energies – will decay long before reaching Earth, if traveling in that direction. In most cases, however, they do not escape from the supernova but are rather caught in a neutron star after a gravitational collapse.

5.4 Cosmic-Ray Sources

Cosmic rays may present a radiation hazard, in particular, for flight personnel. Radiation-relevant aspects concerning cosmic rays are presented in some detail in Sect. 9. Here only the potential of cosmic rays as sources of particles for measurement and calibration of detectors shall be mentioned. Primary cosmic rays consist mainly of protons, α particles, and some heavy nuclei ($Z \geq 3$). By interactions of primary cosmic rays with atomic nuclei of the atmosphere the initial primary particles initiate cascades of secondary and tertiary particles. In this way predominantly pions and kaons are generated. These mesons can either induce further interactions or decay. (Pions and kaons are mesons which belong to the group of strongly interacting particles ("hadrons"). These hadrons include baryons and mesons. Protons, neutrons, and their excitations are baryons. In the naïve quark model baryons are composed of three quarks. For example, the proton (uud) is a bound state consisting of two up quarks (electric charge +2/3 of an elementary charge) and a down quark (electric charge −1/3 of an elementary charge), while the neutron is a udd state. In contrast, mesons are bound states of a quark and an antiquark. For example, the positively charged pion (π^+) is a $u\bar{d}$ state.) The competition between interaction and decay depends on the local density of the atmosphere. The soft component of cosmic rays consisting of electrons, positrons, and photons (initiated by π^0 decay) will be absorbed relatively fast in the atmosphere. In contrast, the decay products of charged pions and kaons, namely, muons and neutrinos, can easily reach sea level. Eighty percent of charged cosmic rays at sea level are composed of muons, which are mainly distributed over an energy range from about 1 GeV up to 1 TeV. Because of the low interaction probability cosmic neutrinos play almost no role for the purposes of radiation protection.

The omnidirectional muon flux at sea level through a horizontal area amounts to about 1 particle per square centimeter and minute, corresponding to a flux for near-vertical directions per solid angle and area of

$$\phi(\mu^{\pm}) = 8 \times 10^{-3}\,\mathrm{cm^{-2}\,s^{-1}\,sr^{-1}}. \tag{47}$$

Because of their low energy loss ($\approx 2\,\mathrm{MeV/(g/cm^2)}$) muons can also reach large depths underground.

In the field of radiation protection, cosmic-ray muons will lead to a background in all radiation monitors, in addition to that from terrestrial radiation. A contamination monitor with a horizontal area of $15 \times 10\,\mathrm{cm^2}$ will measure a background rate of ≈150 particles per minute purely due to cosmic rays. This result should not only be considered as a disadvantage since it

represents at the same time a function test of the measurement devices. For such a test no artificial radiation sources are required. For measurements on radioactive materials this background rate has to be considered in any case.

6 Radiation Detectors

The necessity for the measurement of radiation exposures originates from the fact that this type of radiation has to be surveyed, controlled, and limited. Humans also have to be protected against unexpected exposures. On the one hand, the surveillance of radiation-exposed workers, the measurement of external radiation exposures, contaminations, and incorporations, in particular, in working areas, are very important. On the other hand, the environment has to be protected against unnecessary exposures. The latter point of view includes the determination of radiation exposures of the population, the monitoring of the disposal of radioactive waste into the environment, and the examination of the distribution of radioactive material in the biosphere (atmosphere, soil, water, food). In addition, national radiation-protection authorities also have realized that radiation exposures from natural sources have to be considered. In certain situations, natural radiation can increase the radiation level for individuals of the population quite considerably. Therefore, these natural sources cannot be neglected in the framework of radiation protection.

The radiation detectors used in the field of radiation protection have to be reliable and robust and their measurements have to be reproducible. It is important to note that for different types of radiation adequate detectors must be used. The detailed working principles of standard radiation detectors are described in various articles in this handbook in great detail. Here we only present in the following some general features mostly relevant to the field of radiation protection.

6.1 Ionization Chambers

A very simple radiation detector is the ionization chamber. Radiation incident into an ionization chamber will produce electrons and ions by ionization of the counter-gas filling. These charge carriers are collected in a constant homogeneous electrical field. Since there is no gas gain in this type of chamber, the signals are very small. Therefore, these signals have to be amplified electronically. Ionization chambers are excellently suited for the measurement of α rays which deposit their total energy in the chamber volume. Since the chamber signals are proportional to the energy, these detectors allow α-ray-spectroscopy measurements. Ionization chambers also permit an accurate measurement of the ion dose and the ion-dose rate via a measurement of the chamber current.

6.2 Proportional Counters and Geiger–Müller Counters

The detection technique in sealed proportional counters and Geiger–Müller counters is the same as in ionization chambers. But in contrast to ionization chambers, one measures the current and voltage signals and counts them instead of measuring the chamber current. The

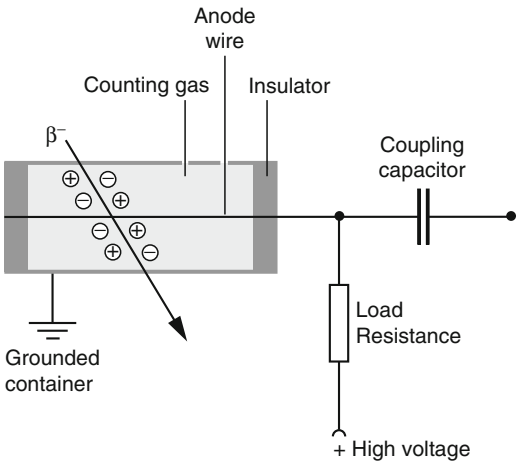


Fig. 5
Sketch of a cylindrical gas counter

mechanical and electrical setup of a cylindrical gas counter is shown in *Fig. 5*. Depending on the high voltage, one distinguishes between proportional counters and Geiger–Müller counters. (The high voltage depends on the counter geometry, the anode-wire diameter, and the gas filling. Typical values for the proportional regime are around 500 V for anode wires with 30 μm diameter for an argon/methane gas filling of 80:20. For otherwise identical conditions Geiger–Müller tubes require somewhat higher voltages, like about 1,500 V.) The electrical field in the cylindrical counter is inhomogeneous – its field strength varies inversely proportional with the distance to the anode wire. The charge carriers produced by ionization will drift depending on the sign of their charge to the anode or cathode. The electrons drift to the anode wire. When they approach the anode they will encounter stronger and stronger field strengths. If they gain on their mean free path, i.e. between two collisions with gas molecules, an energy from the electrical field which is larger than the ionization energy of the gas, then gas amplification starts. This charge-carrier amplification sets off an avalanche. In the proportional range gas-gain factors of about 10^4 are achieved. The discharge in the proportional counter is localized to the position where the particle has passed through the chamber. Since the output signal in the proportional counter is proportional to the energy loss (or the energy, if the particle is completely absorbed in the detector volume), strongly ionizing α particles and weakly ionizing β rays can be distinguished in this operation mode.

For higher anode voltages a transition from the proportional range to the Geiger–Müller regime is observed. The discharge in the counter volume will spread laterally along the whole anode wire and gas gains of 10^8 up to 10^{10} are reached. The signals in the Geiger–Müller mode are independent of the type and energy or energy loss of the incident particle. Therefore, it is impossible to distinguish different particle types with this detector. The advantage is that the signal amplitudes are rather large so that no sensitive amplifiers are needed for the signal analysis.

The detection efficiency for charged particles in both counter types is close to 100%. Photons, however, will only be measured with an efficiency on the order of a few percent. This is because

photons first have to create charged particles in the chamber volume, and the interaction probability is relatively small because of the low target density of the gas filling.

Proportional counters and Geiger–Müller counters allow the measurement of low doses and activities. They are also an excellent detector for dose-rate measurements. A problem with a sealed proportional counter or Geiger–Müller counter is that low-energy α and β radiation might easily be absorbed in the chamber walls. For low-energy α and β rays, therefore, open gas flow detectors are frequently used. Radioactive samples (e.g., ^{14}C for radioactive carbon dating) are installed inside the counter so that an absorption in the chamber walls is avoided. For large-area counters for the measurement of activities and contaminations the gas volume is terminated by an extremely thin plastic foil. This entrance window normally is also transparent for α and β rays. Large-area counters are frequently used as personal radiation and contamination monitors. They mostly contain several parallel anode wires and they are operated in the proportional regime ("multi-wire proportional chamber").

To find out the optimal working point of a gas counter, the count rate as a function of the high voltage is measured. The obtained dependence exhibits a plateau, where the counting rate (which is given by the radioactive source used) is independent of the applied high voltage or where there is only very little variation with the voltage. The length of the count-rate plateau and its slope is a measure of the quality of the counter. The working point is best chosen right at the center of the count-rate plateau.

Activity measurements in the high-rate domain with Geiger–Müller counters must be corrected for dead-time losses. The ionizing power of particles and the following gas amplification produces a large number of charge carriers which will decrease the external field strength. Therefore, the counter is for a certain time insensitive for further particle registration. This time is called dead time. Only when the charge carriers produced by the ionization of the incident particle are removed from the sensitive volume can further particles be recorded. The counter is said to be dead after each particle passage for a certain time τ. The amount of the time where the counter is insensitive for particle measurements per unit time is $N\tau$ if N is the count rate. This means that the counter is only sensitive for the fraction $1 - N\,\tau$. The true count rate N_{true} is obtained after correction for dead-time effects to be

$$N_{\text{true}} = \frac{N}{1 - N\,\tau}\,. \tag{48}$$

Rate measurements with Geiger–Müller counters on board American space probes (Explorer I) in the late 1950s showed suddenly extremely low count rates when they flew through the Van Allen radiation belts, so that the physicists were afraid that the detectors stopped to work properly. The apparent low count rates were, however, related to the extremely high particle fluxes in the Van Allen belts which produced extremely large effective dead times.

For scintillation counters the dead times are much smaller (typically $\tau \approx$ nanoseconds). Therefore, scintillation counters are highly superior over Geiger–Müller counters for high-rate measurements.

6.3 Scintillation Counters

Scintillation counters can not only be used for the measurement of charged particles but also for the detection of γ rays (*Fig. 6*). The photons are absorbed by the scintillation-counter medium undergoing photoelectric effect, Compton scattering, or pair production. The created

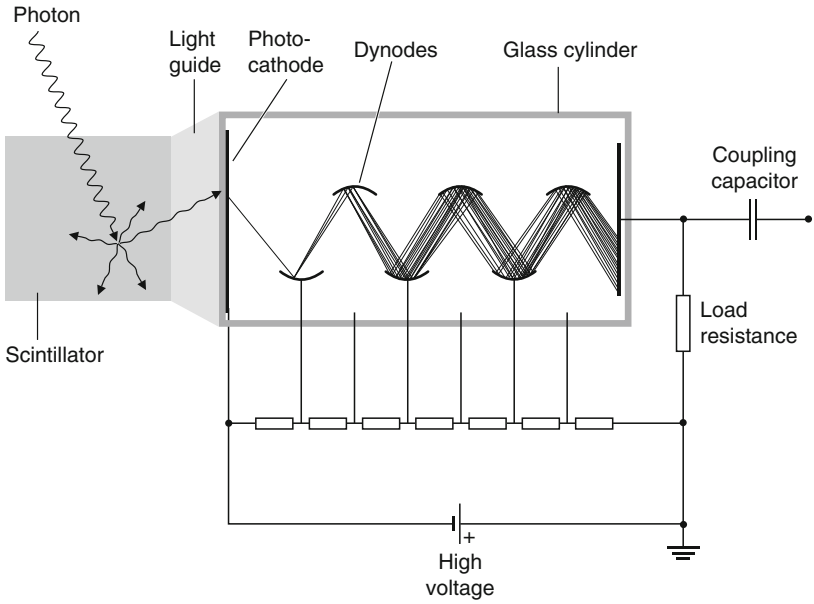


Fig. 6
Sketch of a scintillation counter with photomultiplier readout

electrons or positrons excite in the scintillator medium electrons in atomic shells into higher states. Upon de-excitation the scintillator gives off light. The light yield is proportional to the deposited energy in the scintillator. To produce a photon in the visible spectral range an energy loss of about 100 eV is required. The scintillation photons will be propagated via a light guide to the photocathode of a photomultiplier. This photomultiplier then converts the incoming scintillation photons via the photoelectric effect into photoelectrons, which will be amplified in the photomultiplier by secondary emission on dynodes. In this way signals of about 50 mV at the output of the photomultiplier are obtained. The construction of a scintillation counter with photomultiplier readout is shown in *Fig. 6*.

Suitable scintillation materials are inorganic doped crystals (NaI(Tl), CsI(Tl), LiI(Eu), etc.), organic liquids (*p*-terphenyl, anthracene), or polymerized solids. The organic scintillation counters consist mostly of three components, one of which is the primary scintillator (e.g., naphthalene), a wavelength shifter (e.g., butyl PBD [PBD = 2-(4-tert.-butylphenyl)-5-(4-biphenyl-1,3,4-oxadiazole)]), and a solving agent (e.g., Uvasol [trade name of Merck, Germany] for liquid scintillators or PMMA [polymethylmethacrylate (also known under the name Plexiglas® or Lucite)] for polymerized solids). The primary scintillation material is not transparent for its own scintillation light. Therefore, the wavelength shifter has to shift the primary light into a lower frequency range for which the scintillator is transparent. At the same time, the frequency of the light can be adjusted to the spectral sensitivity of the photomultiplier. Liquid and plastic scintillators have the large advantage that their shape can be adjusted to the required measurement geometry.

A simple procedure for testing the functionality of proportional or scintillation counters is to measure the so-called background rate. The background rate originates from the radiation

of the natural environment. As with all statistical phenomena the result of such a measurement is subject to certain fluctuations which are not related to the inaccuracy of the measurement, but have their origin in the stochastic character of the radioactive decay. For low background the rates per unit time can be described by an asymmetric Poisson distribution (negative values cannot occur):

$$f(N,\mu) = \frac{\mu^N \mathrm{e}^{-\mu}}{N!}, \qquad N = 0,1,2,3,\ldots. \tag{49}$$

In this distribution μ is the average value $\mu = \frac{1}{k}\sum_{i=1}^{k} N_i$ of k measurements. $N!$ is a shorthand for the product of $1 \times 2 \times 3 \times 4 \times \cdots \times N$, called N factorial.

For higher count rates the Poisson distribution can be approximated by a Gaussian distribution. This Gaussian is symmetrical around its average value. The function

$$f(N,\mu) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(N-\mu)^2}{2\sigma^2}\right) \tag{50}$$

describes the form of a Gaussian. The width of the Gaussian can be described by its standard deviation σ. Within $\mu \pm \sigma$ one finds 68.27% of all measurement values, and within $\mu \pm 2\sigma$ there are 95.45% of the measured values. From the fit to the data the full width at half maximum (FWHM) can easily be read. This FWHM is related to the standard deviation by

$$\Delta N_{\mathrm{FWHM}} = 2 \times \sqrt{2 \times \ln 2}\,\sigma = 2.355\,\sigma\,. \tag{51}$$

The statistical error for a single measurement is given by the square root of the count rate. These considerations apply for all count-rate measurements.

For the measurements of the activity of radioactive sources the background effect always has to be considered. This background rate can vary for different places because the environmental activity also depends on where it is measured.

Because of the excellent detection efficiency for photons, scintillation counters (mainly inorganic scintillation counters) are used for the detection of γ rays and for γ-ray spectroscopy. They can also be used for high-precision dose-rate measurements. Even though scintillation counters have a high intrinsic time resolution, saturation effects at very high count rates will occur.

6.4 Semiconductor Counters

Semiconductor counters have a superior energy resolution over scintillation counters. In semiconductor counters, which are mostly based on silicon or germanium, only about 3 eV are required to produce an electron–hole pair. Hence, more electron–hole pairs are produced for a fixed energy deposit compared to the number of photons generated in scintillation counters. The electron–hole pairs created by α, β, or γ rays are collected in an external electrical field before they can recombine ("solid-state ionization chamber").

Because of their high energy resolution, semiconductor counters are ideally suited for γ spectroscopy and the identification of radioisotopes via their characteristic γ-ray lines. This is, for example, demonstrated in ❯ *Fig. 4*.

In addition to silicon and germanium semiconductor detectors also other materials like gallium arsenide (GaAs), cadmium telluride (CdTe), and cadmium–zinc telluride (CdZnTe) are frequently used in the field of radiation protection and nuclear physics.

6.5 Neutron Dosimeters

For the purpose of neutron dosimetry these neutral particles first have to create charged particles in nuclear interactions. For the measurement of neutrons in a Geiger–Müller counter or proportional counter one can use as a counting medium BF_3 to eventually measure the α particles from the reaction $n + {}^{10}B \rightarrow \alpha + {}^{7}Li$. In a LiI(Eu) scintillator one takes advantage of the interaction $n + {}^{6}Li \rightarrow \alpha + {}^{3}H$ for the production of α particles and tritons.

Apart from boron-trifluoride counters and albedo dosimeters also nuclear-track detectors are used to a large extent for fast, energetic neutrons. These track-etch detectors can measure neutrons in the energy range from 2 to 70 MeV. Typical applications are works at accelerators, the handling of radium–beryllium sources in the laboratory, or the measurement of the high-energy neutron component in cosmic rays at large flight altitudes. This last measurement is of particular importance for the flying personnel. Track-etch dosimeters made from dedicated materials are largely insensitive against α, β, and γ rays. Neutrons produce a certain local radiation damage in the material, which can be made visible by etching, for example, with a solvent like NaOH. The evaluation of nuclear-track detectors is quite cumbersome because the etch cones produced by neutrons are relatively small and frequently demand scanning the etched material with a microscope.

6.6 Personal Dosimeters

In the field of personal dosimetry one has to distinguish between directly and indirectly readable dosimeters. Ionization chambers can be constructed as pen-type pocket dosimeters which allow a direct reading of the received dose on a scale calibrated in milli- or microsievert. The sensitive volume of a pen-type pocket dosimeter consists essentially of a capacitor which is charged to a certain voltage. The irradiation of the chamber gas leads to a dose-proportional current which discharges the capacitor slowly. Just as with an electrometer, the charge state of the capacitor is read with the help of a quartz dial which is imaged via a built-in optic to a scale calibrated in milli- or microsievert. If one uses pen-type pocket dosimeters for X rays, one has to consider that low-energy X rays can easily be absorbed in the walls of the dosimeter. Furthermore, the sensitivity of a pen-type pocket dosimeter varies considerably for energies below 300 keV due to the strong dependence of the photoelectric cross section on the photon energy (see *Fig. 7*).

The most popular and well-known indirectly readable dosimeter is the film-badge dosimeter. These dosimeters use the blackening of a photographic film (X-ray film) as a measure for the received dose. By metal absorbers (Cu, Pb) of different thickness information about the intensity, direction of incidence, radiation type, and even the energy can be achieved. A hole in the cassette also allows the measurement of low-energy β rays. Because of the high degree of information, the mechanical stability, and the possibility of permanent documentation, official dosimeters are mostly film badges. There are, however, several disadvantages of film badges, namely, the limited durability of films, the sensitivity to humidity and high temperature, the limited measurement accuracy, and the laborious way of reading the information.

Film dosimeters have their limitations in determining the skin dose parameters $H_p(10)$ and $H_p(0.07)$. New types of dosimeters optimized for the measurement of the penetration of radiation into the skin, like the "sliding-shadow" dosimeters have become available. These dosimeters also allow a determination of the photon energy and its angle of incidence. Furthermore they can discriminate between β and γ rays.

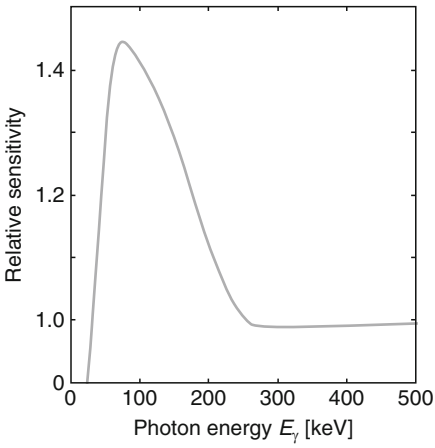


Fig. 7
Relative sensitivity of a pen-type pocket dosimeter in its dependence on the photon energy

Phosphate-glass dosimeters can be used in the field of radiation protection because they will give off a characteristic fluorescence radiation under exposure to ultraviolet light, where this fluorescence radiation is proportional to the intensity of the received energy dose. This type of dosimeter is frequently constructed in a spherical form which partially compensates for the distinct energy dependence of the dose reading ("spherical dosimeter"). By using γ-ray filters (from tin) or neutron filters (from plastic with boron content) the identification of the radiation type also appears possible.

Phosphate-glass dosimeters have a large measurement range, a very low long-term fading, and they can be analyzed repeatedly so that the information they contain is available for multiple documentation. One of the disadvantages of these types of dosimeters is related to the fact that the analysis of these dosimeters requires a rather expensive readout system.

Thermoluminescence dosimeters (TLDs) take advantage of the property of some inorganic compounds (e.g., LiF) to give off light when they are heated up after excitation with ionizing radiation. The light intensity is proportional to the received energy dose. A distinct advantage of thermoluminescence dosimeters is that even for very small dimensions they exhibit a high sensitivity and they can be used just as phosphate-glass dosimeters as finger-ring dosimeters. A disadvantage is certainly the tedious analysis and the fact that the heating-up of the thermoluminescence dosimeter erases the dose information. However, this can also be considered as an advantage because after heating up the thermoluminescence material this type of dosimeter can be used anew.

An interesting candidate giving useful applications for the measurement of low-level radiation are thermoluminescence badges containing the rare-earth metal thulium in the form of thulium acetate or thulium acetylacetonate. This material has a very high radiation sensitivity with the added advantage that low-energy X rays and γ rays can be separately detected. Thulium emits light in the ultraviolet (375 nm) and in the visible range (465 nm).

Albedo neutron dosimeters measure low-energy neutrons which are backscattered from the body of the person who carries such a dosimeter. They mostly consist of thermoluminescence

sheets in which neutrons can be detected via reactions with boron or lithium. The calibration depends, however, on the person that carries this type of dosimeter.

For radon monitoring, plastic detectors can be used (cellulose-nitrate sheets). The α rays which occur in the decay chain of the radon isotope will produce a local radiation damage in the plastic material which can be etched with soda solution and hence made visible.

Occasionally, a problem occurs with the determination of body doses after radiation accidents if no dosimeter information was available (accident dosimetry). A possibility to work out the received body dose after the fact is given by the so-called hair-activation method. Hair contains sulphur with a concentration of 48 mg sulphur per gram hair. By neutron interactions (e.g., after a reactor accident) the sulphur can be activated according to

$$n + {}^{32}\mathrm{S} \longrightarrow {}^{32}\mathrm{P} + p\,. \tag{52}$$

In this way the radioisotope phosphorus 32 is produced, which has a half-life of 14.3 days. In addition to this, radioactive silicon 31 is produced by

$$n + {}^{34}\mathrm{S} \longrightarrow {}^{31}\mathrm{Si} + \alpha\,. \tag{53}$$

The ^{31}Si radioisotope interferes with the phosphorus-activity measurement. The half-life of the ^{31}Si isotope, however, amounts to only 2.6 h. Therefore, one waits until this activity has decayed before the ^{32}P activity is measured. In case of a surface contamination of the hair substantial cleaning is also necessary before the phosphorus activity of the hair is measured.

Phosphorus 32 is a pure β-ray emitter. The maximum energy of the electrons is 1.71 MeV. Because of the low expected count rates, a detector with high efficiency and low background is required. For example, an actively and passively shielded end-window counter can be used. The received radiation dose can be inferred from the phosphorus-32 activity with the help of the known activation cross section.

A further possibility of accident dosimetry is given by the procedure of blood activation. Human blood contains about 2 mg sodium per milliliter. This concentration is about the same for all humans. By the capture of thermal neutrons the stable ^{23}Na isotope transforms into radioactive ^{24}Na,

$$n + {}^{23}\mathrm{Na} \longrightarrow {}^{24}\mathrm{Na} + \gamma\,. \tag{54}$$

Again activities of short-lived radioisotopes produced in this neutron-capture process can interfere with the ^{24}Na-activity measurement. After a suitable decay time the remaining activity of ^{24}Na ($T_{1/2}$ = 15 h, $E_{\beta^-_{\max}}$ = 1.39 MeV with subsequent γ emissions) can be recorded and used as basis for the determination of the received dose.

The presented radiation detectors help to identify radioisotopes, to determine the amount of radioactivity release, to monitor possible illegal clearance of radioactive material, and to observe the effect of radiation on the environment and on humans. In typical surveillance programs for nuclear power plants it is required that the different ways of exposure due to external and internal radiation, the release rates into water and air, the concerned food chains, and the identity of the released radioisotopes have to be determined. The corresponding measurement techniques have to be very sensitive since the permitted dose-rate limits for the release of radioactive material into the environment for the normal population are usually quite low.

Dose-rate measurements for personal dosimetry predominantly employ ionization chambers and Geiger–Müller counters. Because of the energy and directional dependence of these

Table 4

Typical applications of different measurement techniques for personal dosimetry

Dosimeter	Principle of operation	Radiation type measurement range	Advantages and disadvantages
Film badge	Photo-chemical blackening	γ, β 0.1 mSv–5 Sv	Can be documented, insensitive for low-energy γ rays
Pen-type pocket dosimeter	Ionization chamber	γ 0.03–2 mSv, also other ranges of measurement	Very sensitive, permanently readable, insensitive for α and β rays, cannot be documented
Directly readable dosimeter ("pocket dosimeter")	Ionization or proportional chambers and GM counters	γ 0.1 μSv–10 Sv	Permanently readable, cannot be documented
TLD dosimeter	Thermo-luminescence measurement	γ, (β) 0.1 mSv–10 Sv	Suitable for the measurement of low doses, cannot be documented
Phosphate-glass dosimeter	Photo-luminescence measurement	γ 0.1 mSv–10 Sv	Can be documented, can be read repeatedly
Albedo neutron dosimeters	Neutron moderation by the carrier	n, γ 0.1 mSv–10 Sv	Calibration depends on the human carrier
Track-etch dosimeter	Material damage in polycarbonate films	n 0.5 mSv–10.0 mSv	Can be documented, the evaluation requires the knowledge of the radiation field
Radon personal dosimeter	Material damage in cellulose nitrate films	α 75–7000 kBq h/m^3 [a]	Can be documented

[a] For a trimonthly surveillance (corresponding to 500 working hours) this corresponds to an average radon concentration at the workplace of 150–14,000 Bq/m^3

detectors, the frequently unknown mixtures of types of radiation, and the impossibility to record low-energy β-ray emitters, the measurement errors are quite substantial (20–50%). If the activity and dose of a mixture of different γ-ray sources has to be determined, the uncertainty of the measurement can amount to 100% and even more.

Typical domains, where different measurement techniques for personal dosimetry are used, are compiled in ❯ *Table 4*.

It can be noticed that there is only one personal detector for α rays in Table 4, namely, the radon personal dosimeter. Due to the short range of α particles in air or clothing, the external irradiation by α rays mostly represents no radiation risk. An exception consists of radon inhalation in a radon environment.

7 Safety Standards

The ICRP has proposed safety standards to protect the health of workers and the general public against the dangers arising from ionizing radiation. The recommendations are laid down in a European Directive (Council Directive 96/29/EURATOM) which was presented to the Member States of the European Community.

The Directive has defined safety standards for exposed workers in the following way:

- The limit on the effective dose is 100 mSv in a consecutive 5-year period, subject to a maximum effective dose of 50 mSv in any single year. In accordance with this, most Member States have defined an annual limit of 20 mSv.
- The annual limit on the equivalent dose for the lens of the eye is 150 mSv.
- The annual limit on the equivalent dose for the skin is 500 mSv.
- The annual limit on the equivalent dose for the hands, forearms, feet, and ankles is 500 mSv.

The annual limit for the whole-body dose for the general population, for example, from nuclear power plants, is 1 mSv in most countries.

The general recommendation is that reasonable steps must be taken to ensure that the exposure of the population as a whole is kept as low as reasonably achievable (ALARA principle).

In contrast, other countries, for example, the USA, have regulations which differ distinctly from the European Directive. For example, the annual whole-body dose limit for workers exposed to ionizing radiation in the USA is 50 mSv (many laboratories in the USA set lower limits) compared to 20 mSv in European countries. Other differences are that in the USA the old radiation units (rad and rem) are still in use (1 Sv = 100 rem, 1 Gy = 100 rad).

8 Organization of Radiation Protection

The responsibility for the correct integration of the radiation-protection rules in a company, nuclear power plant, research center, or an university lies in the hands of the radiation-protection supervisor. The radiation-protection supervisor has to appoint in a radiation-protection directive an appropriate number of radiation-protection officers for the control and surveillance of the work in question. The radiation-protection officer or, for short, the radiation officer has to be qualified for his/her work in the field of radiation protection. In contrast to this the radiation-protection supervisor need not be an expert in the field of radiation protection. He/she transfers the duty to respect the regulations of radiation protection to the radiation officer.

The regulations in the field of radiation protection require that persons who will handle radioactive material, for example, students in a nuclear physics lab at a university, are instructed about the possible dangers of handling radioactive sources. This instruction must be done annually and has to cover a number of aspects which are:

- Introduction into the local laboratory safety rules.
- Introduction into the standard working procedures with radioactive sources.
- Information with respect to possible dangers.
- Information about radiation exposures.
- Description of the safety rules and possible protection techniques.
- Accurate information about the relevant radiation-protection safety rules.
- Instructions about the organization of radiation protection, and the responsibility of the radiation officer. The workers perform their tasks under the guidance of the radiation officer and they are bound to follow his instructions.

It is the aim of safety measures in the field of radiation protection to avoid unnecessary radiation exposures, contaminations, and ingestion and inhalation of radioactive material ("incorporations"). To a certain extent, of course, there are radiation exposures which are unavoidable, but it is the aim to reduce these unavoidable radiation exposures, contaminations, or incorporations to a level as low as reasonably achievable. This is the so-called ALARA principle. There are, however, national radiation-protection regulations which require the radiation exposure to be kept as low as possible. Of course, it must be ensured that the exposures stay within the limits given by the regulations. To fulfill these requirements is the main task of the radiation-protection officer. (In the early days of the study of radioactivity the concept of "radiation protection" did not even exist and there was no "radiation officer." Physicists like Becquerel, Curie, and Hahn handled relatively large quantities of radioactive substances with their bare hands. In addition to this, any radioactive material given off as gas or airborne dust was frequently inhaled. Even today the logbooks of Marie and Pierre Curie are contaminated by radium (half-life 1600 years) and their decay products, and they are on loan in the Bibliothèque Nationale only with special restrictions.)

In the following some practical aspects of radiation protection are listed, namely,

- Licensing
- Installations
- External facilities
- Design approval
- Fire fighting
- Arrangements for mitigating the consequences of a severe accident
- Training of radiation workers
- Protection of air, water, and soil
- Radiation exposure from specific causes or unforeseen circumstances
- Handling of unsealed radioactive sources
- Contamination and decontamination
- Medical supervision
- Storage and security of radioactive substances
- Bookkeeping and declaration
- Treatment of radioactive waste

One might assume that only radioisotopes are significant sources of radiation. However, as mentioned in Sect. 5, there is quite a variety of possibilities to produce all kinds of particles over a wide energy range.

Regulations for X-ray sources and X-ray facilities are analoguous to the normal radiation-protection standards.

Many radiation accidents in the fields of medicine and technology are caused by losses and careless disposal of radioactive material. The reason for unnecessary exposures is frequently due to improper storage of disused radioactive sources.

Radiation accidents in large manufacturing plants and nuclear-medical sections of hospitals are frequently caused by nonexisting elementary safety rules. In the case of existing safety rules they are often ignored. It is also essential that the maintenance personnel are suitably trained and aware of the radiation risks. Radiation-protection regulations must be meticulously respected, otherwise accidental irradiations or even accidents may occur.

The basic rules in the field of radiation protection are:

- Limit the activity of radioactive material
- Use shielding whenever possible
- Limit the exposure time
- Keep suitable distance to the radioactive material if possible
- Avoid contamination
- Avoid incorporations

As far as shielding is concerned one has to keep in mind that γ rays are effectively absorbed by heavy materials, like lead. In contrast, electrons are best stopped with a sandwich starting with low-Z material followed by lead. In this way one avoids bremsstrahlung by fast electrons which is difficult to shield. When the electrons are slowed down by the low-Z material they are stopped by a layer of lead. Neutrons are best absorbed by materials that contain many protons because neutrons can transfer their energy effectively to partners of the same mass.

An incorrect shielding might even lead to an increased dose, because behind an inadequately designed shield the particle rate can increase due to interactions in the shielding. This built-up effect has to be considered.

9 Environmental Radiation

Natural radioactivity from the environment has three components:

- Cosmic rays ($\approx$ 0.3 mSv/year),
- Terrestrial radiation ($\approx$ 0.5 mSv/year),
- Ingestion (eating, drinking, and breathing) ($\approx$ 1.5 mSv/year).

Cosmic rays from our Sun and our galaxy and terrestrial radiation from the Earth's crust as well as incorporations of radioisotopes from the biosphere represent whole-body exposures. A special role is played by the inhalation of the radioactive noble gas radon which, in particular, represents an exposure for the lungs and the bronchi. In addition to these natural sources further exposures due to technical, scientific, and medical installations developed by modern society

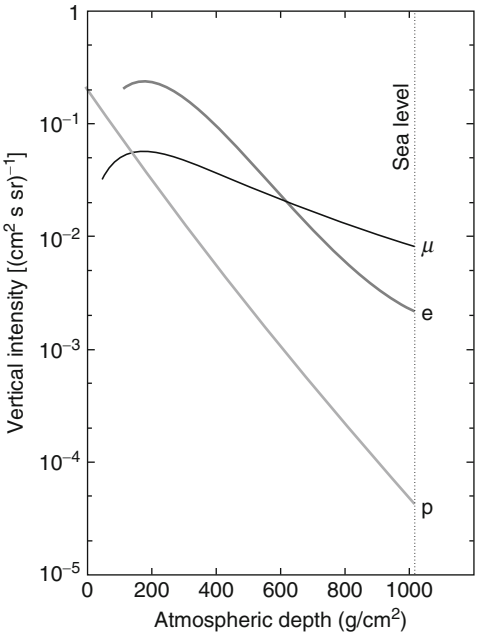


Fig. 8
Altitude dependence of proton, electron, and muon fluxes in the atmosphere

occur. The existence of natural radioactive substances, however, demonstrates that radioactivity and the development of life coexisted since the very earliest times on our planet.

Our Milky Way is the dominant source of high-energy cosmic rays. The low-energy particles predominantly originate from our Sun. Primary cosmic rays consist largely of protons (≈ 85%) and helium nuclei (≈ 12%). Only 3% of primary nuclei are heavier than helium.

The altitude dependence of protons and secondary electrons and muons is shown in *Fig. 8*. Secondary particles are created in interactions of primaries in the atmosphere. At sea level, cosmic-ray muons are the dominant particle species, which present an omnipresent background for radiation detectors.

The soil of the planet Earth contains substances which are naturally radioactive and provide natural radiation exposures. The most important radioactive elements that occur in the soil and in rocks are the long-lived primordial isotope potassium (^{40}K), and the isotopes of radium (^{226}Ra) and thorium (^{232}Th). The radioisotopes ^{40}K, ^{226}Ra, and ^{232}Th also occur in many building materials (such as concrete and bricks).

The most important natural isotopes that occur in air, in drinking water, and in food are the isotopes of hydrogen (tritium: ^{3}H), carbon (^{14}C), potassium (^{40}K), polonium (^{210}Po), radon (^{222}Rn), radium (^{226}Ra), and uranium (^{238}U). These natural radioactive elements accumulate in the human body after being taken in with food, water, and air so that humans themselves become radioactive. The natural radioactivity of the human body is about 9,000 Bq and originates predominantly from ^{40}K and ^{14}C. The dominant contribution to the natural radioactivity originates from the inhalation of radon isotopes. The production of radon and radon daugthers and the exhalation of radon from the ground is sketched in *Fig. 9*. Average and extreme per capita exposures from natural radiation sources are compiled in *Table 5*.

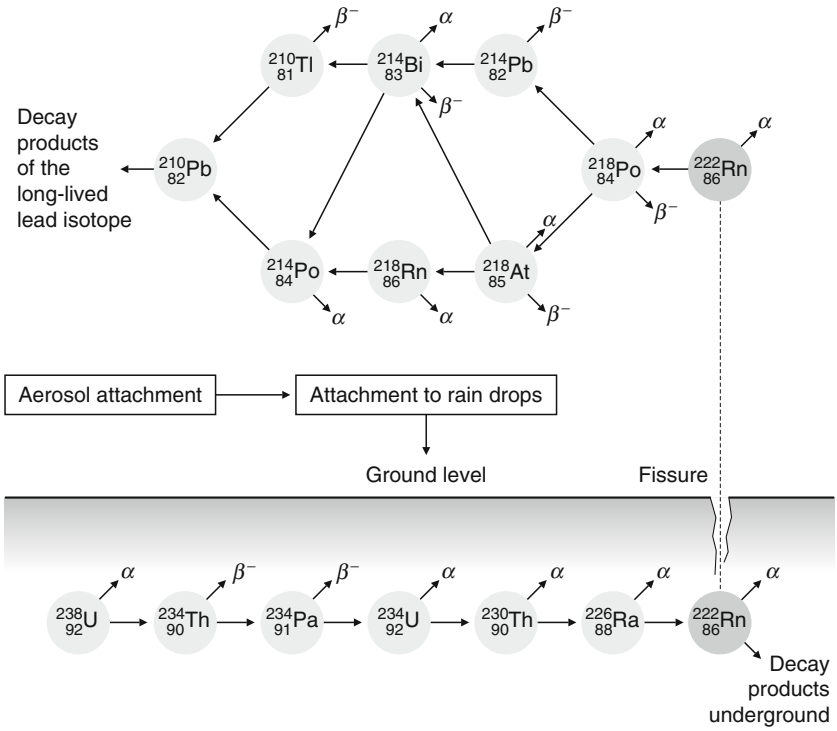


Fig. 9
Production and release of ^{222}Rn

Table 5
Radioactive per capita exposures from natural sources

Source	Average exposure per year	Highest values
Cosmic radiation	≈ 0.3 mSv	10 mSv (At high altitudes)
Terrestrial radiation	≈ 0.5 mSv	450 mSv (Ramsar, Iran)
Incorporation of radioisotopes	≈ 1.5 mSv	5 mSv (For extreme diet)

The exposure of humans from X-ray diagnostics, radiology, and radioactive substances from technical installations amounts to about 2 mSv/year. The dominant source for the per capita exposures is the application of X rays, β, and γ rays in medicine in diagnostics and therapy. Some examples are given for illustration. Taking an X-ray image of the lungs gives a whole-body dose of about 0.1 mSv. An angiography of the arteries or an X-ray of a kidney represent an exposure of 10 mSv. In contrast, an X-ray image of the teeth leads to a dose of only 0.01 mSv.

The total per capita dose for the population from natural sources and from technical installations (mainly medicine) amounts to about 4 mSv/year to 5 mSv/year. The legal limits for radiation-exposed workers and the general population mentioned above do not apply to this unavoidable natural radiation and to exposures from medical diagnostics and therapy. The legal limits only concern exposures additional to environmental radiation and medical exposures.

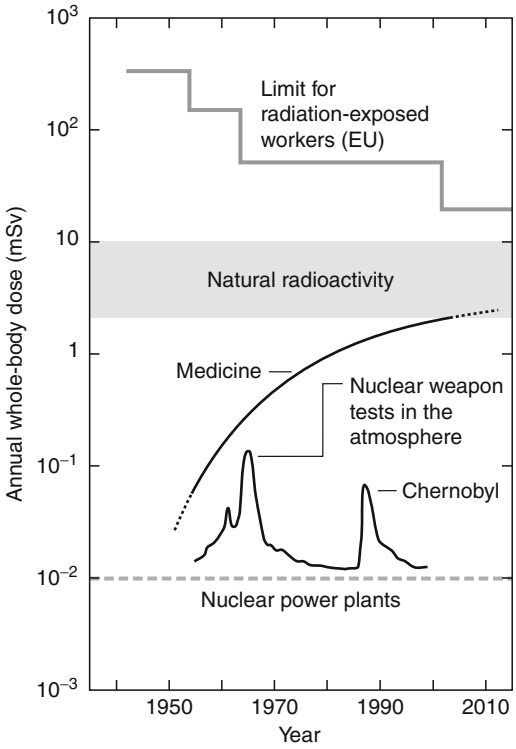


Fig. 10

Comparison of radiation from the natural environment, exposures from nuclear medicine, legal limits, and exposures from nuclear weapon tests in the atmosphere and from the Chernobyl accident (Grupen 2010)

Figure 10 shows the evolution of different contributions to the annual whole-body dose over the last 60 years.

10 Biological Effects of Radiation

Any radiation exposure might have negative effects on health. This can be considered as the basic principle of radiation protection. It is therefore no surprise that radiation damage due to ionizing radiation was first observed right after the discovery of radioactivity by Becquerel. The biological effect of ionizing radiation is a consequence of the energy transfer by ionization and excitation to body cells.

- Early effects: This radiation damage occurs immediately after the irradiation. From a whole-body dose of 0.25 Sv upward a modification of the hemogram is visible. From 1 Sv on clear symptoms of radiation sickness are to be expected. However, the recovery of the patients is

Table 6
Risk factors for radiation-induced cancer

Concerned organ or tissue	Risk factor for 10 mSv whole-body irradiation
Red bone marrow (leukemia)	50×10^{-6}
Periosteum, surface of bones	5×10^{-6}
Colon	85×10^{-6}
Liver	15×10^{-6}
Lung	85×10^{-6}
Esophagus	30×10^{-6}
Skin	2×10^{-6}
Stomach	110×10^{-6}
Thyroid gland	8×10^{-6}
Bladder	30×10^{-6}
Chest	20×10^{-6}
Ovaries	10×10^{-6}
Other organs or tissue	50×10^{-6}
Total radiation-induced cancer risk	500×10^{-6}
Genetic risk	100×10^{-6}

nearly guaranteed. For a whole-body dose of 4 Sv the chance of survival is 50%. This dose is called the lethal dose. For a dose of 7 Sv the mortality is nearly 100%.

- Delayed radiation damage: A typical late effect is cancer after a period of latency, which can amount to several decades. In contrast to prompt damage, whose effect is proportional to the received dose, delayed radiation-damage effects represent a stochastic risk, which means the probability of a damage to occur depends on the dose, but nothing can be said about whether the sickness will be serious or not. The total cancer risk per absorbed dose of 1 Sv is estimated to be about 5×10^{-2}. Detailed risk factors for specific types of cancer are given in *Table 6*.
- Genetic damage: Radiation absorption in germ cells can result in mutations. For the irradiated person mutations are not recognizable. They will only manifest themselves in the following generations. During the genetically significant age of humans (up to the age of 35) about 140 genetic mutations occur due to environmental factors. A radiation exposure of 10 mSv will add another two mutations; this corresponds only to 1 or 2% of the natural rate of mutations. The average risk factor for radiation effects, which can be inherited in the first two generations, is estimated to be 10^{-2} per 1 Sv.

Apart from damage due to ionizing radiation favorable effects after radiation exposures have also been observed. This effect is called hormesis. It is suggested that low doses of nonnatural radiation might increase the lifetime of cells. The idea is that cells are able to repair minor damage as caused by natural radioactivity and that cells become more resistant if they are regularly stimulated to repair themselves by being exposed to additional nonnatural low-level radiation. For the purposes of radiation protection, however, one must assume that any additional irradiation should be avoided if possible.

11 Conclusions

Radiation is everywhere. Air, soil, rocks, animals, and plants are radioactive. Life has developed in this environment of natural radiation. Therefore it is good to know about possible biological hazards if additional radioactive material is used in research and technology. Most countries rely on the so-called ALARA principle, which states that by using radioactive material the effect on humans and the environment should be "as low as reasonably achievable." Limits for additional radiation have to comply with exposures from natural radiation. In most places these doses are around a few millisieverts per year, sometimes with large fluctuations. Radiation doses from medical diagnosis and treatment on average are comparable to the exposure from natural radiation. The ICRP has recommended a limit for the annual whole-body dose of 20 mSv, which has been implemented in most national regulations. To survey and control such limits, suitable detector equipment must be available. The majority of detectors, as described in this article and in this handbook, can be used for this purpose. A special point for radiation detection is that these instruments must be robust and reliable. Qualified personnel is required to guarantee that the legal aspects of radiation protection are respected.

Acknowledgment

It is a pleasure to thank Dr. Tilo Stroh for a very careful reading of the manuscript and his efficient help in layouting this article in a professional way in LaTeX.

References

Amsler C et al (2008) Review of particle physics. Phy Lett B 667

Grupen C (2010) Introduction to radiation protection. Springer, Heidelberg, New York

Grupen C, Shwartz B (2008) Particle detectors. Cambridge University Press, Cambridge

Kalthoff O (1996) Berechnung der Photopeakeffizienz für koaxiale Reinstgermaniumdetektoren. Diploma Thesis, Siegen

Kraft G (1996) Radiobiology of heavy charged particles. GSI preprint, pp 96–60, offprint of the Gesellschaft für Schwerionenforschung GSI, Darmstadt, Germany

Kraft G (2000) Tumor therapy with ion beams. Nucl Instr Meth A 454:1–10

Krieger H (2002) Strahlenphysik, Dosimetrie und Strahlenschutz. Teubner Verlag, Stuttgart

Sauter E (1982) Grundlagen des Strahlenschutzes. Thiemig, München

Unger LM, Trubey DK (1982) Specific gamma-ray dose constants for nuclides important to dosimetry and radiological assessment. ORNL/RSIC-45/R1

Further Reading

Burchfield LA (2008) Radiation safety, protection and management: for homeland security and emergency response. Wiley, New York

Charles MW, Greening JR (2008) Fundamentals of radiation dosimetry, 3rd edn. Taylor & Francis, London

Eisenbud M, Gesell ThF (1997) Environmental radioactivity. Academic, San Diego

Lederer CM, Shirley VS (1979) Table of isotopes. Wiley, New York

Martin A, Harbison SA (2006) An introduction to radiation protection. Oxford University Press, A Hodder Arnold, New York
Martin JE (2006) Physics for radiation protection: a handbook. Wiley-VCH, Weinheim
Stabin MG (2007) Radiation protection and dosimetry: an introduction to health physics. Springer, Heidelberg
The International Commission on Radiological Protection, ICRP (2008) www.icrp.org/

Suppliers of Radiation-Protection Equipment

Berthold Technologies, Germany; www.bertholdtech.com/ww/en/pub/home.cfm
Bicron Radiation Measurement Products, USA; www.bicron.com/
Canberra, USA; www.canberra.com/
EG&G, USA, Germany; www.berthold.com.au/radiation_pages/berthold_radiation.html
GRAETZ, Germany; www.graetz.com/englisch/sonden_en.htm
Nuclitec, Germany; www.nuclitec.de/
Thermo Eberline, Germany; www.esm-online.de/sm/contact/index.html
Saint Gobain, France; www.bicron.com/

Part 2

Specific Types of Detectors

11 Gaseous Detectors

Maxim Titov
CEA Saclay Centre d'Études de Saclay, Gif-sur-Yvette Cedex, France

1 Introduction 240

2 Basic Principles: Ionization, Transport Phenomena, and Avalanche Multiplication 241

3 The Multi-Wire Proportional, Drift, and Time Projection Chambers 244

4 Micro-Pattern Gas Detectors (MPGD) 247

5 Micro-Pattern Gas Detector Applications 252

6 Development of Large-Area MPGDs 256

7 Pixel Readout for Micro-Pattern Gas Detectors 257

8 Summary and Outlook 261

9 Cross-References 261

References 261

C. Grupen, I. Buvat (eds.), *Handbook of Particle Detection and Imaging*,
DOI 10.1007/978-3-642-13271-1_11, © Springer-Verlag Berlin Heidelberg 2012

Abstract: Since long time, the compelling scientific goals of future high-energy physics experiments were a driving factor in the development of advanced detector technologies. A true innovation in detector instrumentation concepts came in 1968, with the development of a fully parallel readout for a large array of sensing elements – the Multi-Wire Proportional Chamber (MWPC), which earned Georges Charpak a Nobel prize in physics in 1992. Since that time radiation detection and imaging with fast gaseous detectors, capable of economically covering large detection volumes with low mass budget, have been playing an important role in many fields of physics. Advances in photolithography and microprocessing techniques in the chip industry during the past decade triggered a major transition in the field of gas detectors from wire structures to Micro-Pattern Gas Detector (MPGD) concepts, revolutionizing cell-size limitations for many gas detector applications. The high radiation resistance and excellent spatial and time resolution make them an invaluable tool to confront future detector challenges at the next generation of colliders. The design of the new micro-pattern devices appears suitable for industrial production. Novel structures where MPGDs are directly coupled to the CMOS pixel readout represent an exciting field allowing timing and charge measurements as well as precise spatial information in 3D. Originally developed for the high-energy physics, MPGD applications have expanded to nuclear physics, photon detection, astroparticle and neutrino physics, neutron detection, and medical imaging.

1 Introduction

The single-wire proportional counter, invented more than 100 years ago by E. Rutherford and H. Geiger (Geiger and Rutherford 1908), and its high-gain successor, the Geiger–Mueller counter first described in 1928 (Geiger and Mueller 1928), can be considered the ancestors of all modern gaseous detectors and were for many decades a major tool for the study of ionizing radiation. Forty years ago, in 1968, the instrumentation in experimental particle physics was revolutionized with the advent of the Multi-Wire Proportional Chamber (MWPC) (Charpak et al. 1968). With its excellent accuracy and modest rate capability, the MWPC allowed large areas to be instrumented with fast tracking detectors and were able to localize particle trajectories with sub-mm precision. Confronted by the increasing demands of particle-physics experiments, MWPCs have continuously improved over the years (Sauli 1977; Charpak and Sauli 1979, 1984; Grupen 1996; Sauli 2004; Blum et al. 2008). Gradually replacing slower detectors, numerous generations of gaseous devices, with novel geometries and exploiting various gas properties, have been developed: Drift Chamber (Walenta et al. 1971), Multi-Drift Module (Bouclier et al. 1988), JET Chamber (Drumm et al. 1980), Time Projection Chamber (TPC) (Nygren and Marx 1978), Time Expansion Chamber (Walenta 1979), Multi-Step Chamber (Charpak and Sauli 1978), Ring-Imaging Cherenkov Counter (Seguinot and Ypsilantis 1977), Resistive Plate Chamber (Santonico and Cardarelli 1981), and many others. However, limitations have been reached for MWPC-based devices in terms of maximum rate capability and detector granularity. A fundamental rate limitation of wire chambers, due to positive-ion accumulation in the gas volume, was overcome with the invention of Micro-Strip Gas Chamber (MSGC) (Oed 1988), capable of achieving position resolution of few tens of microns at particle fluxes exceeding

the MHz/mm^2 range (Barr et al. 1998). Developed for the projects at high-luminosity colliders, MSGCs promised to fill a gap between the performing but expensive solid-state detectors and cheap but rate-limited traditional wire chambers. Despite their impressive performance, detailed studies of long-term behavior at high rates revealed two possible weaknesses of the MSGC technology: formation of deposits on the electrodes, affecting gain and performances ("aging effects"), and appearance of rare but destructive discharges in the presence of highly ionizing particles (Bagaturia et al. 2002).

The invention of Micro-Pattern Gas Detectors (MPGD), in particular the Gas Electron Multiplier (GEM) (Sauli 1997), the Micro-Mesh Gaseous Structure (Micromegas) (Giomataris et al. 1996), and other micro-pattern detector schemes, offers the potential to develop new gaseous detectors with unprecedented spatial resolution, high-rate capability, large sensitive area, and operational stability (Sauli and Sharma 1999; Titov 2007). Recent developments in radiation-hardness research with state-of-the-art MPGDs revealed that they might be even less vulnerable to the radiation-induced aging effects than standard silicon micro-strip detectors, if reasonable precautions are taken on the components' quality (Titov et al. 2002; Titov 2003). In some applications, requiring very large-area coverage with moderate spatial resolutions, more coarse macro-patterned detectors, for example, thick-GEMs (THGEM) (Periale et al. 2002; Chechik et al. 2004; Breskin et al. 2009) or patterned resistive thick GEM devices (RETGEM) (Di Mauro et al. 2007) could offer an interesting and economic solution. In addition, the availability of highly integrated amplification and readout electronics allows for the design of gas-detector systems with channel densities comparable to that of modern silicon detectors. An elegant solution is the use of a CMOS pixel ASIC (Application Specific Integrated Circuit), assembled directly below the GEM or Micromegas amplification structures (Costa et al. 2001; Bellazzini et al. 2004; Campbell et al. 2005; Bamberger et al. 2007b). Modern wafer post-processing also allows for the integration of a Micromegas grid directly on top of a CMOS pixelized readout chip (Chefdeville et al. 2006). In 2008, the RD51 collaboration at CERN has been established to further advance technological developments of MPGDs and associated electronic-readout systems, for applications in basic and applied research (http://rd51-public.web.cern.ch/RD51-Public).

2 Basic Principles: Ionization, Transport Phenomena, and Avalanche Multiplication

The low density of gaseous media sets basic limitations to the performances of detectors. The process of detection in gas proportional counters starts with the inelastic collisions between the incident particle and gas molecules. These collisions lead to excitation of the medium (followed by the emission of the light, the basis of scintillation detectors) and ionization, the primary signal for tracking devices. The number N_P and the space distribution of the primary ionization clusters depend on the nature and energy of the radiation. The primary electrons can often have enough energy to further ionize the medium; the total number of electron–ion pairs (N_T) is usually a few times larger than the number of primaries (N_P). ➤ *Table 1* provides values of relevant parameters in some commonly used gases at NTP (normal temperature and pressure) for unit-charge minimum-ionizing particles (MIPs) (Sauli and Titov 2010).

The primary statistics determines several intrinsic performances of gas detectors, such as efficiency, time resolution, and localization accuracy. The actual number of primary interactions follows the Poisson's statistics; the inefficiency of a perfect detector with a thin layer of gas is

Table 1

Properties of noble and molecular gases at normal temperature and pressure (NTP: 20° C, one atm). E_X, E_I: first excitation, ionization energy; W_I: average energy per ion pair; $dE/dx|_{min}$, N_P, N_T: differential energy loss, primary and total number of electron–ion pairs per cm, for unit-charge minimum-ionizing particles. Values often differ, depending on the source, and those in the table should be taken only as approximate (Sauli and Titov 2010)

Gas	Density (mg cm^{-3})	E_x (eV)	E_I (eV)	W_I (eV)	$dE/dx\|_{min}$ (keV cm^{-1})	N_P (cm^{-1})	N_T (cm^{-1})
He	0.179	19.8	24.6	41.3	0.32	3.5	8
Ne	0.839	16.7	21.6	37	1.45	13	40
Ar	1.66	11.6	15.7	26	2.53	25	97
Xe	5.495	8.4	12.1	22	6.87	41	312
CH_4	0.667	8.8	12.6	30	1.61	28	54
C_2H_6	1.26	8.2	11.5	26	2.91	48	112
iC_4H_{10}	2.49	6.5	10.6	26	5.67	90	220
CO_2	1.84	7.0	13.8	34	3.35	35	100
CF_4	3.78	10.0	16.0	54	6.38	63	120

given by e^{-N_P}. Therefore, in one mm of Ar/CO_2 (70:30), approximately 6% of all MIPs do not release a single primary electron cluster and therefore cannot be detected. The total energy loss, sum of primary and secondary ionization, follows a statistical distribution described by a Landau function, with characteristic tails toward higher values. A simple composition law can be used for gas mixtures: for example, the number of primary (N_P) and total (N_T) electron–ion pairs produced by MIP in a 1 cm of Ar/CO_2 (70:30) mixture at NTP is:

$$N_P = 25 \cdot 0.7 + 35 \cdot 0.3 = 28 \ \frac{\text{pairs}}{\text{cm}}; \quad N_T = \frac{2530}{26} \cdot 0.7 + \frac{3350}{35} \cdot 0.3 \approx 97 \ \frac{\text{pairs}}{\text{cm}}. \tag{1}$$

While charged particles release an ionization trail of primary electron clusters, low-energy X rays undergo a single localized interaction, usually followed by the emission of the photoelectron, accompanied by a lower-energy photon or Auger electron. For example, a 5.9 keV X ray converts in argon mainly on a K shell (3.2 keV); the emitted photoelectron with energy $E_\gamma - E_K \sim 2.7$ keV has a practical range in the detector of ~200 μm. In addition, with 85% probability another (Auger) electron with energy ~3 keV (~250 μm range in argon) is ejected; in the remaining cases, a 3 keV K–L fluorescence photon is produced with a mean absorption length of 40 mm. The sum of the energies of photoelectron and Auger electron is responsible for the main 5.9 keV peak, while fluorescence mechanism leads to the Ar escape peak. The total number of electron–ion pairs created by X rays absorbed in argon can be evaluated by dividing its energy by the W_I: $\frac{5900}{26} \approx 227$.

Once released in the gas, and under the influence of an applied electric field, electrons and ions drift in opposite directions and diffuse toward the electrodes. The scattering cross section is determined by the details of atomic and molecular structure. Therefore, the drift velocity and diffusion of electrons depend very strongly on the nature of the gas, specifically on the inelastic cross section involving the rotational and vibrational levels of molecules. In noble gases, the inelastic cross section is zero below excitation and ionization thresholds. Large drift velocities are achieved by adding polyatomic gases (usually CH_4, CO_2, or CF_4), having large inelastic cross sections at moderate energies, which results in "cooling" electrons into an energy range

of the Ramsauer–Townsend minimum (located at ~0.5 eV) of the elastic cross section of argon. The reduction in both the total electron scattering cross section and the electron energy results in a large increase of the electron drift velocity (for a compilation of electron–molecule cross sections, see http://rjd.web.cern.ch/rjd/cgi-bin/cross). Another principal role of the polyatomic gas is to absorb the ultraviolet (UV) photons emitted by the excited inert gas atoms. The quenching of UV photons occurs through the photodecomposition of polyatomic molecules. Extensive collections of experimental data (Peisert and Sauli 1984) and theoretical calculations based on transport theory (Biagi 1999) permit estimates of drift and diffusion properties in pure gases and their mixtures. In a simple approximation, gas kinetic theory provides the following relation between drift velocity, v, and the mean collision time between electron and molecules, τ (Townsend's expression): $v = eE\tau/m$. Values of drift velocity for some commonly used gases at NTP, computed with the MAGBOLTZ program (see http://consult.cern.ch/writeup/magboltz), are given in ❯ *Fig. 1*. Using fast CF_4-based mixtures at fields around kV/cm^{-1}, the electron drift velocity is around 10 cm · μs^{-1}. Since the collection time is inversely proportional to the drift velocity, diffusion is less in gases such as CF_4 that have high drift velocities. In the presence of an external magnetic field, the Lorentz force acting on electrons between collisions deflects the drifting electrons and modifies the drift properties. For parallel electric and magnetic fields, drift velocity and longitudinal diffusion are not affected, while the transverse diffusion can be strongly reduced: $\sigma_T(B) = \sigma_T(B = 0)/\sqrt{1 + \omega^2\tau^2}$, where ω is the (angular) cyclotron frequency of the electron. This reduction is exploited in a TPC to improve spatial resolution.

In mixtures containing electronegative molecules such as O_2, H_2O, or CF_4, electrons can be captured to form negative ions. Capture cross sections are strongly energy dependent, and therefore the capture probability is a function of the applied field. For example, the electron is attached to the oxygen molecule at energies below 1 eV. The three-body electron attachment coefficients may differ greatly for the same addition in different mixtures. As an example,

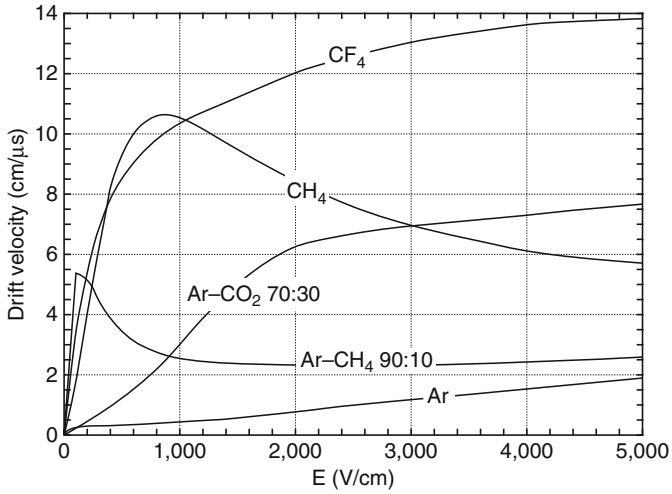


Fig. 1

Computed electron drift velocity with the MAGBOLTZ program as a function of the electric field in several gases at NTP and $B = 0$. For different conditions, the horizontal axis must be scaled inversely with the gas density (Sauli and Titov 2010)

at moderate fields (up to 1 kV/cm), the addition of 0.1% of oxygen to an Ar/CO_2 mixture results in an electron-capture probability about 20 times larger than for the same addition to Ar/CH_4. Carbon tetrafluoride is not electronegative at low and moderate fields, making its use attractive as drift gas due to its very low diffusion. However, CF_4 has a large cross section for dissociative attachment in the 6–8 eV electron energy range (Christophorou et al. 1996). Depending on geometry, some signal reduction and resolution loss can be expected using this gas.

The primary ionization signal is very small in a gas layer: in 1 cm of Ar/CO_2 (70:30) at NTP only ~100 electron–ion pairs are created. Therefore, one has to use an "internal gas amplification" mechanism to generate a detectable signal in gas counters; excitation and subsequent photon emission participate in the avalanche-spread processes and can also be detected by optical means. If the electric field is increased sufficiently, electrons gain enough energy between collisions to ionize molecules. Above a gas-dependent threshold, the mean free path for ionization, λ_i, decreases exponentially with the field; its inverse, $\alpha = 1/\lambda_i$, is the first Townsend coefficient. In wire counters, most of the increase of the avalanche particle density occurs very close to the anode wires, and a simple electrostatic consideration shows that the largest fraction of the detected signal is due to the motion of positive ions receding from the wires. The electron component, although very fast, contributes very little to the signal. This determines the characteristic shape of the detected signals in the proportional mode: a fast rise followed by a gradual increase. The slow component, the so-called 'ion tail' that limits the time resolution of the counter, is usually removed by differentiation of the signal. In uniform fields, N_0 initial electrons multiply over a length x forming an electron avalanche of size $N = N_0 e^{\alpha x}$; N/N_0 is the gain of the counter. Gas amplification of 10^3–10^4 is usually required in order to provide signals with sufficient amplitudes for conventional electronics.

Positive ions released by the primary ionization or produced in the avalanches drift and diffuse under the influence of the electric field. Negative ions may also be produced by electron attachment to gas molecules. The drift velocity of ions in the fields encountered in gaseous counters (up to few kV/cm) is typically about three orders of magnitude lower than for electrons. The ion mobility, μ, the ratio of drift velocity to electric field, is constant for a given ion type up to very high fields (McDaniel and Mason 1973; Shultz et al. 1977). For mixtures, due to a very effective charge-transfer mechanism, only ions with the lowest ionization potential survive after a short path in the gas. The diffusion of ions, both σ_L and σ_T, are proportional to the square root of the drift time, with a coefficient that depends on temperature but not on the ion mass. Accumulation of ions in the gas volume may induce gain reduction and field distortions.

3 The Multi-Wire Proportional, Drift, and Time Projection Chambers

The invention of the Multi-Wire Proportional Chambers (MWPC) revolutionized the field of radiation detectors (Charpak et al. 1968; Charpak and Sauli 1984). In the original design, the MWPC consists of a set of parallel, evenly spaced, anode wires stretched between two cathode

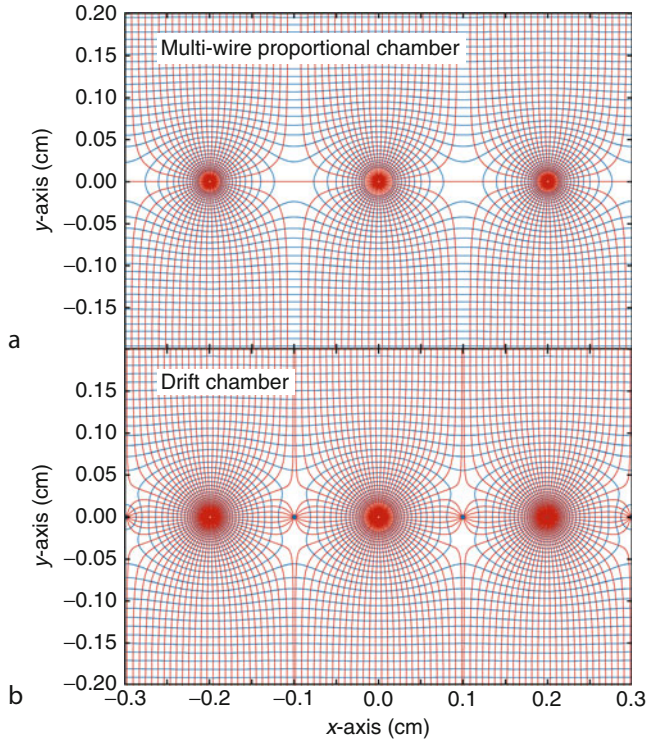


Fig. 2

Electric field lines and equipotentials in (a) a multi-wire proportional chamber and (b) a drift chamber. Thin anode wires (running perpendicular to the page) are placed equidistantly between two parallel cathode planes and act as a set of independent proportional counters (Charpak and Sauli 1979, 1984)

planes. Applying a potential difference between anodes and cathodes, field lines and equipotentials develop as shown in *Fig. 2a*. Typical values for the anode-wire spacing range between 1 and 5 mm, the anode-to-cathode distance is 5–10 mm. The operation gets increasingly difficult at smaller wire spacings. For example, the electrostatic repulsion for thin (10 μm) anode wires causes mechanical instability above a critical wire length, which is less than 25 cm for 1-mm wire spacings (Sauli and Titov 2010). In an MWPC, electrons formed by ionization of the gas drift toward the plane of anode wires, initially in a nearly uniform field. The signal multiplication process, which begins a few radii from the anode, is over after a fraction of a nanosecond, leaving the cloud of positive ions receding from the wires. A large negative-polarity-induced pulse appears on the anode on which the avalanche is collected, while the neighboring wires show smaller positive-amplitude pulses. Detection of charge over a predefined threshold provides the transverse coordinate to the anode wire with an accuracy comparable to that of the wire spacing. With a digital readout and $s = 1$ mm wire spacing, the spatial resolution is limited to $\sigma = \frac{s}{\sqrt{12}} = 300$ μm. The cathode planes can be also fabricated in the form of isolated strips

or group of wires (Sauli 1994). Making use of the positive charge signals induced by avalanches on segmented cathodes, the so-called electronic center-of-gravity (COG) method, allows bidimensional localization of the ionizing event. Due to the statistics of energy loss and asymmetric ionization clusters, the position accuracy is ~50 μm rms for tracks perpendicular to the wire plane, but degrades to ~250 μm at 30° to the normal (Charpak et al. 1979).

Drift chambers, developed in the early 1970s, can estimate the position of a track by exploiting the arrival time of electrons at the anodes if the time of interaction is known (Walenta et al. 1971). The distance between anode wires is usually several centimeters allowing coverage of large areas at reduced cost. In the original design, a thicker wire at proper voltage between anodes (field wire) reduces the field at the middle point between anodes and improves charge collection (◉ *Fig. 2b*). In some drift-chambers design, and with the help of suitable voltages applied to field-shaping electrodes, the electric field structure is adjusted to improve the linearity of the space-to-drift-time relation, resulting in better spatial resolution (Breskin et al. 1975). Drift chambers can reach a spatial resolution from timing measurement of order 100 μm (rms) or better for minimum-ionizing particles, depending on geometry and operating conditions. However, a degradation of resolution is observed (Breskin et al. 1978) due to primary-ionization statistics for tracks close to the anode wires, caused by the spread in arrival time of the nearest ionization clusters. For an overview of detectors exploiting the drift time for coordinate measurement, see Grupen (1996) and Blum et al. (2008).

The "ultimate" drift chamber is the TPC concept invented in 1976 (Nygren and Marx 1978), which combines a measurement of drift time and charge induction on cathodes. It has been the prime choice for large tracking systems in e^+e^- colliders (PEP-4 (Nygren et al. 1976), ALEPH (Decamp et al. 1990), DELPHI (Sacquin et al. 1992)) and proved its unique resolving power in heavy-ion collisions (NA49 (Afanasev et al. 1999), STAR (Wieman et al. 1997)). A TPC consists of a large gas volume, with an uniform electric field applied between the central electrode and a grid at the opposite side. The ionization trails produced by charged particles drift toward the end plate, segmented into 2D readout pads; the third coordinate is measured using the drift-time information. A good knowledge of electron drift velocity and diffusion properties is required, which has to be combined with the modeling of electric fields in the structures (http://consult.cern.ch/writeup/magboltz, http://www.ansoft.com). Gaseous TPCs are often designed to operate within a strong magnetic field (typically parallel to the drift field) so that particle momenta can be estimated from the track curvature. A conventional readout structure, based on MWPC and pads, is a benchmark for the most modern ALICE TPC, which incorporates innovative and state-of-the-art technologies, from the mechanical structures to the readout electronics and data processing chain (Meyer 2003; Antonczyk et al. 2006). TPC performance has been recently improved by replacing wire planes with MPGDs; T2K TPC is an example of the large-volume TPC based on the Micromegas amplification structure (Anvar et al. 2009).

The production of positive ions in the avalanches and their slow drift before neutralization result in a rate-dependent accumulation of positive charge in the detector. This may result in significant field distortion, gain reduction, and degradation of spatial resolution. As shown in ◉ *Fig. 3*, the gain of an MWPC starts to drop at particle rates above $10^4\ \mathrm{mm}^{-2}\ \mathrm{s}^{-1}$, leading to a loss of detection efficiency (Breskin et al. 1975). Together with the practical difficulty to manufacture detectors with sub-mm wire spacing, this has motivated the development of new-generation gaseous detectors for high-luminosity accelerators.

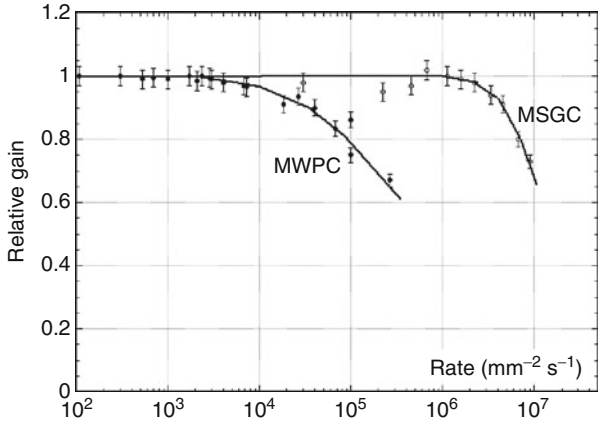


Fig. 3
Normalized gas gain as a function of particle rate for Multi-Wire Proportional Chamber (MWPC) (Breskin et al. 1975) and Micro-Strip Gas Chamber (MSGC) (Barr et al. 1998)

4 Micro-Pattern Gas Detectors (MPGD)

Despite various improvements, position-sensitive detectors based on wire structures are limited by basic diffusion processes and space-charge effects in the gas to localization accuracies of 50–100 μm (Aleksa et al. 2000). Modern photolithographic technology led to the development of novel Micro-Pattern Gas Detector (MPGD) concepts (Sauli and Sharma 1999), revolutionizing cell-size limitations for many gas detector applications. By using pitch size of a few hundred microns, an order of magnitude improvement in granularity over wire chambers, these detectors offer an intrinsic high-rate capability (>10^6 Hz/mm^2), excellent spatial resolution (~30 μm), multiparticle resolution (~500 μm), and single-photoelectron time resolution in the ns range.

The Micro-Strip Gas Chamber (MSGC), a concept invented in 1988, was the first of the microstructure gas detectors (Oed 1988). The principle of MSGCs resembles a multi-anode proportional counter, with a set of parallel metal strips laid on a thin resistive substrate, and alternately connected as anodes and cathodes (see *Fig. 4*). Through an accurate and simple photolithography process, the anode strips can be made very narrow (~10 μm) with a typical pitch (distance between strips) of ~100 μm. When appropriate potentials are applied to the electrodes, negative with respect to the anode on both drift electrode and cathodes, electrons released in the drift volume move toward the strips and start to multiply, as they approach the high-field region. All field lines from the drift volume terminate on the anodes, resulting in full electron-collection efficiency. Owing to the small anode-to-cathode distance, the fast removal of positive ions by nearby cathode strips reduces space-charge build-up, and provides a greatly increased rate capability of the MSGC, compared with wire chambers (Barr et al. 1998) (see *Fig. 3*). As in the conventional proportional counter, a large fraction of the negative signal on the anodes is induced by the rapidly moving ions, resulting in a fast rise time. Signals of opposite polarity are induced on neighboring cathodes and on the back-plane electrode. Although their primary use has been as position-sensitive detectors for particle tracking, MSGCs can also provide energy information for spectroscopic measurements. Their energy resolution is enhanced

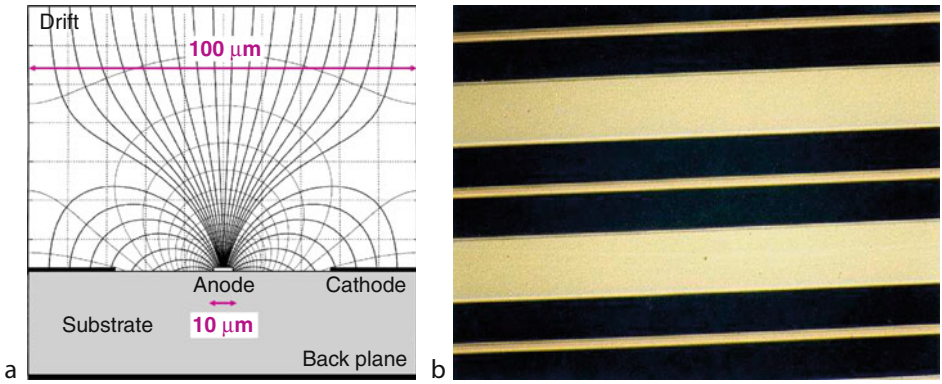


Fig. 4

(a) Schematic view, equipotentials, and field lines in the MSGC (Sauli and Sharma 1999). The back-plane potential has been selected to prevent field lines entering the dielectric; (b) Microscopic image of the MSGC electrodes. On a resistive substrate, narrow (~10 μm) anode strips alternate with wider cathodes; the pitch is 100 μm (http://gdd.web.cern.ch/GDD/)

compared to a wire counter, because the avalanche fluctuations are minimized, as a result of the sharp gradient in the electric field strength near the anode surface (Miyamoto and Knoll 1997).

Despite their promising performance, experience with MSGCs has raised serious concerns about their long-term behavior. There are several major processes, particularly at high rates, leading to the following MSGC operating instabilities: substrate charging-up and time-dependent distortions of the electric field, surface deposition of polymers ("aging") during sustained irradiation, and destructive micro-discharges under exposure to heavily ionizing particles (Charpak and Sauli 1984; Bouclier et al. 1996). The physical parameters used to manufacture and operate these detectors (substrate material, metal of strips, type and purity of the gas mixture) appeared to play dominant roles in determining the medium- and long-term stability. To avoid surface charging, the substrate must have some finite electrical conductivity, and a number of different recipes for slightly conducting glass and/or coatings on other materials have emerged in the development of these devices.

The problem of discharges is the intrinsic limitation of all single-stage micro-pattern detectors in hadronic beams (Peskov et al. 1997; Bressan et al. 1999a; Ivaniouchenkov et al. 1999). Whenever the total charge in the avalanche exceeds a value of 10^7–10^8 electron–ion pairs (Raether limit), an enhancement of the electric field in front of and behind the primary avalanche induces a fast growth of a filament-like streamer followed by breakdown. This has been confirmed under a wide range of operating conditions and multiplying gaps (Ivaniouchenkov et al. 1998; Fonte et al. 1999; Iacobaeus et al. 2002; Peskov and Fonte 2009). In the high fields and narrow gaps, the MSGC turned out to be prone to irreversible discharges induced by heavily ionizing particles and destroying the fragile electrode structure (Bagaturia et al. 2002).

Nevertheless, the detailed studies on their properties, and, in particular, on the radiation-induced processes leading to discharge breakdown, led to the development of more powerful devices with improved reliability and radiation hardness. Modern MPGD structures

can be grouped into two large families: hole-type structures and micromesh-based detectors. The hole-type structures are GEMs, THGEM, RETGEM, and Micro-Hole and Strip Plate (MHSP) elements. The micromesh-based structures include: Micromegas, "Bulk" Micromegas, "Microbulk" Micromegas, and "InGrid."

Introduced in 1996 by F. Sauli (1997), a gas electron multiplier (GEM) detector consists of a thin-foil copper-insulator-copper sandwich chemically perforated to obtain a high density of holes in which avalanche formation occurs. The GEM manufacturing method, developed at CERN, is a refinement of the double-side printed circuit technology. The copper-clad polymer is engraved on both sides with the desired hole pattern; controlled immersion in a kapton-specific solvent opens the channels in the insulator. The hole diameter is typically between 25 μm and 150 μm, while the corresponding distance between holes varies between 50 μm and 200 μm. The central insulator is usually (in original design) the polymer kapton, with a thickness of 50 μm. Application of a potential difference between the two sides of the GEM generates the electric field indicated in ❯ *Fig. 5a*. Each hole acts as an independent proportional counter. Electrons released by the primary ionization particle in the upper drift region (above the GEM foil) are drawn into the holes, where charge multiplication occurs in the high electric field (50–70 kV/cm). Most of avalanche electrons are transferred into the gap below the GEM (Bressan et al. 1999b; Bachmann et al. 1999). Several GEM foils can be cascaded (see ❯ *Fig. 5b*), allowing the multilayer GEM detectors to operate at an overall gas gain above 10^4 in the presence of highly ionizing particles, while strongly reducing the risk of discharges (Bachmann et al. 2001, 2002). This is a major advantage of the GEM technology. Systematic measurements with the multiple-GEM structures confirm that the gains and charge-transfer processes are predictable from electrostatic considerations and an avalanche-development model (Bachmann et al. 1999; Killenberg et al. 2003; Villa et al. 2010). A unique property of GEM detectors is a full decoupling

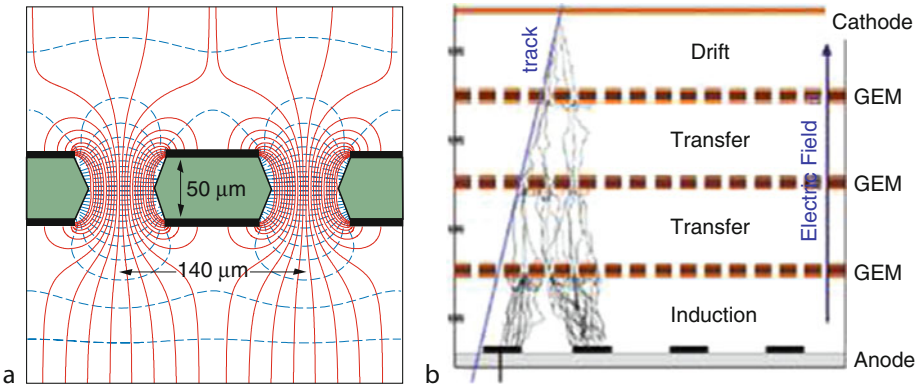


Fig. 5

(a) Schematic view and typical dimensions of the hole structure in the GEM amplification cell. Electric field lines (*solid*) and equipotentials (*dashed*) are shown (Sauli and Sharma 1999); (b) Schematic view of the triple-GEM detector. The original ionization occurs in the region labeled "Drift," and the ionization electrons are drawn downward to the GEM foil. The amplified electrons emerging from the first GEM are drifted to the second GEM, where they again multiplied. One more GEM stage provides further amplification, and its output is collected at the readout plane (Sauli 2007)

of the amplification stage (GEM) and the readout electrode (PCB), which is kept at ground potential. The signal detected on the PCB is entirely due to electrons, without ion tail, and is typically few tens of nanoseconds for 1 mm-wide induction gap. The 2D imaging of the primary ionization with a spatial resolution of about 50 μm can be achieved by collecting the charge on a two-dimensional readout board, placed below the last GEM. Cascaded GEMs reach gains above 10^5 with single electrons; this permitted the use of a photocathode in conjunction with multiple GEM foils to develop gaseous imaging photomultipliers (GPM) (Breskin et al. 2003). Moreover, with an appropriate choice of GEM fields and geometry, both photon and ion feedback can be strongly suppressed (Garty et al. 1999).

Several groups reported progress in hybrid amplification structures, where the principles of different charge-multiplication devices are integrated in one detector. The MHSP is a GEM-like hole structure with thin anode and cathode strips patterned on the bottom electrode and GEM-like holes on the top side. Avalanche electrons are multiplied within the holes and additionally on the anode strips (Veloso et al. 2000). A significant part of avalanche ions is collected on the MHSP cathode strips (Lyashenko et al. 2004). A real breakthrough in ion-blocking capability has been achieved with cascades combining GEMs and MHSPs operated in different modes regarding the voltages and orientation schemes: MHSP, reversed-MHSP (R-MHSP) (Lyashenko et al. 2006), and flipped-reversed-MHSP (F-R-MHSP) (Lyashenko et al. 2007).

Introduced in 1996 by Y. Giomataris, the Micromegas is a thin parallel-plate avalanche counter, as shown in ❯ *Fig. 6* (Giomataris et al. 1996). It consists of a few millimeters of drift region and a narrow multiplication gap (25–150 μm) between a thin metal grid (micromesh) and the readout electrode (strips or pads of conductor printed on an insulator board). The mesh itself, a standard component in high-resolution TV screens, is commercially available in various sizes and shapes. Regularly spaced supports (insulating pillars) guarantee the uniformity of the gap between the anode plane and the micromesh, at the expense of a small, localized loss of

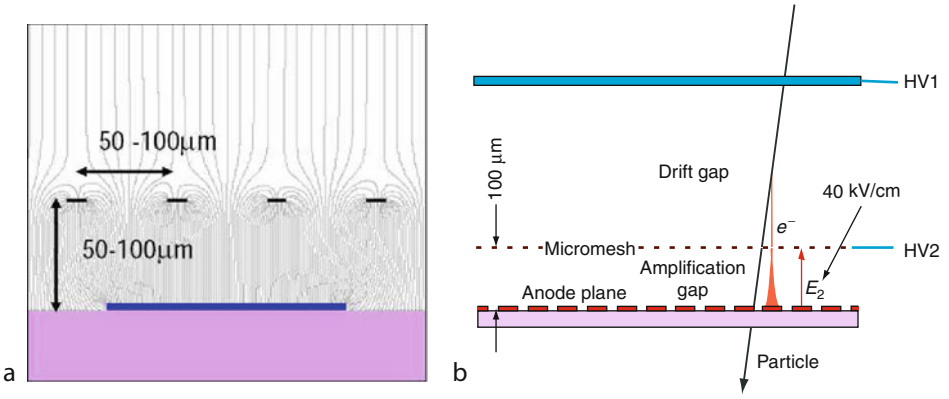


Fig. 6

(a) Electric field map in Micromegas (Charpak et al. 2002); (b) Basic principle of a Micromegas detector. A metallic micromesh separates a few millimeters of low-field (~1 kV/cm) drift region from the high-field (~50 kV/cm) amplification region. Ionization electrons drift downward and pass through the micromesh into the multiplication region (Charpak et al. 1998)

efficiency. Electrons from the primary ionization drift through the holes of the mesh into the narrow multiplication gap, where they are amplified. The electric field is homogeneous both in the drift (electric field ~1 kV/cm) and amplification (~40–80 kV/cm) gaps and only exhibits a funnel-like shape close to the openings of the micromesh: field lines are compressed into a small diameter of the order of few microns, depending on the electric field ratio between the two gaps (see *Fig. 6a*). The Micromegas retains the rate capability and energy resolution of the parallel-plate counter. The Micromegas energy resolution depends on the uniformity of the amplification field, the drift/multiplication field ratio, which influences the electrical transparency of the mesh, and the gain. In the narrow multiplication region, small variations of the amplification gap are compensated by an inverse variation of the Townsend coefficient, resulting in a more uniform gain (Giomataris 1998). By proper choice of the applied voltages, most of the positive ions are collected by the micromesh; this prevents space-charge accumulation and induces very fast signals with a small ion tail (50–100 ns length). The small amplification gap produces a narrow avalanche, giving rise to excellent spatial resolution: 12 μm accuracy, limited by the micromesh pitch, has been achieved for MIPs (Derre et al. 2001; Derre and Giomataris 2002), and single-photoelectron time resolution better than 1 ns (Derre et al. 2000).

Efforts have been focused on producing the Micromegas amplification region as a single piece, using the newly developed "Bulk" technology (Giomataris et al. 2006). A woven mesh is laminated on a PCB covered by a photo-imageable polyamide film, and the pillars are made by a photochemical technique with insulation through the grid. Such an "all-in-one" detector, called "Bulk" Micromegas, is robust and allows regular production of large, stable, and inexpensive detector modules. A new Micromegas manufacturing technique "Microbulk," based on kapton etching technology, has been recently developed, resulting in further improvement of the detector characteristics, such as flexible structure, low material budget, high radio-purity, and time stability, but is delicate because of the way it is manufactured (Andriamonje et al. 2010). The novelty of the "Microbulk" Micromegas is the absence of pillars, the whole detector being constructed out of a thin kapton foil doubly clad with copper. Excellent energy resolution has been obtained for these detectors, reaching 11% FWHM for the 5.9 keV X rays and 1.8% FWHM of the 5.5 MeV α peak for the Am^{241} source (Dafni et al. 2009). It differs only slightly from the accuracy obtained with gaseous scintillation proportional counters (Policarpao et al. 1972), which is ultimately limited by the Fano factor .

The success of GEMs and glass capillary plates triggered the development of coarse and more robust structures, "optimized GEM" (Periale et al. 2002, 2003) followed by the THGEM (Chechik et al. 2004; Shalem et al. 2006; Breskin et al. 2009) gaseous multiplier. These are produced by standard printed circuit technology: mechanical drilling of 0.3–1 mm diameter holes, etched at their rims to enhance high-voltage stability (see *Fig. 7a*); different PCB materials can be used, of typical thicknesses of 0.4–1 mm and hole spacings of 0.7–1.2 mm. These multipliers exhibit specific features: the electron collection and transport between cascaded elements is more effective than in GEM because the THGEM's hole diameter is larger than the electron's diffusion range when approaching the hole. Systematic studies are ongoing to assess the operating properties of such devices. It appears that the presence of a rim, a circular region around the holes where metal is etched away, plays a major role in their performance. A small or zero rim allows to achieve better gain stability in time, while large rims permit to attain larger maximum gain and to reduce discharges, at the cost of significant charging-up effects after detector is powered. The THGEM detectors have been studied in gaseous-photomultiplier configurations – coupled to semitransparent (Chechik et al. 2003) or reflective (Mormann et al. 2004) CsI photocathodes (Chechik et al. 2005). Effective gas-amplification factors of 10^5 and 10^7

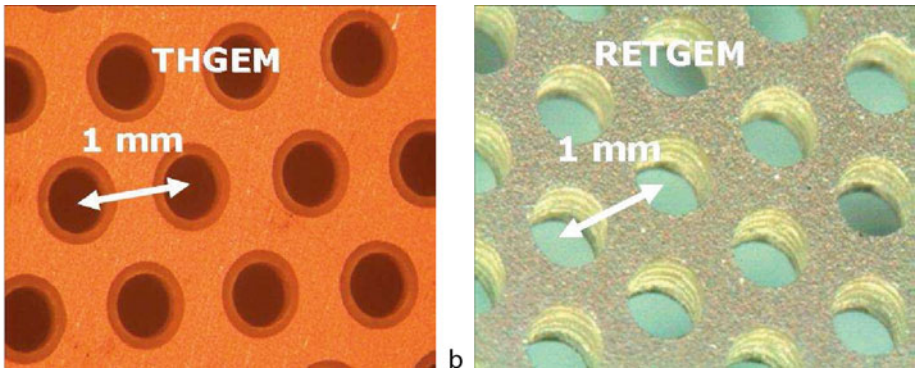


Fig. 7

(a) Photo of the Thick GEM (THGEM) multiplier. A rim of 0.1 mm is chemically etched around mechanically drilled holes to reduce discharges (Breskin et al. 2009); (b) Photo of the RETGEM detector with resistive kapton electrodes (Peskov private communication)

can be reached in single and cascaded double-THGEM elements, which permits efficient imaging of light at single-photon level (Shalem et al. 2006). Stable operation at photon fluxes exceeding 1 MHz/mm^2 was recorded together with sub-mm localization accuracy and timing in the 10 ns range (Breskin et al. 2009; Azevedo et al. 2010).

A novel spark-protected version of thick GEM with resistive electrodes (RETGEM) has been recently developed, where the Cu-clad conductive electrodes are replaced by resistive materials (Di Mauro et al. 2007; Oliveira et al. 2007; Di Mauro et al. 2009b). Sheets of carbon-loaded kapton 50 μm thick or screen-printed resistive surfaces are attached onto both surfaces of the PCB to form resistive-electrode structures; holes 0.3 mm in diameter with a pitch of 0.6 mm are mechanically drilled (see *Fig. 7b*). At low counting rates, the detector operates as a conventional THGEM with metallic electrodes, while at high intensities and in case of discharges the behavior is similar to a resistive-plate chamber.

5 Micro-Pattern Gas Detector Applications

The performance and robustness of MPGDs have encouraged their use in high-energy and nuclear physics, UV- and visible-photon detection, astroparticle and neutrino physics, X-ray imaging and neutron detection, and medical physics. Common themes for applications are low mass, large active areas (~m^2), high spatial resolution (<50 μm), high-rate capability, and radiation hardness.

Due to the variety of geometries and flexible operating parameters, MPGDs are widely used for high-rate particle tracking and triggering in nuclear- and particle-physics experiments. COMPASS, a high-luminosity experiment at CERN, pioneered the use of large-area (~40 × 40 cm^2) GEM and Micromegas detectors close to the beam line with particle rates of 25 kHz/mm^2. Both technologies achieved a tracking efficiency of close to 100%, a spatial resolution of 70–100 μm, and a time resolution of ~10 ns (Ketzer et al. 2004; Bernet et al. 2005). For the long-term future (2012 and beyond), a set of triple-GEM and

Micromegas detectors with a hybrid readout, consisting of pixels of ~1 mm^2 in the beam area and 2D strips in the periphery, was built (Austregesilo et al. 2009; Neyret et al. 2009). High-resolution planar triple-GEM detectors are already used in the LHCb Muon System (Alfonsi et al. 2004), for the TOTEM Telescopes (Lami 2009) and being developed for the PANDA experiment at the FAIR facility (Vandenbroucke 2009), in the SBS spectrometer for the Hall A at JLAB (http://hallaweb.jlab.org/12GeV/SuperBigBite), and for the forward tracker of the STAR experiment at RHIC (Simon et al. 2009). Using fast CF_4-based mixtures, a time resolution of about 5 ns rms is achieved (Barouch et al. 1999), adequate to resolve two bunch crossings at the Large Hadron Collider (LHC). Therefore, large-area Micromegas and GEMs are currently being developed for the upgrade of the ATLAS and CMS Muon Systems. Although normally used as flat detectors, GEMs and Micromegas can be bent to form cylindrically curved ultra-light tracking systems, without support and cooling structures, as preferred for inner tracker (barrel) and vertex applications (Balla et al. 2009; Aune et al. 2009a). *Figure 8a* shows a photo of the curved flexible "Bulk" Micromegas prototype for the CLAS12 experiment at JLAB.

For the future International Linear Collider applications, both GEM and Micromegas devices are foreseen as one of the main options for the TPC (Arogancia et al. 2009; De Lentdecker 2009; Matsuda 2010). Compared to wire chambers, they offer a number of advantages: negligible $\mathbf{E}\times\mathbf{B}$ track-distortion effects, the narrow Pad Response Function (PRF) and the intrinsic suppression of ion feedback, relaxing the requirement on gating of the devices, and, depending on the design, possibly allowing non-gated operation of the TPC. Large-area MPGD systems are also studied as a potential solution for the highly granular digital hadron calorimeter. Implementations in GEM (Yu 2009), Micromegas (Adloff et al. 2009), and THGEM technologies have been proposed.

X-ray imaging detectors, based on MPGDs, are being used for diffraction experiments at synchrotron radiation facilities (Smith 2006) and could serve as a powerful diagnostic tool for magnetic fusion plasmas (Pacella et al. 2003, 2004). An innovative GEM-based system, which combines fast imaging of X-ray emissions with spectral resolution in the VUV range (0.2–10 keV), has been developed to study 2D dynamics of plasma instabilities and to control the core plasma position, both being crucial issues for fusion researches (Pacella et al. 2003).

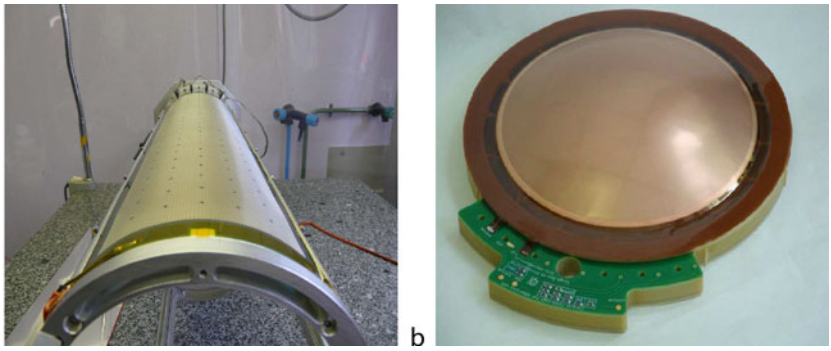


Fig. 8
Images of: (a) curved cylindrical "Bulk" Micromegas (Procureur private communication); (b) spherical GEM detector (Duarte Pinto et al. 2009b)

Another recent effort is the construction of the spherical GEM detector for parallax-free X-ray diffraction measurements, developed in collaboration with industry (see *Fig. 8b*) (Duarte Pinto et al. 2009b).

For the field of rare-event searches, "Bulk" and "Microbulk" Micromegas are used in searching for solar axions (CAST), where the expected signal comes from solar axion conversions into low-energy photons of 1–10 keV range (Dafni et al. 2008). The background level in the CAST experiment is determined by the detector's radio purity and particle discrimination efficiency: the main contribution comes from cosmic ray muons, which can be suppressed by applying a pattern-recognition analysis, since muons have a completely different signature than a few keV X rays. "Microbulk" Micromegas with high-granularity 2D readout can largely reduce the background event rate down to 5×10^{-5} keV^{-1} cm^{-2} s^{-1}, exploiting its excellent energy resolution and time stability (Aune et al. 2009b). Similar detectors are being studied for the future neutrinoless double beta decay experiment (NEXT) using a high-pressure gas xenon TPC (Dafni et al. 2009).

There has been a considerable interest in the field of gaseous photomultipliers (GPM), by combining MPGD with semitransparent or reflective CsI photocathodes (PCs), to localize ultraviolet photons emitted in radiators by Cherenkov effect (Chechik et al. 2004; Mormann et al. 2004; Chechik and Breskin 2008; Buzulutskov 2008). In this case, the multiple GEM foils or Micromegas mesh in a gas replicate the function of multiple dynodes in a conventional vacuum photomultiplier tube. In a reflective mode, a thin layer of CsI is deposited by vacuum evaporation on the upper surface of the first GEM in a cascade or on the Micromegas mesh, facing the UV-transparent window. The operation of MPGD-based photomultipliers in CF_4 with CsI PCs could form the basis of new-generation windowless Cherenkov detectors, where both the radiator and the photo-sensor operate in the same gas. Exploiting this principle, originally proposed for the Parallel Plate Avalanche Chamber (Giomataris and Charpak 1991), a Hadron Blind Detector was successfully operated using a triple-GEM amplification system with CsI PCs for the PHENIX experiment at RHIC (Fraenkel et al. 2005; Woody et al. 2009). A hadron-blindness property is achieved by reversing the direction of the drift field E_D, therefore pushing primary ionization produced by charged particles towards the mesh (see *Fig. 9a*). In this configuration photoelectrons released from the CsI PCs surface are still effectively collected into the GEM holes due to the strong electric field inside the holes. For many applications in photon detectors, a rather coarse (sub-mm) spatial resolution is usually sufficient. Therefore, THGEM and RETGEM are under study for RICH-detector upgrades for COMPASS and ALICE experiments (Di Mauro et al. 2009a).

The development of large-area position-sensitive photon detectors, with visible sensitive photocathodes (e.g., bialkali), could lead to numerous spin-offs, beyond the Cherenkov-light imagers. Most commonly used in the visible range are vacuum PhotoMultiplier Tubes (PMTs), with rather limited module size and bulky geometry due to mechanical constraints on the glass vacuum envelope. A possible alternative is to use gas-filled photomultipliers at atmospheric pressure; a proof of principle of visible-sensitive GPM was demonstrated with MPGDs (Chechik and Breskin 2008; Lyashenko et al. 2009).

Recent progress in the operation of cascaded GEM, Micromegas, THGEM, and RETGEM GPMs at cryogenic temperatures (down to 80 K) (Periale et al. 2005) and in two-phase mode (Lightfoot et al. 2005; Bondar et al. 2006, 2007, 2009; Periale et al. 2007) could pave the road toward their potential applications for the next-generation neutrino physics and proton decay experiments (Hagmann et al. 2004; Rubbia 2004), direct dark-matter searches (Rubbia 2005), positron emission tomography (PET) (Buzulutskov 2007), and noble-liquid Compton

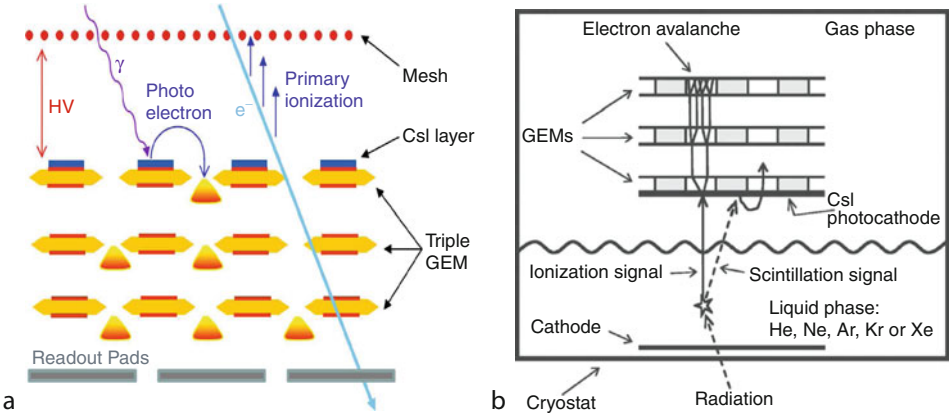


Fig. 9

(a) Basic principle of triple-GEM with CsI PCs in the Hadron Blind Detector (HBD) at RHIC (Woody et al. 2009); (b) Basic principle of the two-phase avalanche detector with a triple-GEM and reflective CsI PCs; both ionization and scintillation signals from the liquid are detected (Bondar et al. 2006)

telescope, combined with a micro-PET camera (Grignon et al. 2007; Duval et al. 2009). The operation principle of the cryogenic two-phase avalanche detector with GEM readout is demonstrated in *Fig. 9b*: the ionization produced in the noble liquid by radiation is extracted from the liquid into the gas phase by an electric field. The multi-GEM detector, operated at cryogenic temperature in saturated vapor above the liquid phase, can detect both the ionization signal, extracted from the liquid, and the scintillation signal, generated in the noble liquid by a particle (Bondar et al. 2007). The latter is achieved by depositing an UV-sensitive photocathode, namely CsI, on top of the first GEM (see *Fig. 9b*). The detection of both scintillation and ionization signals could allow efficient background rejection in rare-event experiments; in PET applications, the detection of scintillation signals could provide a fast trigger for coincidences between two collinear gamma quanta.

There were several attempts to use GEM detectors for medical physics and portal imaging. In particular, a GEM-based prototype was used to detect simultaneously the position of the therapeutic radiation beam (γ's) and the position of the patient tumor, using X rays, in order to provide feedback to the cancer treatment machine and to correct online the position of the beam with respect to the patient (Iacobaeus et al. 2001; Ostling et al. 2003). In this device, the X-ray image was obtained using conversions in gas and the γ-beam profile was determined using solid converters placed in between several GEM foils.

The avalanches formed within the GEM holes emit light with wavelengths in the visible and near-infrared regions (from 400 to 1,000 nm). This property has be exploited to read out the pattern of particle tracks or radiographic images by viewing a GEM with an optical output device such as CDD (Fetal et al. 2003). Another application of the GEM scintillation detector is radiation therapy, which demands new online beam-monitoring systems with ~1 mm position resolution and 3D dosimetry of delivered doses with an accuracy of ~5%.

There are many applications of the Micromegas in the neutron-detection domain, which include neutron-beam diagnostics (Pancin et al. 2004), inertial-fusion experiments

(Houry et al. 2006), thermal-neutron tomography (Andriamonje et al. 2004), and a sealed Piccolo-Micromegas detector, designed to provide in-core measurements of the neutron flux and energy (from thermal to several MeV) in the nuclear reactor (Andriamonje et al. 2006). Neutrons can be converted into charged particles to detect ionization in Micromegas by two means: either using the detector gas filling or a target with appropriate deposition on its entrance window. In such applications, the fast timing and good spatial resolution intrinsic to the Micromegas design can be exploited. Under the proper conditions (Pancin et al. 2007), these detectors can also show very low sensitivity to gamma rays, often an important part of the background in neutron measurements.

The MPGDs can also be used for a variety of security applications: detection of dangerous cargo, radon detection in the air as an early warning of earthquakes, and UV-sensitive early forest fire detection system. Muon tomography using large-area GEMs, based on the precise measurement of multiple scattering of cosmic ray muons traversing cargo of vehicles that contain high-Z materials, is a promising passive interrogation technique to detect hidden nuclear materials (Hohlmann et al. 2009). Among the planetary disasters, the most common and often happening are earthquakes and forest fires. A sharp increase in the Rn concentration before earthquakes has been observed in the air regions associated to rocks and caves, and its detection can serve as a basis of an early earthquake warning system. A network of RETGEM-based devices, capable to operate in the air, can be installed in the sensitive regions to provide daily assessment of the Rn concentration (Charpak et al. 2010). Similar RETGEM detectors could be used for the detection of forest fires at distances up to 1 km, compared with a range of 200 m for commercially available UV-flame devices (Charpak et al. 2003, 2009).

6 Development of Large-Area MPGDs

The current size of MPGDs is around $40 \times 40\,\text{cm}^2$, limited by existing tools and materials. A big step in the direction of the industrial manufacturing of large-size MPGDs with a unit size of a few square meters and spatial resolution typical of silicon micro-strip devices (30–50 μm) is the development and improvement of the fabrication technologies – single-mask GEM (Duarte Pinto et al. 2009a) and "Bulk" Micromegas (Giomataris et al. 2006).

Recent developments on large-area GEMs are focused on two new techniques to overcome the existing limitations: a single-mask technology and a splicing method for GEM foils. The standard technique for creating the GEM hole pattern, involving accurate alignment of two masks, is replaced by a single-mask technology to pattern only the top copper layer. The bottom copper layer is etched after the polyamide, using the holes in the polyamide as a mask. A single-mask technique overcomes the cumbersome practice of alignment of two masks between top and bottom films within 5–10 μm, which limits the achievable lateral size to ~50 cm. In a splicing procedure, foils are glued over a narrow seam, obtaining a larger foil. Both techniques were successfully implemented in the large-area prototype of $66 \times 66\,\text{cm}^2$ size, shown in *Fig. 10a* (Alfonsi et al. 2010).

The basic idea of the "Bulk" Micromegas technology is to build the whole detector in a single process: the anode plane with copper strips, a photo-imageable polyamid film, and the woven mesh are laminated together at a high temperature forming a single object. At the end, the micromesh is sandwiched between two layers of insulating material, which is removed after UV exposure and chemical development. The "Bulk" Micromegas technique has been

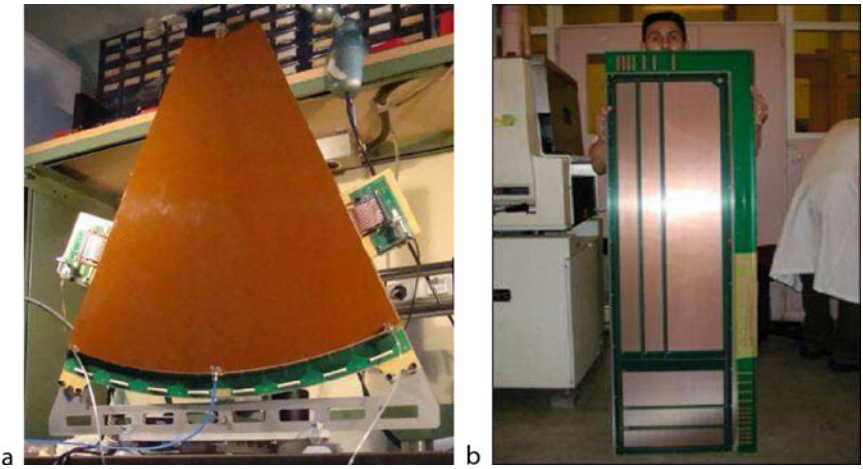


Fig. 10
(a) Triple-GEM prototype of 66 × 66 cm^2 active area, produced using single-mask technology, for the TOTEM experiment (Alfonsi et al. 2010); (b) Large-area "Bulk" Micromegas prototype of 40 × 130 cm^2 size for the ATLAS muon-system upgrade (De Oliveira private communications)

recently extended to produce large-area detectors, up to 40 × 130 cm^2 in a single piece (see *Fig. 10b*) (Alexopoulos et al. 2009, 2010). This industrial assembly process allows regular production of large, robust, and inexpensive detector modules.

7 Pixel Readout for Micro-Pattern Gas Detectors

Coupling of the microelectronics industry together with advanced PCB technology has been very important for the development of modern gas detectors with increasingly smaller pitch size. The fine granularity and high-rate capability of MPGDs can be further exploited by introducing a high-density pixel readout with a size corresponding to the intrinsic width of the detected avalanche charge. While the standard approach for the readout of MPGDs is a segmented strip or pad plane with front-end electronics attached through connectors from the back side, an attractive possibility is to use CMOS pixel chips, assembled directly below the GEM or Micromegas amplification structures (Costa et al. 2001; Bellazzini et al. 2004; Campbell et al. 2005; Bamberger et al. 2007b). These detectors use the input pads of a pixel chip as an integrated charge-collecting anode. With this arrangement signals are induced at the input gate of a charge-sensitive preamplifier (top metal layer of the CMOS chip). Every pixel is then directly connected to the amplification and digitization circuits, integrated in the underlying active layers of the CMOS technology, yielding timing and charge measurements as well as precise spatial information in 3D.

The combination of GEM detector and an analog, low-noise and high-granularity (50 μm pitch) CMOS pixel ASIC can bring large improvement in sensitivity, compared to traditional X-ray polarimeters (based on Bragg diffraction or Compton scattering) (Bellazzini et al. 2003;

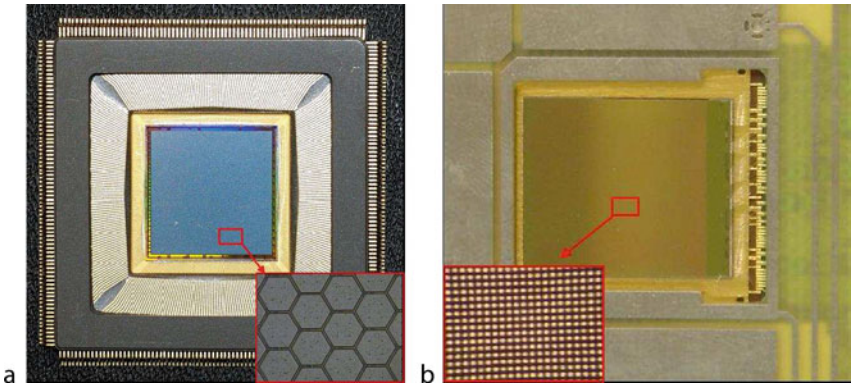


Fig. 11
(a) Photo of the analog CMOS ASIC with hexagonal pixels, bonded to the ceramic package (Bellazzini et al. 2004). (b) Photo of the Medipix2 chip (Llopart and Campbell 2002) – the 25 μm wide conductive bump bond openings, used for electron collection, are seen as a matrix of dots (*in the inset*) (Titov 2007)

Muleri et al. 2008). The scattering polarimeters are practically insensitive below 5 keV and are background limited, while Bragg crystal polarimeters are efficient only around a narrow band fulfilling the Bragg condition. In contrary, single GEM detectors coupled to a CMOS pixel chip can convert photons in the energy range from a few keV up to tens of keV, by choosing the appropriate gas mixture, and is able to simultaneously produce high-resolution images (50 μm), moderate spectroscopy (15% FWHM at 6 keV), and fast timing (30 ns) signals. Three ASIC generations of increased complexity and size, reduced pitch, and improved functionality have been designed and built (Bellazzini et al. 2004, 2006a, b, c) (see *Fig. 11a*). The high detector granularity allows to reconstruct individual low-energy (2–3 keV) photoelectron tracks, before they are distorted by Coulomb scattering, through the localization of the absorption point of the photon and then by estimation of the photoelectron emission direction. The degree of X-ray polarization is computed from the distribution of reconstructed track angles, since the photoelectron is emitted mainly in the direction of the photon's electric field. At the focal plane of the large-area mirror, like at the XEUS telescope, such a novel device will allow to perform energy-resolved polarimetry at the level of a few percent on many galactic and extragalactic sources with photon fluxes down to one milliCrab in one day (Bellazzini et al. 2006b, c).

The original motivation of combining a MPGD with Medipix2 (Llopart and Campbell 2002) and Timepix (Llopart et al. 2007) chips was the development of a new readout system for a large TPC at the future linear collider. The digital Medipix2 chip was originally designed for single-photon counting by means of a semiconductor X-ray sensor coupled to the chip. In gas detector applications, the chip is placed in the gas volume without any semiconductor sensor, with a GEM or Micromegas amplification structure above it (Colas et al. 2004; Campbell et al. 2005; Bamberger et al. 2007b). The avalanche electrons are collected on the metalized input pads, exposed to the gas (see *Fig. 11b*). The Timepix chip, which is a modification of the Medipix2 chip, is designed and manufactured in a six-metal 0.25 μm CMOS technology. The Timepix ASIC sensitive area is arranged as a square matrix of 256 × 256 pixels of 55 × 55 μm^2 size, resulting in a total detection area of ~14 × 14 mm^2. Each pixel in the chip matrix can be

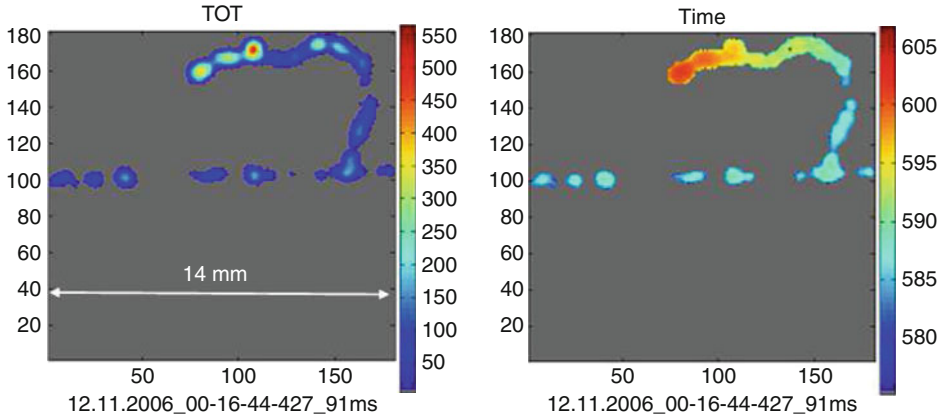


Fig. 12

Image of a 5 GeV electron track recorded with a triple-GEM detector and a TimePix chip operated in a mixed mode: the pixels are alternately operated in the "TOT" or "TIME" mode, resp., in a "chess board" fashion. The horizontal and vertical axes represent the chip sensitive area, obtained by mapping the original data (matrix of 256 × 256 pixels of 55 μm pitch) into a matrix of 181 × 181 pixels with a pitch of 78 μm (Titov 2007)

programmed to record either the arrival time of the avalanche charge signal with respect to an external shutter ("TIME" mode) or the 14-bit counter is incremented as long as the signal remains above the threshold (Time Over Threshold "TOT" mode), thus providing pulse-height information. The operation of the MPGD with a Timepix chip has demonstrated the possibility to reconstruct 3D-space points of individual primary electron clusters with ~30–50 μm spatial resolution and event-time resolution with nanosecond precision (Bamberger et al. 2007a, c; Brezina et al. 2009). This becomes indispensable for tracking and triggering and also for discriminating between ionizing tracks and photon conversions. Thanks to these developments, a micro-pattern device with CMOS readout can serve as a high-precision "electronic bubble chamber" (see *Fig. 12*).

An elegant solution for the construction of the Micromegas with pixel readout is the integration of the amplification grid and the CMOS chip by means of an advanced "wafer post-processing" technology (Chefdeville et al. 2006). The process uses standard photo-lithography and wet etching techniques and is CMOS compatible. It can be used to equip both single chips and chip wafers with a Micromegas grid. With this technique, the structure of a thin (1 μm) aluminum grid is fabricated on top of an array of insulating pillars, which stand ~50 μm above the CMOS chip. This novel concept is called "InGrid" (see *Fig. 13a*). The sub-μm precision of the grid dimensions and avalanche gap size results in a uniform gas gain. The grid hole size, pitch, and pattern can be easily adapted to match the geometry of any pixel readout chip. Among the most critical items that may affect the long-term operation of the "InGrid" concept is the appearance of destructive discharges across the 50 μm Micromegas amplification gap. One way to achieve protection is to cover the chip with a thin layer (few μm) of silicon nitride (Si_3N_4) deposited on top of the Timepix ASIC (Aarts et al. 2006; Bosma et al. 2007).

Due to its high sensitivity, "InGrid" detectors can resolve single primary electrons (Campbell et al. 2005; Blanco Carballo et al. 2007). The performance of the "InGrid"

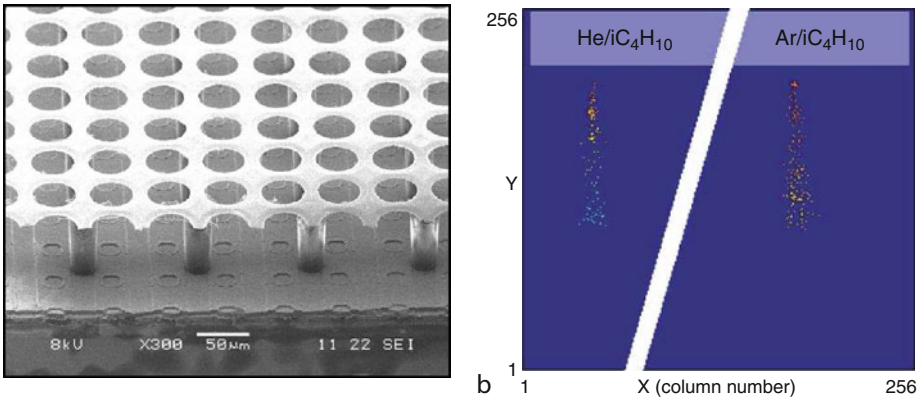


Fig. 13

(a) Photo of the Micromegas "InGrid" detector. The grid holes can be accurately aligned with readout pixels of CMOS chip. The insulating pillars are centered between the grid holes, thus avoiding dead regions (Melai et al. 2010) . (b) Images of 2 GeV electron tracks recorded with an "InGrid" detector in Ar/C_4H_{10} and He/C_4H_{10} mixtures in the "TIME" mode. The color is a measure of the arrival time of electrons (Bilevych et al. 2009)

device with a 4 µm layer of Si_3N_4 is illustrated in *Fig. 13b* for 2 GeV electron tracks. The color is a measure of the electron arrival time in "TIME" mode, red color corresponds to the primary ionization clusters produced close to the chip surface. The spread of electrons indicates that the diffusion increases with the distance from the chip. The time range in He/iC_4H_{10} mixture is larger than in Ar/iC_4H_{10}, which confirms a lower drift velocity in a He-based mixture, while the primary ionization density (number of active pixels) is higher in an Ar-based mixture (Bilevych et al. 2009). The "InGrid" device with a few mm drift gap is currently proposed for the upgrade of the ATLAS tracking system at the super-LHC (van der Graaf et al. 2009). Another potential application of GEM and Micromegas devices with CMOS pixel readout is the single-photon detection (Bellazzini et al. 2007).

A key element that has to be solved to allow CMOS pixel readout of MPGDs for applications in various fields of science is the production of large-area detectors. Present prototypes under construction rely at maximum on 2 × 4 CMOS integrated readout (De Lentdecker 2009; Kaminski 2009). Going to large surfaces requires a solution for the dead area on one side of the Timepix chip, where electrical connections enter the chip (see *Fig. 11b*). A first solution relies on etching of "through-silicon vias" that allows to bring signals and services from the back side of the chip using "through-wafer-vias" technology (Takahashi et al. 2003; Heijne 2005). This is known as a "via-last" operation: the existing Timepix wafers could be modified by etching "in-vias" connections after the wafer production step is completed. In order to move from three-side tileable detectors to 4-side ones, all common circuitry (bias, converters, . . .) needs to be spread over the chip area, which has nowadays become possible thanks to the progress in 3D micro-electronics. However, a major R&D effort is required in the future to fully exploit this potential.

8 Summary and Outlook

A century after the invention of the basic principle of gas amplification, gaseous detectors are still the first choice whenever the large-area coverage with low material budget is required. Advances in photo-lithography and microprocessing techniques during the past decade triggered a major transition in the field from wire chambers to micro-pattern gas-amplification devices. Today's MPGDs have opened a new era of state-of-the-art technologies and are the benchmarks for the gas detector developments beyond the LHC. They could eventually enable a plethora of new radiation detection concepts in fundamental science, medical imaging, and industry. Modern sensitive and low-noise electronics will enlarge the range of applications. Micro-pattern gaseous detectors with finely segmented CMOS readout can be used as a ultra-high-precision "electronic bubble chamber," opening a window to new physics for many applications, especially for low-energy charged-particle and photon detection.

The interest in the technological development and use of the novel micro-pattern gas detector technologies has led to the establishment of the research collaboration RD51 at CERN, with more than 75 participating institutes world wide (http://rd51-public.web.cern.ch/RD51-Public). It is a common platform for sharing of information, results, and experiences, and it supports efforts to make MPGDs suitable for large areas, increase cost efficiency, improve ease of use, and to develop portable detectors for industrial applications.

9 Cross-References

- Chapter 12, "Tracking Detectors"
- Chapter 13, "Photon Detectors"
- Chapter 19, "Muon Spectrometers"

References

Aarts AA et al (2006) Discharge protection and ageing of micromegas pixel detectors. In: IEEE NSS conference record

Adloff C et al (2009) JINST 4:P11023

Afanasev S et al (1999) Nucl Instrum Meth A 430:210

Aleksa M et al (2000) Nucl Instrum Meth A 446:435

Alexopoulos T et al (2009) JINST 4:P12015

Alexopoulos T et al (2010) Nucl Instrum Meth A 617:161

Alfonsi M et al (2004) IEEE Trans Nucl Sci (TNS) 51(5):2135

Alfonsi M et al (2010) Nucl Instrum Meth A 617:151

Andriamonje S et al (2004) IEEE NSS/MIC conference record

Andriamonje S et al (2006) Nucl Instrum Meth A 562:755

Andriamonje S et al (2010) JINST 5, P02001

Antonczyk D et al (2006) Nucl Instrum Meth A 565:551

Anvar S et al (2009) Nucl Instrum Meth A 602:415

Arogancia DC et al (2009) Nucl Instrum Meth A 602:403

Aune A et al (2009a) Nucl Instrum Meth A 604:53

Aune S et al (2009b) Nucl Instrum Meth A 604:15

Austregesilo A et al (2009) Nucl Phys Proc Suppl 197:113

Azevedo C et al (2010) JINST 5:P01002

Bachmann S et al (1999) Nucl Instrum Meth A 438:376

Bachmann S et al (2001) Nucl Instrum Meth A 470:548

Bachmann S et al (2002) Nucl Instrum Meth A 479:294

Bagaturia Y et al (2002) Nucl Instrum Meth A 490:223

Balla A et al (2009) Status of the cylindrical GEM project for the KLOE-2 inner tracker. In: IEEE NSS conference record, Orlando

Bamberger A et al (2007a) arXiv: 0709.2837
Bamberger A et al (2007b) Nucl Instrum Meth A 573:361
Bamberger A et al (2007c) Nucl Instrum Meth A 581:274
Barouch G et al (1999) Nucl Instrum Meth A 423:32
Barr A et al (1998) Nucl Phys B (Proc. Suppl.) 61B, 264
Bellazzini R et al (2003a) Proc SPIE, Polarimetry in Astronomy 4843:372, 394
Bellazzini R et al (2004) Nucl Instrum Meth A 535:477
Bellazzini R et al (2006a) Nucl Instrum Meth A 560:425
Bellazzini R et al (2006b) Nucl Instrum Meth A 566:552
Bellazzini R et al (2006c) Nucl Instrum Meth A 572:160
Bellazzini R et al (2007) Nucl Instrum Meth A 581:246
Bernet C et al (2005) Nucl Instrum Meth A 536:61
Biagi S (1999) Nucl Instrum Meth A 421:234
Bilevych Y et al (2009) The performance of GridPix detectors. In: IEEE NSS conference record
Blanco Carballo VM et al (2007) Nucl Instrum Meth A 583:42
Blum W, Riegler W, Rolandi L (2008) Particle detection with drift chambers. Springer, Berlin
Bondar A et al (2006) Nucl Instrum Meth A 556:273
Bondar A et al (2007) Nucl Instrum Meth A 581:241
Bondar A et al (2009) Nucl Instrum Meth A 598:121
Bosma M et al (2007) Results from MPGDs with a Protected Timepix or Medipix2 pixel sensor as active anode. In: IEEE NSS conference record
Bouclier R et al (1988) Nucl Instrum Meth A 265:78
Bouclier R et al (1996) Nucl Instrum Meth A 381:289
Breskin A et al (1975) Nucl Instrum Meth A 124:189
Breskin A et al (1978) Nucl Instrum Meth A 156:147
Breskin A et al (2003) Nucl Instrum Meth A 513:250
Breskin A et al (2009) Nucl Instrum Meth A 598:107
Bressan A et al (1999a) Nucl Instrum Meth A 424:321
Bressan A et al (1999b) Nucl Instrum Meth A 425:262
Brezina C et al (2009) JINST 4:P11015
Buzulutskov A (2007) Instr Exp Tech 50:287
Buzulutskov A (2008) Phys Part Nucl 39:424
Campbell M et al (2005) Nucl Instrum Meth A 540:295
Charpak G et al (1968) Nucl Instrum Meth A 62:262
Charpak G et al (1979) Nucl Instrum Meth A 167: 455
Charpak G et al (1998) Nucl Instrum Meth A 412:47
Charpak G et al (2002) Nucl Instrum Meth A 478:26
Charpak G et al (2003) IEEE Trans Nucl Sci (TNS) 55:1657
Charpak G et al (2009) JINST 4:P12007
Charpak G et al (2010) Performance of wire-type Rn detectors operated with gas gain in ambient air in view of its possible application to early earthquake predictions, arXiv: 1002.4732
Charpak G, Sauli F (1978) Phys Lett B 78:523
Charpak G, Sauli F (1979) Nucl Instrum Meth A 162:405
Charpak G, Sauli F (1984) Ann Rev Nucl Part. Sci 34:285
Chechik R et al (2003) Nucl Instrum Meth A 502: 195
Chechik R et al (2004) Nucl Instrum Meth A 535:303
Chechik R et al (2005) Nucl Instrum Meth A 553:35
Chechik R, Breskin A (2008) Nucl Instrum Meth A 595:116
Chefdeville M et al (2006) Nucl Instrum Meth A 556:490
Christophorou LG et al (1996) J Phys Chem Ref. Data 25(5):1341
Colas P et al (2004) Nucl Instrum Meth A 535:506
Costa E et al (2001) Nature 411:662
Dafni T et al (2008) MICROMEGAS for rare event searches. PoS IDM2008 106
Dafni T et al (2009) Nucl Instrum Meth A 608:259
De Lentdecker G (2009) A large TPC prototype for an ILC detector. In: IEEE NSS conference record
De Oliveira R private communications
Decamp D et al (1990) Nucl Instrum Meth A 294:121
Derre J et al (2000) Nucl Instrum Meth A 449:314
Derre J et al (2001) Nucl Instrum Meth A 459:523
Derre J, Giomataris I (2002) Nucl Instrum Meth A 477:23
Di Mauro A et al (2007) Nucl Instrum Meth A 581:225
Di Mauro A et al (2009a) IEEE Trans Nucl Sci (TNS) 56(3):1550
Di Mauro A et al (2009b) Nucl Instrum Meth A 610:169
Drumm H et al (1980) Nucl Instrum Meth A 176:333
Duarte Pinto S et al (2009a) JINST 4:P12009
Duarte Pinto S et al (2009b) Spherical GEMs for parallax-free detectors, arXiv:09113255
Duval S et al (2009) JINST 4:P12008
Fetal S et al (2003) Nucl Instrum Meth A 513:42
Fonte P et al (1999) IEEE Trans Nucl Sci (TNS) 46(3):321
Fraenkel Z et al (2005) Nucl Instrum Meth A 546:466
Garty G et al (1999) Nucl Instrum Meth A 433:476
Geiger H, Mueller W (1928) Phys Zeits 29:839
Geiger H, Rutherford E (1908) Proc Royal Soc A 81:141
Giomataris Y (1998) Nucl Instrum Meth A 419:239
Giomataris Y et al (1996) Nucl Instrum Meth A 376:29
Giomataris Y et al (2006) Nucl Instrum Meth A 560:405
Giomataris Y, Charpak G (1991) Nucl Instrum Meth A 310:585
Grignon C et al (2007) Nucl Instrum Meth A 571:142

Grupen C (1996) Particle detectors. Cambridge University Press
Hagmann C et al (2004) IEEE Trans Nucl Sci (TNS) 51(5):2151
Heijne E (2005) Nucl Instrum Meth A 541:274
Hohlmann M et al (2009) Design and construction of a first prototype muon tomography system with GEM detectors for the detection of nuclear contraband, arXiv: 0911.3203
Houry M et al (2006) Nucl Instrum Meth A 557:648
http://consult.cern.ch/writeup/magboltz
http://gdd.web.cern.ch/GDD/
http://hallaweb.jlab.org/12GeV/SuperBigBite
http://rd51-public.web.cern.ch/RD51-Public
http://rjd.web.cern.ch/rjd/cgi-bin/cross
http://www.ansoft.com
Iacobaeus C et al (2001) IEEE Trans Nucl Sci (TNS) 48:1496
Iacobaeus C et al (2002) IEEE Trans Nucl Sci (TNS) 49(4):1622
Ivaniouchenkov Y et al (1998) IEEE Trans Nucl Sci 45(3):258
Ivaniouchenkov Y et al (1999) Nucl Instrum Meth A 422:300
Kaminski J (2009) Talk at the 4th RD51 collaboration meeting, 23–25 November 2009
Ketzer B et al (2004) Nucl Instrum Meth A 535:314
Killenberg M et al (2003) Nucl Instrum Meth A 498:369
Lami S (2009) First test results for the TOTEM T2 telescope. In: IEEE NSS conference record
Lightfoot PK et al (2005) Nucl Instrum Meth A 554:266
Llopart X et al (2007) Nucl Instrum Meth A 581:485
Llopart X, Campbell M (2002) IEEE Trans Nucl Sci (TNS) 49(5):2279
Lyashenko A et al (2004) Nucl Instrum Meth A 523:334
Lyashenko A et al (2006) JINST 1, P10004
Lyashenko A et al (2007) JINST 2, P08004
Lyashenko A et al (2009) JINST 4:P07005
Matsuda T (2010) JINST 5, P01010
McDaniel E, Mason E (1973) Mobility and diffusion of ions in gases. Wiley, New York
Melai J et al (2010) arXiv: 1003.2083
Meyer T (2003) System aspects of (gaseous) tracking detectors. In: Proceedings of the 42nd workshop of the INFN ELOISATRON project on Innovative Detectors For Supercolliders, Erice, Italy, 28 Sept 2003–4 Oct 2003
Miyamoto J, Knoll G (1997) Nucl Instrum Meth A 399:85
Mormann D et al (2004) Nucl Instrum Meth A 530:258
Muleri F et al (2008) Nucl Instrum Meth A 584:149
Neyret D et al (2009) JINST 4:P12004
Nygren D et al (1976) PEP-PROPOSAL-004, Appendix A6
Nygren DR, Marx JN (1978) Physics Today 31, Vol. 10
Oed A (1988) Nucl Instrum Meth A 263:351
Oliveira R et al (2007) Nucl Instrum Meth A 576:362
Ostling J et al (2003) IEEE Trans Nucl Sci (TNS) 50:809
Pacella D et al (2003) Nucl Instrum Meth A 508:414
Pacella D et al (2006) JINST 1:P09001
Pancin J et al (2004) Nucl Instrum Meth A 524:102
Pancin J et al (2007) Nucl Instrum Meth A 572:859
Peisert A, Sauli F (1984) Drift and diffusion of electrons in gases, CERN 84-08
Periale L et al (2002) Nucl Instrum Meth A 478:377
Periale L et al (2003) IEEE Trans Nucl Sci (TNS) 50(4):809
Periale L et al (2005) IEEE Trans Nucl Sci (TNS) 52(4):927
Periale L et al (2007) Nucl Instrum Meth A 573:302
Peskov V et al (1997) Nucl Instrum Meth A 397:243
Peskov V private communication
Peskov V, Fonte P (2009) Research on discharges in micropattern and small gap detectors, arXiv: 0911.0463
Policarpao A et al (1972) Nucl Instrum Meth A 102:337
Procureur S private communication
Rubbia A (2004) Experiments for CP violation: a giant liquid argon scintillation, Cerenkov and charge imaging experiment, arXiv: hep-ph/0402110
Rubbia A (2005) ArDM: a ton-scale liquid argon experiment for direct detection of dark matter in the universe, arXiv: hep-ph/0510320
Sacquin Y et al (1992) Nucl Instrum Meth A 323:209
Santonico R, Cardarelli R (1981) Nucl Instrum Meth A 187:377
Sauli F (1977) Principles of Operation of Multiwire Proportional and Drift Chambers, CERN 77-09
Sauli F (1994) Potentials of advanced gas detectors for health physics. CERN-PPE/94-195
Sauli F (1997) Nucl Instrum Meth A 386:531
Sauli F (2004) From bubble chambers to electronic systems: 25 years of evolution in particle detectors at CERN, CERN PH-EP 2004-040
Sauli F (2007) Nucl Instrum Meth A 580:971
Sauli F, Sharma A (1999) Ann Rev Nucl Part Sci 49:341
Sauli F, Titov M (2010) in Review of Particle Physics (Particle Data Group), chapters 28.6.1–28.6.4, Gaseous Detectors, J Phys G 37 075021:308
Seguinot J, Ypsilantis T (1977) Nucl Instrum Meth A 142:377
Shalem C et al (2006) Nucl Instrum Meth A 558:475
Shultz G et al (1977) Rev Phys Appl 12:67
Simon F et al (2009) The forward GEM tracker of STAR at RHIC. In: IEEE NSS conference record

Smith G (2006) Gas-based detectors for synchrotron radiation. J Synchrotron Radiat 13(2):172

Takahashi K et al (2003) Microelectron Realib 43:1267

Titov M (2003) Radiation damage and long-term aging in gas detectors, arXiv: physics/0403055; also in Proceedings of the 42nd workshop of the INFN ELOISATRON project on Innovative Detectors for Supercolliders, Erice, Italy, 28 Sept 2003–4 Oct 2003

Titov M (2007) Nucl Instrum Meth A 581:25

Titov M et al (2002) IEEE Trans Nucl Sci (TNS) 49(4):1609

van der Graaf H et al (2009) Performance and prospects of GridPix and gossip detectors, RD51 Note, RD51-2009-006

Vandenbroucke M (2009) A GEM-based TPC prototype for PANDA. In: IEEE NSS conference record

Veloso JFCA et al (2000) Rev Sci Instr 71(6):2371

Villa M et al (2010) arXiv: 1007.1131

Walenta AH (1979) IEEE Trans Nucl Sci (TNS) 26:73

Walenta AH et al (1971) Nucl Instrum Meth A 92:373

Wieman H et al (1997) IEEE Trans Nucl Sci (TNS) 44(3):671

Woody C et al (2009) Initial performance of the PHENIX hadron blind detector at RHIC. In: IEEE NSS conference record

Yu J (2009) Development of GEM based digital hadron calorimeter. In: IEEE NSS conference record

12 Tracking Detectors

Manfred Krammer · Winfried Mitaroff
Austrian Academy of Sciences, Vienna, Austria

1 Introduction 267

2 Gas Detectors 268
2.1 Wire Chambers 268
2.2 Micro Pattern Gas Detectors 269
2.3 Time Projection Chambers 270

3 Silicon Detectors 270
3.1 Strip Detectors 271
3.2 Hybrid Pixel Detectors 271

4 Other Tracking Detectors 273

5 Integration in Experiments 274
5.1 Fixed-Target Experiments 274
5.2 Collider Experiments 275
5.3 Detector Alignment 279

6 Event Reconstruction 279
6.1 Pattern Recognition 280
6.2 Track Reconstruction 281
6.3 Vertex Reconstruction 282

7 Performance Optimization 284

8 Summary 287

9 Appendix: Formulae 287
9.1 Kalman Track Fitting 287
9.2 Kalman Vertex Fitting 289
9.3 Robust Vertex Fitting 291
9.4 Helix Coordinate Systems 292

Acknowledgment 293

References 293

C. Grupen, I. Buvat (eds.), *Handbook of Particle Detection and Imaging*,
DOI 10.1007/978-3-642-13271-1_12, © Springer-Verlag Berlin Heidelberg 2012

Further Reading 295

Software Portals (Selection) 295

Abstract: Tracking detectors are devices to measure and reconstruct the trajectories of charged particles. They are developed for and used by physics experiments in the fields of nuclear physics, particle physics, and astro-particle physics. To understand and analyze the physics processes under study at these experiments the reconstruction and precise determination of the particles flight path is important. From these particle tracks, parameters such as the particle momentum, the particle type, its origin, etc., can be deduced. Several detector technologies have been invented and are being constantly improved. The most important ones are the various types of gas detectors, detectors based on semiconductor material, and scintillating detectors. In a realistic experiment, several tracking (and other) devices are arranged to a complex set-up. Charged particle tracks are reconstructed by making use of all available information from all tracking detectors. The actual reconstruction of events from raw measurements is a nontrivial task involving pattern recognition, track and vertex fitting. The performance of both hardware and software must be optimized for the benefit of follow-on physics analyses.

1 Introduction

Experiments for nuclear physics, elementary particle physics, and astro-particle physics investigate the properties of subatomic particles and the laws of nature responsible for the interaction among them. These sciences need detectors not only to detect the presence of particles but also to measure the various particle parameters. Dedicated detectors are used to measure the interaction cross sections, the energy of particles, the type of particles, etc. For many experiments the reconstruction of the trajectory of particles is essential. From the trajectories, the origin of the particle, the direction of the flight, and even more sophisticated parameters such as the particle's momentum or the sign of the particle charge can be deduced if the detector is placed inside a magnetic field. Such tracking detectors are usually constructed using position-sensitive detector elements which make use of the signals particles induce in the detectors when interacting with the detector material. In tracking detectors, the individual detector elements are thin and absorb only a fraction of the particle energy in order to make several position measurements possible on the same particle. The individual measurements (space points) are then used to reconstruct the particle's trajectory. This chapter is restricted to the description of tracking detectors for charged particles. Neutral particles also undergo interaction with the detector material and the position of the interaction can be measured. However, the interactions of photons usually lead to the absorption of these and hence a tracking cannot be performed. Other neutral particles, such as neutrons or neutrinos, produce in the interaction with matter charged particles and, consequently, a tracking of these particles is again performed by charged particle detectors.

To measure the position of charged particles several technologies have been developed, the most important ones being gaseous detectors described in Sect. 2, and semiconductor detectors described in Sect. 3. Their integration into larger experimental set-ups is covered in Sect. 5, with emphasis on modern colliding beam experiments.

Event reconstruction, i.e., calculation of the parameters needed for physics analyses from the individually measured space points, requires sophisticated statistical methods; the most

important ones are described in Sect. 6, followed by a discussion on detector performance and optimization. A summary of useful formulae is given in Sect. 9.

2 Gas Detectors

Detectors based on the ionizing effect of charged particles in a gaseous medium are in use since almost one century, e.g., Geiger–Müller counters. Very early after the discovery of radiation, simple tubes with a strung wire have been used to detect radiation and to measure the rate and intensity (Ortner and Stetter 1928). The principle is very simple – between the wire acting as anode and the metal tube acting as cathode, a high voltage is applied. The field strength within the tube is

$$|E(r)| = \frac{1}{r} \cdot \frac{V}{\ln(b/a)}, \tag{1}$$

where V is the applied voltage and a, b are the radii of the wire and outer tube, respectively. Close to the thin wires (thickness in the order of 100 μm) the field becomes very high. Charged particles ionize the gas along their path through the detector, and the liberated electrons and ions move to the electrodes due to the electric field. The electrons approaching the anode wire are accelerated in the increasing field strength eventually reaching a velocity sufficient for secondary ionization. The resulting avalanche of secondary electrons and ions amplify the signal by many orders of magnitude. Gas detectors have therefore a built-in amplification mechanism. Based on this principle a multitude of different gas detectors has been developed.

2.1 Wire Chambers

A gas detector with a single wire can detect the presence of radiation. By operating several counters next to each other also the position of charged particles can be deduced. The tube surrounding each wire can be omitted resulting in a structure called multi-wire proportional chamber (Charpak et al. 1968). The resulting field geometry is shown in *Fig. 1*. Each wire is connected to the input of an electronic amplifier. An electrical signal indicates the passage of a charged particle in a particular position. The obtained position resolution is in the order of 300 μm assuming a typical distance of 1 mm between adjacent wires (variance of a uniform distribution).

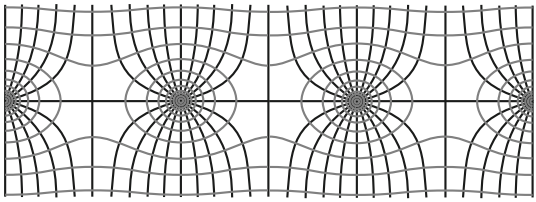


Fig. 1
Field geometry of a multi-wire proportional chamber

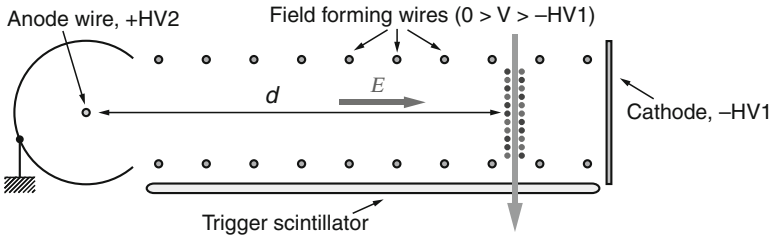


Fig. 2
Principle of a gas drift chamber

Further developments of gas detectors lead to the construction of drift chambers. The principle is illustrated in *Fig. 2*. The primary ionization electrons and ions produced by a charged particle drift to an anode wire (electrons) and to a cathode (ions). The electrons create secondary ionization in the field of the wire. The precise position of the throughgoing particle x is now calculated from the time the electrons need to drift to the anode. x is given by

$$x = \int v_e(t)\,\mathrm{d}t \tag{2}$$

with v_e, the drift velocity of the electrons. To reach high precision, v_e along the drift path has to be known and monitored precisely. The scintillator detector shown in *Fig. 2* delivers the start signal for the time measurement. A large advantage of this construction is the reduction of electronics channels for large-area chambers with respect to multi-wire proportional chambers. The position resolution is limited by the precision of the wire placement, properties of the gas such as diffusion, and the time resolution of the electronics. A long drift time causes also a severe limitation on the rate capability of drift chambers. The achievable spatial resolution of wire-based gas detectors is in the order of 50–100 μm (Riegler et al. 2000). For examples of tracking systems based on drift chambers see Adorisio et al. (2009), Qin et al. (2007), and Adinolfi et al. (2002).

2.2 Micro Pattern Gas Detectors

To overcome some of the limitations mentioned above micro pattern gas detectors were developed. In this large class of gaseous detectors the classical wires are replaced by metal strip structures deposited on high-resistivity substrates. The first detector of this kind was the micro-strip gas chamber (MSGC) (Oed 1988). The strips are defined by photolithography and can therefore be separated within distances much smaller than wires (typically 70 μm). *Figure 3* shows the field lines and the typical dimensions of an MSGC cell. From the vast variety of different micro pattern gas detectors (see Sauli and Sharma 1999; Titov 2007 for an overview), the most used structures are the GEM (Gas Electron Multiplier) and the Micromegas detectors. In a GEM detector (Sauli 1997), the amplification of the primary ionization takes place in the high-field region of holes in a foil metalized on both surfaces and placed on top of an MSGC. In a Micromegas detector (Giomataris et al. 1996), the drift region is separated by a micromesh in an ionization region and a thin region above the MSGC structure in which the amplification takes place.

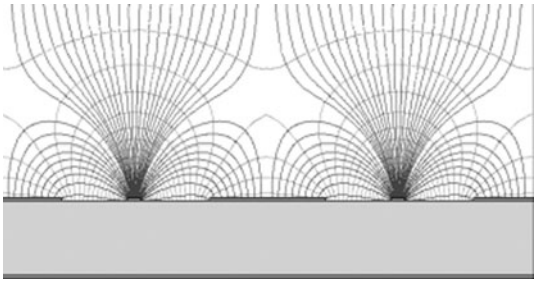


Fig. 3
Field geometry of a microstrip gas chamber

The position resolution of micro pattern gas detectors can be as excellent as a few tenth μm (Derré et al. 2001; Zerguerras et al. 2007). Due to the small sizes and the fast collection of positive ions very high rate capability of MHz/mm^2 is achievable (Ivaniouchenkov et al. 1999; Alfonsi et al. 2004). Examples for large tracking detector systems based on micro pattern gas detectors are described in Altunbas et al. (2002), Ketzer (2002), and Bagliesi et al. (2010).

2.3 Time Projection Chambers

Time projection chambers (TPCs) are large gas-filled volumes with readout structures on the end plates only (see *Fig. 4*). Between the end plates and a central electrode (the cathode) a large electric field (100–400 V/cm) is applied. Charged particles traversing the gas volume ionize the gas. The electric field forces the electrons to drift to the end plates where the position and the time of arrival is measured. From these measurements the three-dimensional track of the particle is reconstructed. In collider experiments such time projection chambers are used in connection with a magnetic field parallel to the electric field. The throughgoing particle follows a curved track then. The position detectors on the end plates are usually wire chambers. In very large TPCs (up to a length of 5 m and a radius of 4 m), a space point resolution of 1 mm to few 100 μm is achievable (Larsen 2010; Anderson 2003). Future TPCs will use micro pattern gas detectors such as GEM or Micromegas as end plate detectors because of their superior performance (Colas 2004; Kobayashi 2007).

3 Silicon Detectors

Silicon detectors as position detectors were first developed in the early 1980s. They are produced from thin wafers (approximately 300 μm thick) of high-resistivity silicon. For most applications n-type silicon is selected. With photolithographic methods fine-patterned p^+-doped implantations are introduced on one side of the wafer forming pn junctions. Each implant is connected to an individual channel of the readout electronics. With respect to the pn junction a reverse bias voltage is applied. This bias voltage has to be large enough to fully deplete the silicon bulk.

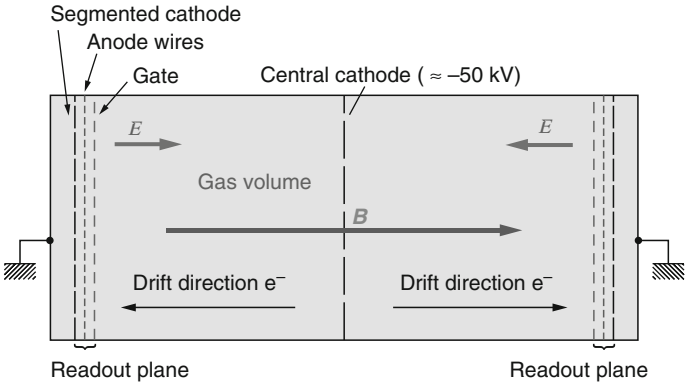


Fig. 4
Geometry of a time projection chamber

Charged particles traversing the detector create electron–hole pairs along their flight path. Due to the applied electrical field these electron–hole pairs are separated and drift toward the electrodes, thus inducing signals in the closest readout channels (Hartmann 2009).

3.1 Strip Detectors

In the case of the most frequently employed silicon strip detector the implants are thin strips (typically 20 μm wide with an interstrip distance of 100 μm). A sketch of a silicon strip detector is shown in *Fig. 5*. This cut through a detector shows the p^+-implants on top, covered by a layer of aluminum. The input of an amplifier is connected to the aluminum pad through an AC coupling capacitor. In so-called AC coupled silicon sensors this capacitor is integrated into the detector with an additional layer of deposited isolator (SiO_2 or Si_3N_4).

With a proper choice for the strip geometry a precision of a few μm can be easily achieved. By placing several silicon detectors in the path of the charged particles the trajectories of these particles can be reconstructed. The first experiment using a silicon strip detector was the experiment NA11 at CERN in 1983 (Hyams et al. 1983). Following a rapid development, large silicon tracking detectors of up to 200 m^2 of silicon area have been built (Chatrchyan et al. 2008; Adam et al. 2009a).

3.2 Hybrid Pixel Detectors

In parallel to more and more sophisticated strip detectors hybrid pixel detectors were developed. The implants in these detectors are small pixels, e.g., 300 × 300 μm^2. The difficulty of such devices is the connection of the large number of pixels to the individual electronic channels. Each pixel of the sensor realized on high-resistivity silicon material has to be electrically connected to the corresponding input channel of the electronics chip. In hybrid pixel detectors this connection is done by so-called bump bonding. A schematic drawing of a cell of

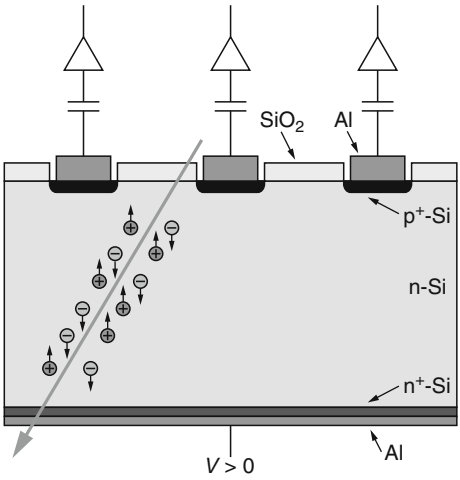


Fig. 5
Schematics showing a single-sided silicon strip detector

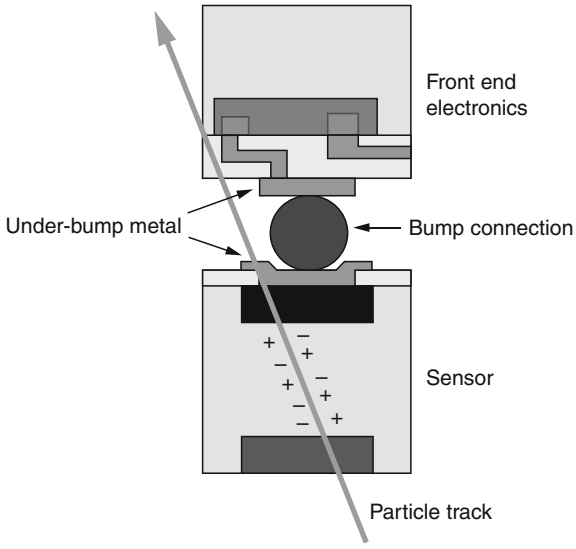


Fig. 6
Schematics showing a cell of a hybrid pixel detector

a hybrid pixel detector is shown in *Fig. 6*. In this drawing the silicon pixel sensor is at the bottom. Small conductive bump balls (e.g., Cu or In) connect the pixels to the input pads of the electronics chip at the top. The layout of the electronics chip has to match the pattern of the pixel sensor. Examples for large hybrid pixel detector systems are Dominguez (2007) and Klingenberg (2007). The achievable position resolution varies if the signal height measured on a pixel is processed and used to interpolate between the pixels. In such a configuration a position

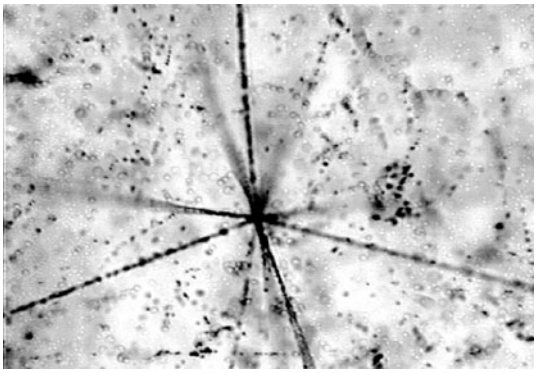


Fig. 7
Micro photograph of tracks in a nuclear emulsion (Eskut et al. 1997); the grain size is about 1 μm

resolution of a few μm is possible. The strength of hybrid pixel detectors is their capability to operate in a high track density environment.

4 Other Tracking Detectors

Apart from gaseous and semiconductor detectors other technologies are used for tracking charged particles. These are nuclear emulsions for very-high-precision applications, detectors based on cryogenic liquid noble gases, and detectors using scintillation phenomena for particle detection. A short description of these technologies follows.

Nuclear emulsions are a special kind of photographic films. The materials used are typically micro crystals (size 0.1–1 μm) of silver halides (mostly AgBr) embedded in gelatine. Ionizing particles produce a latent perturbation along the tracks, which becomes visible as tracks of silver grains after development and fixation. The development agent reduces the AgBr to metallic Ag, while the fixation agent removes the remaining silver halides. Under a microscope the silver grains indicating the tracks of the particles are finally visible. This technology was first used in 1911 for the tracking of alpha particles (Reingamum 1911; Beiser 1952). It was further developed in the early days of particle physics by Blau (1961/1963) and is nowadays exploited on a large scale in present experiments (Arrabito et al. 2007). Emulsions are still the tracking detectors that enable the most precise reconstruction resolution in the sub-micrometer range. An example of tracks in a nuclear emulsion is shown in *Fig. 7*.

The principle of a time projection chamber (TPC) as explained in Sect. 2.3 for gas detectors can also be used with liquids. Especially in noble liquids, such as in cryogenic liquid Ar or Xe, the lifetime of electrons produced by the ionization of charged particles can be large (about 2 ms in LAr (Arneodo et al. 2003)) and as a consequence the drift length of these electrons within an electric field can be large enough to allow the construction of detectors with dimensions of more than one meter. To achieve these values the liquid has to be of extreme purity. The drift time of the electrons and the position is measured by wire chambers embedded in the liquid.

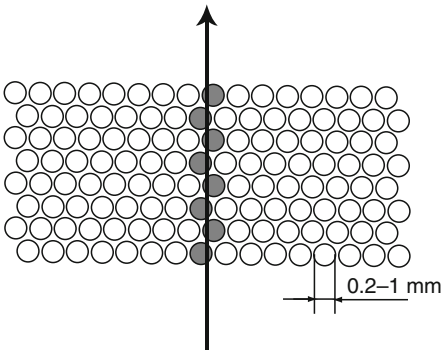


Fig. 8
Principle of a scintillating fiber detector

In various materials charged particles excite atomic or molecular levels, which during the subsequent decay to a lower state emit scintillation light. In scintillating fiber detectors, thin fibers (core thickness less than 1 mm) made out of scintillating material are bundled and read out by photosensitive detectors. The principle of a scintillating fiber detector is illustrated in *Fig. 8*. Photodetectors used are multi-anode photomultipliers (Achenbach et al. 2008), CCD cameras (Kim et al. 2003), or silicon Avalanche Photo Diodes (APDs) (De Gerone et al. 2009). A position resolution of a few 100 μm and below is achievable (Achenbach et al. 2008; Beischer et al. 2010).

5 Integration in Experiments

The tracking detectors described in Sects. 2–4 are most often not used stand-alone, but integrated into larger assemblies forming an experimental set-up. In particular, experiments at high-energy particle accelerators are big and complex, composed of track-sensitive detectors, calorimeters, and other active and passive components (e.g., a superconducting magnet). Two basic set-ups exist, for *fixed-target* and *colliding beam* experiments. Here and in the following the units [mass] = GeV/c^2, [momentum] = GeV/c, and [energy] = GeV are understood with $c = 1$ (dimensionless).

5.1 Fixed-Target Experiments

Accelerated beam particles (momentum P_B) hit the nucleons (mass m_T) of a solid or liquid target at rest. The collision energy in the center-of-mass frame and the forward boost factor η_{cm} are, for a relativistic beam,

$$E_{\mathrm{cm}} = \sqrt{2 m_T P_B}\,, \qquad \eta_{\mathrm{cm}} = P_B / E_{\mathrm{cm}}\,.$$

The particles, either directly accelerated (p, e^-, ions) or produced on a target upstream, arrive in "spills" synchronized with the accelerator cycle. A fast multi-level *trigger* selects reactions of

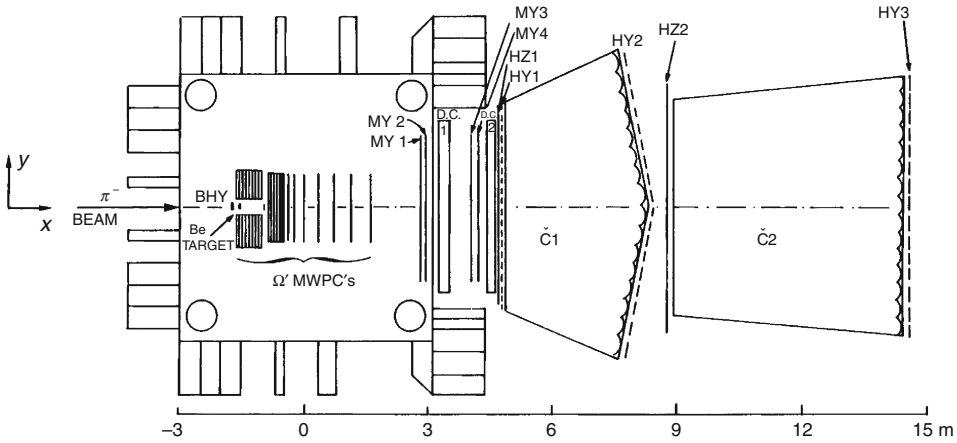


Fig. 9
The Ω′ detector at the CERN-SPS, experiment WA77 (top view)

interest. The total number of events collected per unit cross section is called *sensitivity* of an experiment.

The experimental set-ups differ considerably among the experiments. Early single- or double-arm magnet spectrometers have been replaced by general-purpose ones with 4π acceptance. A typical example is shown in *Fig. 9*: the central tracking and recoil part is a flexible arrangement of multi-wire proportional chambers inside a large superconducting 1.8 T dipole magnet; forward track resolution is improved by two lever-arm drift chambers; two threshold Cherenkov counters discriminate forward $\pi^{\pm} \leftrightarrow K^{\pm} \leftrightarrow p/\bar{p}$; trigger detectors select high-p_T reactions of incident 350 GeV/c π^- (Jacob and Quercigh 1997).

5.2 Collider Experiments

Two oppositely accelerated beam particles (energies E_1, E_2) collide with each other. The center-of-mass (CM) energy and the boost factor η_{cm} are, for relativistic beams and head-on collisions at zero crossing angle,

$$E_{\mathrm{cm}} = 2 \cdot \sqrt{E_1 E_2}\,, \qquad \eta_{\mathrm{cm}} = |E_1 - E_2|/E_{\mathrm{cm}}\,.$$

If $E_1 = E_2$ (*symmetric collider*), laboratory and CM frame coincide.

There are *hadron colliders* (pp, $p\bar{p}$, ion–ion), *lepton colliders* (e^-e^+, e^-e^-), and *hybrid colliders* (e^-p). The beams are structured into "bunches" or the like, run in vacuum tube(s), and are focused by quadrupole magnets to collide inside the detector. The average collision rate per unit cross section is called *luminosity* ($\mathcal{L}$). Colliders integrate their accelerators, and are either *circular* with stored beams, or *linear* with "one shot" collisions. A circular option is not feasible for $e^{\pm}$ beams above some 100 GeV because of synchrotron radiation $\propto (E/m)^4$.

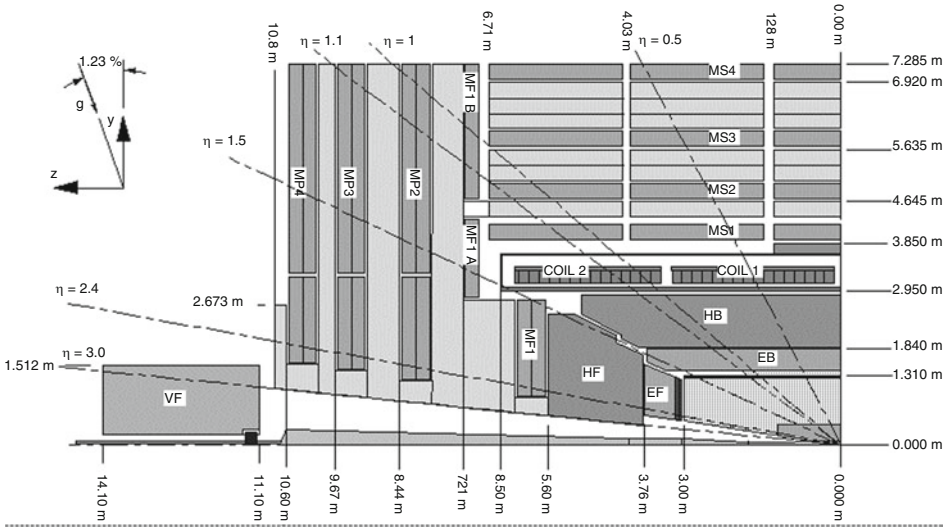


Fig. 10
The CMS detector at CERN's LHC (longitudinal quadrant)

Ideas exist for circular $\mu^-\mu^+$ colliders at TeV energies; problems are the $\mu^\pm$ lifetime, and decay ν background.

Hadron Collider Experiments

They are characterized by huge data rates dominated by soft QCD background. Hard processes between the constituents (quarks, gluons) are rare and kinematically undefined. The low signal-to-background ratios enforce sophisticated triggering. Flagship is the "Large Hadron Collider" (LHC) at CERN, designed for (7 + 7) TeV pp collisions at $\mathcal{L} = 10^{34}$ cm^{-2} s^{-1}. (Evans and Bryant 2008).

One of the detectors is the "Compact Muon Solenoid" (CMS), shown in *Fig. 10*. Its overall layout is cylinder symmetric, with a large all-silicon tracker consisting of pixels (in the very inside) and microstrips. Both the pixel and microstrip detectors are arranged on cylinders in the barrel direction, and on disks in the forward and backward directions. Still inside the superconducting 4 T solenoid magnet are the electromagnetic and hadron calorimeters. Outside the magnet is the muon detection system, which is made of various gas detectors interspersed with the magnet yoke (Chatrchyan et al. 2008).

Lepton Collider Experiments

They provide a clean environment: collisions exploit the full center-of-mass energy and are kinematically well defined; beam energies are tunable, and the beams may be polarized; data rates are moderate and backgrounds can be rejected by a simple interaction trigger (at higher energies, pair production from beamstrahlung becomes significant), thus allowing for unbiased data collection. These characteristics qualify for precision studies.

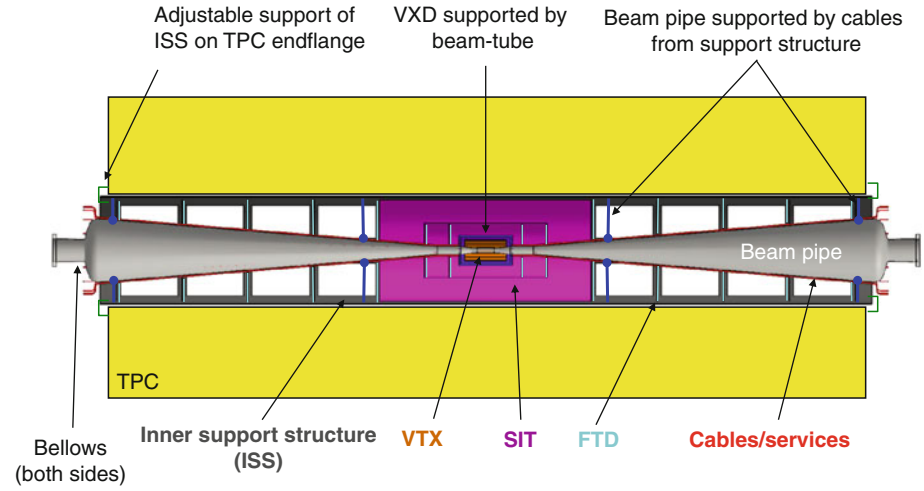


Fig. 11
The ILD detector at the ILC (longitudinal view of inner part)

The circular e^-e^+ collider of highest energy was LEP at CERN (circumference 26.7 km, maximal beam energies (104 + 104) GeV) (Hübner 2004).

The most advanced next generation project is the "International Linear Collider" (ILC), based on superconducting RF cavities for center-of-mass energies in the range 0.2 . . . 1 TeV, and peak luminosity $\mathcal{L} = 2 \times 10^{34}$ cm^{-2} s^{-1}. Options, besides operating e^-e^+, are e^-e^-, $e^-\gamma$, and $\gamma\gamma$ (Phinney et al. 2007).

Two detectors will share one interaction region ("push-pull"). One of them is the "International Large Detector" (ILD), partly shown in *Fig. 11*. The cylinder-symmetric overall layout contains both a silicon and a gaseous tracking system: in the barrel direction the multi-layer pixel vertex detector (VTX) and microstrip inner tracker (SIT), and in the forward and backward directions a stack of pixel and microstrip disks (FTD); the central tracking device is a large TPC with ≈200 pad-rows, enveloped by a layer of microstrips (aiding track precision and pattern recognition). Still inside the superconducting 3.5 T solenoid magnet are the electromagnetic and hadron calorimeters. The muon detectors are integrated into the magnet yoke. The beam-tube has conical shape in order to minimize the e^-e^+-pair background created by beamstrahlung in the forward and backward directions (Stoeck et al. 2010).

Experiments at e^-e^+ Factories

So-called factories are high-luminosity e^-e^+ colliders, running at a center-of-mass energy equivalent to the mass of a vector particle resonance, for the purpose of studying its copiously produced decay products. Examples are DAFNE at Frascati (Φ factory), CESR-C at Cornell (tau–charm factory), PEP II/BaBar at SLAC and KEKB/Belle at KEK (beauty factories), and LEP I at CERN and the "Giga Z" option of ILC (Z^0 factories).

The Beauty Factory KEKB at KEK is an asymmetric e^-e^+ collider with beams around (8 + 3.5) GeV, and tuned to the $b\bar{b}$ resonances $\Upsilon_{4S}(10.58) \rightarrow B^0\bar{B}^0, B^+B^-$ and

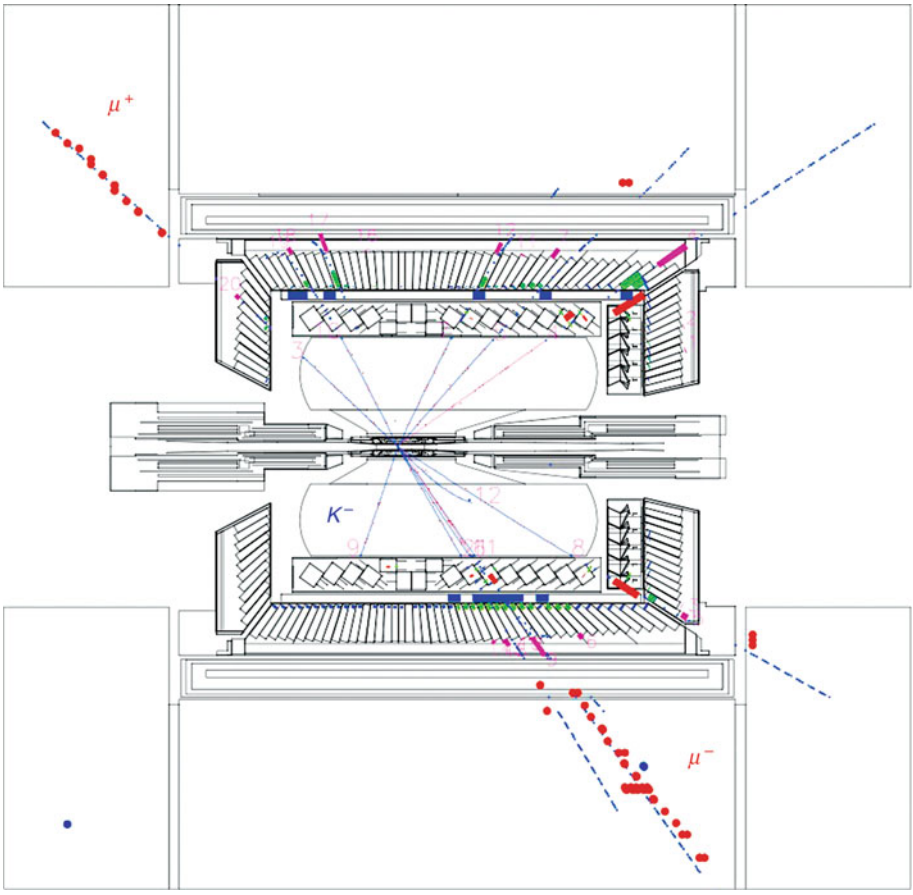


Fig. 12
The Belle detector at KEKB (longitudinal section)

$\Upsilon_{5S}(10.86) \to B_s^{0(*)}\bar{B}_s^{0(*)}, \ldots$. The boost factor $\eta_{\rm cm} = 0.425$ allows spatial resolution of B-meson decay vertices in the laboratory frame. The peak luminosity achieved is $\mathcal{L} = 2.1 \times 10^{34}$ cm^{-2} s^{-1} (Kurokawa and Kikutani 2003).

The Belle detector is shown in *Fig. 12*. The overall layout is cylinder symmetric but not mirror symmetric, thus coping for the forward boost. The superconducting 1.5 T solenoid magnet contains the silicon vertex detector (SVD) and the gaseous central tracker – a large drift chamber with stereo wires (CDC) – surrounded by aerogel Cherenkov and time-of-flight counters for particle identification, and an electromagnetic calorimeter. A dedicated $K_{\rm L}^0$ and muon detector is combined with the magnet yoke (Abashian et al. 2002).

An upgrade is underway, aiming to increase the luminosity of the collider to $\mathcal{L} \approx 8 \times 10^{35}$ cm^{-2} s^{-1} (Super-KEKB, Flanagan and Ohnishi 2004), and to enhance the performance of the detector and the reconstruction software (Belle II, Doležal and Uno 2010).

5.3 Detector Alignment

The alignment of an assembled detector aims at the geometry of its track-sensitive parts to be corrected for small lateral and angular displacements. As a general rule, the error caused by misalignment should be significantly inferior w.r.t. the intrinsic resolution of the sensitive elements. In order to achieve this goal, various strategies can be used. Sensor positions can be measured in the lab or in situ by lasers. To obtain ultimate precision, however, reconstructed tracks have to be used (Strandlie and Frühwirth 2010).

This task can be performed in two stages: in "local alignment," the individual sub-detectors are internally aligned; thereafter, "global alignment" determines the positions and orientations of entire sub-detectors w.r.t. each other. Advantage of this strategy is a reduction of the number of alignment parameters that are to be estimated concurrently. A possible disadvantage is the problem of achieving a consistent global alignment while keeping the internal alignments fixed. It may therefore be necessary to envisage a final adjustment of all alignment parameters of all sub-detectors.

The required frequency of re-alignments depends mainly on the mechanical stability of the various components. Distortions by gravity or magnetic forces, if significant, must be taken into account. The mounting of, e.g., a high-resolution silicon detector directly onto the beam-tube may necessitate one or even several global realignments per day.

With a sufficiently large number of tracks, chosen from a "clean sample," the statistical errors of the estimated alignment parameters can be made as small as required. The challenge is to control the systematic errors to the required level (Brown et al. 2009). This is because for any kind of tracks there are several degrees of freedom that are not constrained, usually referred to as weak modes. A mixture of tracks that is as diverse as possible is required to constrain all or at least most of the weak modes.

The requirements of past and current collider experiments on alignment performance, both in terms of precision and in terms of the sheer number of parameters to be estimated, have instigated the development of a generic algorithm (Millepede, Blobel and Kleinwort 2002; Blobel 2006; Blobel 2007). However, the largest LHC experiments, ATLAS and CMS, which meet the most difficult alignment tasks, do not rely on a single method, but have implemented several alternatives; this is extremely useful for debugging and cross-validation (Morley 2008; Widl and Frühwirth 2008; Adam et al. 2009b).

As a side benefit, alignment can be complemented by additionally providing a precise estimation of the material budgets, using reconstructed tracks. This is particularly important for low momenta, which are strongly affected by energy loss and multiple Coulomb scattering. Estimation methods for this task are currently under development (R. Frühwirth, private communication).

6 Event Reconstruction

The reconstruction of events acquired in the track-sensitive parts of a detector is a chain of tasks, each of which contributes to detector abstraction and data reduction. Care must be taken in order to avoid loss or distortion of relevant information, and keeping the sample representative.

Starting with raw measurements, the tasks to be performed are calibration, local pattern recognition in individual sub-detectors (track search), global pattern recognition

(track finding), track parameters fitting (based on an appropriate track model), vertex finding, and geometric and kinematic vertex fitting – thus ending up with the estimation of the event parameters relevant for subsequent physics analyses.

Calibration and usually local pattern recognition are tasks of sub-detector-specific software, which is not further discussed here.

This section deals with the next steps of the reconstruction chain. Their implementation will be based on established state-of-the-art algorithms and techniques, and will take advantage of available abstract interfaces to the framework, to the data model, and to the geometry and materials information. Access to the "simulation truth" and interfaces to the event graphics are required for software testing and optimization studies.

A broad range of approaches exist to fulfill the tasks mentioned above. This overview is intended to show the sophistication of techniques available. It is not exhaustive, and is focused mainly on colliding beam experiments. More thorough reviews can be found, e.g., in Frühwirth et al. (2000), Strandlie and Frühwirth (2010).

6.1 Pattern Recognition

The task of global pattern recognition, a.k.a. track finding, is to determine which hits in the track-sensitive parts of a detector belong to a common track. Information from a preceding local pattern recognition, yielding a track element candidate with associated hits, is used if available.

A large variety of techniques exist for track finding. Which one to choose depends among other things on the track density, the background noise level, the distance to the interaction point, and the sub-detector in question. Some track finding algorithms rely on the circular shape of the track in the transverse projection, and those will work only in regions where the magnetic field is rotationally symmetric and approximately homogeneous.

In a collider's radial region, due to smaller track density in the outer sub-detectors, an "inward/outward" strategy may be appropriate, e.g.,

- Track element candidates found by local pattern recognition in a large-volume central tracking detector (CTD) are extrapolated inward to high-precision vertex and/or inner tracking detectors (VITD), as available. Tracks that can be found by this technique have transverse momenta sufficiently high for crossing most or all of the CTD.
- Each VITD hit is tested against each extrapolation. If it can be associated to a track candidate, it is removed from the hit collection.
- The remaining VITD hits are the input to a local pattern recognition in these detectors. Various algorithms can be used in this "stand-alone track search." Among them are the conformal transformation (Frühwirth et al. 2000), the Hough transform (Hough 1959), the elastic net (Kisel et al. 1997; Gorbunov and Kisel 2006), the combinatorial Kalman filter (Mankel 1997), and related methods of progressive track search.
- Track element candidates found in this way are extrapolated from the VITD outward into the CTD. If unused CTD hits are found that fit to the extrapolation, they are associated to this track candidate.

A different strategy starts with stand-alone track searches in both the VITD and CTD. Track element candidates found by one are extrapolated and matched against those found by the other, and linked if passing the test. In a final step, unused hits are tried again to be associated.

In the forward region, defined by tracks crossing only a small part or none of the CTD, dedicated forward tracking detectors (FTD) should be available. This may lead to complex situations where no general strategy can be given. In such cases, individual solutions might be appropriate.

Pattern recognition results in a set of global track candidates, each consisting of a crude estimate of its parameters together with an ordered list of associated sub-detector hits, plus a set of unassociated hits.

A more sophisticated approach allows hits to be shared among two or more track candidates; in this case the final hit association is relegated to the next task, track reconstruction.

6.2 Track Reconstruction

The trajectory of a charged particle, moving in a stationary magnetic field and in the absence of matter, is determined by the Lorentz force law. It is uniquely described by five independent parameters, representing the initial conditions at, e.g., the point of intersection with a chosen reference surface: two positions, two directions, and a quantity depending on the momentum.

The global track candidates that have been found in the previous task are subjected to single track fitting based on some track model, with the measurements and their error matrix being derived from the hit coordinates, and contributions from the detector material correctly taken into account. The purpose of this step is twofold: (1) find a "best estimate" for the track parameters, yielding a 5-tuple and a corresponding 5×5 covariance matrix; and (2) test and possibly update the track hypothesis, i.e., the association of hits to this track, by identifying and removing or down-weighting "outliers" (true measurements with errors larger than shown, or false ones not belonging to the track at all) and, if present, resolving ambiguous associations of hits. These two tasks may be achieved by an iterative procedure.

The choice of an appropriate track model depends on the magnetic field. At many experiments, it is sufficiently close to being homogeneous in the tracking region, implying a helix track model. Significant deviations from homogeneity can be handled either by analytical corrections (if rotationally symmetric), or by stepwise numerical integration (e.g., Runge–Kutta–Nyström algorithm (Abramowitz and Stegun 1965)). Moreover, the detector layout is often approximately rotationally symmetric w.r.t. an axis parallel to the magnetic field. Track parameters suggested to be used internally or externally are given in Sect. 9.4.

The free particle trajectory is disturbed by material effects in the detector. The most important one is multiple Coulomb scattering off nuclei. This is a stochastic process with a core distribution of two local projections of the scattering angle well described by the Rossi–Greisen–Highland formula (Highland 1975), except for extremely thin material with few scatterings (for details see Frühwirth and Regler (2001)). If the material budget is concentrated in thin layers (e.g., silicon detector layers, gas detector walls), those may be approximated by geometrical zero thickness; in that case only a kink occurs in the layer, and the track's covariance matrix is "blown up" only in the two direction parameters.

Another important material effect is energy loss. For particles heavier than electrons, this is dominated by ionization; if taken into account by the track model (in particular for low momenta), it can be described by a deterministic step in ΔE as derived from the Bethe–Bloch formula with Fermi plateau (Frühwirth et al. 2000). For electrons at high momenta, energy loss is dominated by bremsstrahlung, which is a process of stochastic occurrence that can be described stochastically by the Bethe–Heitler model (Bethe and Heitler 1934). As it is extremely

non-Gaussian, this poses problems that can be solved by special modifications of the fitting algorithm, e.g., the Gaussian Sum Filter (Adam et al. 2005) or alternatives (Kartvelishvili 2007). However, if the bremsstrahlung photon can be identified and measured, a deterministic ΔE correction on the electron may be appropriate (Urquijo and Barberio 2005).

State of the art in track fitting are techniques based on the Kalman filter (Frühwirth 1987), a recursive and locally linear estimator, which is equivalent to the global least squares method, but is superior in coping with the requirements of a complex modular detector. It is applied on a linear approximation to the track model, the expansion point being a "reference track" sufficiently close to the true track; the crude estimate from track finding may be a good choice. If necessary, a Newton method can be applied, with the new reference track being the track fitted in the previous iteration. Improvement may be gained by local adjustments of the reference track that introduces nonlinear terms into the model.

The best estimate of the track parameters is achieved only at the end of the filter, after all measurements having contributed. In order to obtain best estimates all along the trajectory, the filter must be complemented by a smoother. This can be achieved either iteratively, or with help of a second Kalman filter running in the opposite direction. The smoothed χ^2s may serve as test criteria for outlier measurements; in the realistic case of multiple outliers, however, the power of these tests is limited. For details, see the formulae shown in Sect. 9.1.

Linear estimators are inherently sensitive to the influence of outliers. Robustification can be achieved by using nonlinear estimators, which in the case of track fitting are most easily implemented as iterative extensions of the Kalman filter: measurements with "big" normalized residuals are down-weighted in the next iteration. Such adaptive filters are largely resistant against wrong measurements; thus they are able to defer the final association of hits from the track finding step to the track fitting step, where more complete information about the track is available for the test criteria. This final association need not be a "hard" one, as the final weights of the measurements can be anywhere between 0 and 1.

There exist several adaptive filter methods. In particular, the Deterministic Annealing Filter (DAF) (Frühwirth and Strandlie 1999; Strandlie and Frühwirth 2010) has successfully been implemented and refined by the CMS experiment at LHC.

A nontrivial challenge is flexible track propagation in complex sequences of cylindrical and plane detectors, as is often the case in the forward region (Regler et al. 2008a). Recently, a detector-independent skeleton toolkit for track reconstruction has been developed (GENFIT, (Höppner et al. 2010)).

6.3 Vertex Reconstruction

Vertex reconstruction consists of (1) vertex finding and track bundling (a pattern recognition task); (2) geometric vertex fitting (statistical estimation of the vertex position and the track parameters at the vertex); and (3) re-fitting with the kinematic constraints of energy–momentum conservation. These tasks are in practice often intertwined.

The "virtual measurements" of the vertex fit are the tracks fitted independently in the previous step, and conveniently extrapolated inward into the beam-tube; thus no material effects need to be taken into account anymore. Their five parameters and 5×5 covariance matrix may be defined on a reference surface (e.g., the inside of the beam-tube), or as perigee parameters w.r.t. a fixed pivot point close to the beam interaction (see Sect. 9.4). In case of a primary vertex

the beam interaction profile, if known, may be included as another virtual measurement; it may be determined from the vertex fits of a sample of selected "clean" low-multiplicity events, collected over a period of stable beams. Note that the global covariance matrix of all measurements is block-diagonal.

A vertex fitted from n tracks consists of $(3 + 3n)$ parameters: three position coordinates, and three track parameters at the vertex for each track. For a subsequent kinematic fit the full $(3+3n) \times (3+3n)$ covariance matrix is required; otherwise, only some 3×3 sub-matrices may be of interest.

Geometric vertex fitting starts with a Kalman filter (Frühwirth 1987; Frühwirth et al. 1996), with the linear expansion point being chosen from the beam interaction profile and/or crude track intersections. A smoother is required for obtaining the best estimates of all tracks added before the last one. The smoothed χ^2s may be used as test criteria for outliers, i.e., tracks not belonging to this vertex; however, the power of these tests suffers for the same reasons as discussed with track fitting above. Formulae are given in ❯ Sect. 9.2.

The remedies are robustified, nonlinear, adaptive filters – in particular the Deterministic Annealing Filter (DAF) (Waltenberger et al. 2007; Strandlie and Frühwirth 2010) which downweights the influence of outlier tracks on the fitted vertex. In addition, it introduces an annealing schedule that serves two purposes: it prevents the fit from falling into a local minimum, and allows progressively moving from "soft" to ultimate "hard assignments." Formulae are given in ❯ Sect. 9.3.

This leads in a natural way to a technique of vertex finding, called Adaptive Vertex Reconstruction (AVR) (Waltenberger 2008): if the DAF is applied to all tracks, with the proper setting of the annealing schedule it converges to the vertex with the largest number of tracks. The tracks attached to this vertex are removed from the track collection, and the DAF is applied to the remaining tracks. This is iterated until no more vertices can be found. The AVR has been evaluated against other vertex finders, in particular ZvTop (Jackson 1997). The DAF can be further generalized by introducing a Multi-Vertex Filter (MVF) (Frühwirth and Waltenberger 2004), simultaneously fitting n tracks to m competing vertices by "soft assignment" of each track to more than one vertex (see ❯ Sect. 9.3).

The inclusion of electron tracks fitted with a Gaussian Sum Filter (GSF) causes the vertex fit also to be modeled as a Gaussian sum. This can result in exponential bloat unless compensated for by a "collapsing strategy" (Speer and Frühwirth 2006). GSF is compatible with the adaptive DAF and MVF.

Kinematic fitting imposes the constraints of energy–momentum conservation on each vertex. The method used is that of Lagrangian multipliers (Avery 1999; Frühwirth et al. 2000). It can be achieved either as a separate re-fit step after pure geometric vertex fitting, or directly by a simultaneous geometry and kinematic fit. Problems arise in cases of insufficient information (unknown masses, unseen particles, undefined jet momenta, etc.). In $e^- e^+$ collisions, the well-defined initial state is stochastically disturbed by initial-state radiation photons off a beam particle before its interaction (Beckmann et al. 2010). Complex cascade decay chains are difficult to handle in a general way. And kinematic fitting may more likely be subject to the requirements of a subsequent physics analysis.

Vertex reconstruction, being the last step in the event reconstruction chain, may be boxed in a detector-independent toolkit. Such a toolkit exists (RAVE, Moser et al. 2008; Waltenberger 2011). It includes the Kalman filter, the DAF, the MVF, the GSF, the AVR, and an interface to ZvTop.

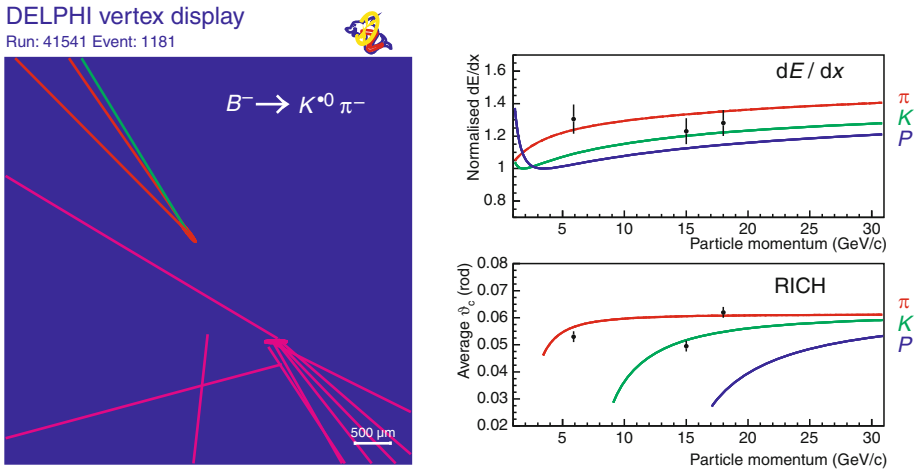


Fig. 13
Reconstructed charmless B^--meson decay (*left*) and associated particle identification (*right*) in the DELPHI experiment (Abreu et al. 1996) at LEP

The importance of vertex reconstruction has impressively been demonstrated in the experiments at LEP (Hübner 2004); an example is shown in *Fig. 13*. At "Beauty Factories" (BaBar, Belle), physics studies like CP violation in $B\bar{B}$ systems rely on precise measurements of Δz between the two B decay vertices, and TeV-scale colliders (LHC, ILC) strive for efficient heavy-flavour tagging. This can only be achieved by complementing the high-resolution silicon vertex detectors with sophisticated reconstruction software, based on statistical methods fully exploiting the information available.

7 Performance Optimization

Big complex detectors are used in experiments at modern particle accelerators and colliders, in order to "squeeze out" a maximum of information about the underlying physics. Performance challenges on behalf of the track-sensitive parts include precise momentum, direction, and spatial resolutions of the trajectories of charged particles, both on their own and in support of other detector parts, like calorimetry by particle flow analysis (Thomson 2009).

In order to achieve these goals, care must be taken from the very beginning in the design of the experimental set-up. This includes:

- Choice of the overall design (e.g., volume available for the tracking and the calorimeters, constrained by the size of the magnet);
- Layout of the components of the tracking parts (vertex detector, central tracking detector, forward detectors, supplementing detectors, etc.);
- Technology: silicon trackers (high spatial resolution of sensors, appreciable amount of multiple scattering) versus gaseous track chambers (worse spatial resolution of single measurements, high momentum resolution due to many measurements, and little multiple

scattering in the gas volume, however, much multiple scattering in the inner wall and end plates);
- Impact of material budgets, magnetic field characteristics, radius of the beam-tube, expected event rates, background radiation, etc;
- Constraints from the machine–detector interface, and costs.

The choices taken will of course strongly depend on the type of experiment (fixed-target vs. colliding beams), on the type of collider (hadrons vs. e^-e^+ vs. e^-p), and on its beam energies and luminosity.

Three regions can be distinguished in the tracking parts of a detector:

- *Central region* (including a "recoil part" in fixed-target experiments)
- *Forward and backward regions* (forward prominent in fixed-target exp.)
- *Transition region*, which is most challenging, albeit often neglected

For detector optimization in view of tracking performance, several techniques are available: analytical calculations from simple to sophisticated; and Monte-Carlo studies from rough but fast to elaborated full simulations, followed by event reconstruction at various levels of detail. Other reasons for Monte-Carlo studies include software debugging, detector acceptance corrections, signal and background estimations for physics analyses, and adjustment of test-beam results for realistic conditions.

The rest of this section refers mainly to colliding beams. When starting to design a new experiment, a first guess of the track resolution in the central region of an approximately cylindrical set-up inside a homogeneous magnetic field may be based on the well-known Gluckstern formulae (Gluckstern 1963). For example the transverse momentum resolution, neglecting multiple scattering, is given by

$$[\sigma(p_T)/p_T]_{\text{det}} \propto \frac{\sigma_{R\Phi}}{B_z L^2} \cdot p_T$$

with $\sigma_{R\Phi}$ being the azimuthal spatial resolution of a single measurement in a stack of layers over a lever arm L, and B_z, the magnetic field (for coordinates and track parameters, see the Sect. 9.4).

Recent improvements have extended the validity of those formulae to:

- Detectors not homogeneous w.r.t. resolution and material budget (Regler and Frühwirth 2008a)
- The track's azimuthal direction angle w.r.t. the cylinder normal, $|\beta|$, not being small (which is the case for low-momentum tracks) (Regler and Frühwirth 2008a)
- The track's dip angle w.r.t. the transverse plane, $|\lambda|$, not being small (covering substantial part of the forward and backward directions) (Valentan et al. 2009)

In the extreme forward and backward regions of the detector, application of Gluckstern-like formulae requires a lot of skill. In these cases, it is advised to use right from the beginning a fast simulation and reconstruction tool. This is anyway a must in the transition region mentioned above. It is also true for estimating vertex precision, and a rough secondary vertex separation.

In the "barrel region" of a cylinder-symmetric set-up (approximately characterized by $\pi/4 < \vartheta < 3\pi/4$), resolutions are commonly parametrized as functions of the transverse momentum

p_T and the polar angle ϑ:

$$\text{Transverse momentum:} \quad \sigma(p_T)/p_T = a \cdot p_T \oplus b/\sqrt{\sin\vartheta}$$
$$\text{Transv. impact parameter:} \quad \sigma(\delta_T) = c \oplus d/(P \cdot \sin^{3/2}\vartheta)$$

(quadratic addition of two terms, arising from detector resolution "det" and multiple scattering "m/s," respectively). The ILD detector at ILC aims at $a < 2 \cdot 10^{-5}$/GeV, $b < 10^{-3}$, $c < 5\,\mu$m, and $d < 10\,\mu$m·GeV (Stoeck et al. 2010).

Figure 14 shows example resolutions of transverse momentum p_T and impact parameter δ_T as functions of p_T, for polar angles $\vartheta \geq 15°$:

Resolution of the absolute momentum can be calculated from those of transverse momentum and polar angle (known from the fit) (Valentan et al. 2009):

$$[\sigma(P)/P]^2 = [\sigma(p_T)/p_T]^2 + [\sigma(\vartheta) \cdot \cot\vartheta]^2 + \mathrm{cov}(1/p_T, \vartheta) \cdot p_T \cdot \cot\vartheta$$
$$\approx [\sigma(p_T)/p_T]^2 + [\sigma^2(\vartheta)_{\mathrm{det}} - \sigma^2(\vartheta)_{\mathrm{m/s}}] \cdot \cot^2\vartheta\,.$$

In an early design phase, detector optimization must be flexible and fast enough for frequent quick discussions concerning local modifications. This calls for lightweight fast simulation and reconstruction tools, capable of closely modeling reality, like LDT (Regler et al. 2008a; Regler et al. 2008b), used for *Fig. 14*.

However, their limits should always be kept in mind: Precise knowledge of the magnetic field may not be necessary now, but the penalty for incorporating field inhomogeneities must be paid later. A magnetic field is crucial for the performance of a TPC, but the effect of its forces may require new global alignment. Beam radiation background and high particle flux densities degrade significantly the performance of detectors close to the beam-tube, like vertex detectors; these effects can only roughly be simulated by heuristic corrections of their resolutions and efficiencies.

A danger is "over-instrumentation" of the inner silicon tracking regions: any gain in direction and momentum resolution by adding layers is reversed by increased multiple Coulomb scattering. Minimizing the material budget of the sensors makes an all-silicon central tracker feasible. However, good momentum resolutions at low momenta can at present only be achieved by a gaseous central tracker, like a TPC. Although its inner wall is a massive scatterer and effectively breaks the track information into two, the absolute momentum (accurately measured

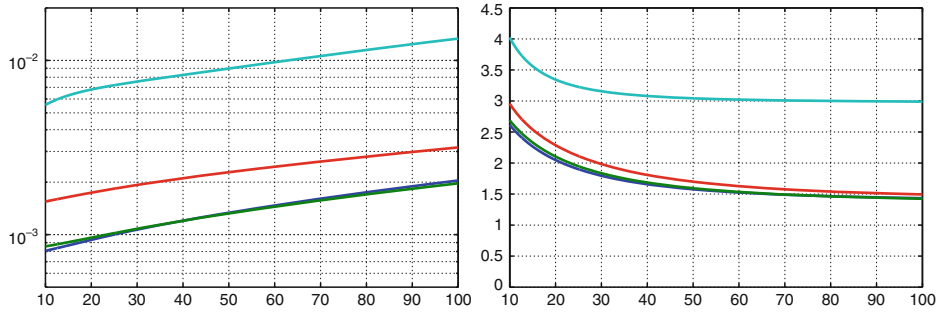


Fig. 14

ILD detector: rms$[\Delta(p_T)/p_T]$ (log scale, *left*) and rms$[\delta_T/\mu\mathrm{m}]$ (*right*) versus p_T/GeV, for ϑ = 15° (*cyan*), 30° (*red*), 60° (*green*), and 90° (*blue*)

in the TPC) is preserved. For the impact parameter, the degradation of direction and transverse momentum by the wall can be compensated by precise measurements at the innermost layer of a vertex detector close to the beam-tube as small as possible. On the other hand, although some detector layers in the intermediate region between vertex detector and TPC may not substantially contribute to the track resolution, they might be beneficial for pattern recognition.

In an advanced design phase, a full detector simulation (e.g., based on the GEANT4 tool, Agostinelli et al. 2003), followed by an adequately detailed reconstruction including realistic pattern recognition, should be used. Only by this way a correct knowledge of the influence of backgrounds (originating from beamstrahlung or detector noise), which are not correlated to the tracks, can be obtained. Other influences arise from, e.g., delta-ray ghosts, high track multiplicities, or multiply overlaid event data (as expected at the LHC). These affect pattern recognition and track reconstruction. Such simulation studies provide important input for the required digital resolutions, "time stamp" information, and pulse-height measurements of the detectors.

Colliders at TeV energies require a precise reconstruction of hadron jets, composed of charged and neutral particles. Particle flow analysis (Thomson 2009) combines the measurements of tracking detectors with those of fine-grained electromagnetic and hadronic calorimeters. The jet energy resolution, neglecting leakage and confusion, is parametrized as quadratic addition of terms from intrinsic calorimeter resolution and from imperfect track reconstruction:

$$\sigma(E_{\text{jet}})/E_{\text{jet}} = x \,/\, \sqrt{E_{\text{jet}}} \;\oplus\; y\,.$$

The ILD detector at ILC aims at $x < 30\%\sqrt{\text{GeV}}$, $y < 1\%$ (Stoeck et al. 2010).

8 Summary

Reconstruction of the trajectory of charged particles is essential for experiments at high-energy accelerators, in astro-particle physics, and for medical applications (diagnostics and particle therapy). Tracking detectors consist of position-sensitive elements, which respond to the signals induced by the passage of charged particles. For measuring these positions, several technologies have been developed: various types of gaseous detectors, semiconductor detectors, scintillating fiber detectors, and nuclear emulsions for special purposes. Further track parameters such as momentum and charge can be measured if the detectors operate together with a magnetic field. Precise reconstruction from raw measurements involves sophisticated statistical methods.

9 Appendix: Formulae

9.1 Kalman Track Fitting

The Kalman filter is used for recursively fitting a trajectory by the measurements of n independent sub-detectors. Multiple scattering is treated locally, thus only small matrices have to be inverted. Eventually, the fitted track parameters may be re-computed anywhere along the trajectory ("smoother").

The *system equations* describe the propagation of the track parameter vector $\boldsymbol{p}_k$ through a set of reference surfaces $k = 1, \ldots, n$ by deterministic functions $\boldsymbol{f}_{k-1}$ derived from the track model, plus random "process noise" $\boldsymbol{\omega}_k$ due to multiple scattering between the surfaces for each step $k-1 \to k$. The *measurement equations* describe the transformation $\boldsymbol{p}_k \to \boldsymbol{m}_k$ (the measurements associated with reference surface k) by deterministic functions $\boldsymbol{h}_k$, plus random "measurement noise" $\boldsymbol{\epsilon}_k$. Both $\boldsymbol{\omega}_k$ and $\boldsymbol{\epsilon}_k$ are assumed to be unbiased and defined by the covariance matrices $\boldsymbol{Q}_k$ and $\boldsymbol{V}_k$, respectively:

$$\begin{aligned} \boldsymbol{p}_k &= \boldsymbol{f}_{k-1}(\boldsymbol{p}_{k-1}) + \boldsymbol{\omega}_k, & \operatorname{cov}(\boldsymbol{\omega}_k) &\equiv \boldsymbol{Q}_k \\ \boldsymbol{m}_k &= \boldsymbol{h}_k(\boldsymbol{p}_k) + \boldsymbol{\epsilon}_k, & \operatorname{cov}(\boldsymbol{\epsilon}_k) &\equiv \boldsymbol{V}_k = \boldsymbol{G}_k{}^{-1} \end{aligned}$$

A *linearization* of the track model functions $\boldsymbol{f}_k$ and $\boldsymbol{h}_k$ is achieved by first-order Taylor expansions around an undisturbed "reference track" $\boldsymbol{p}_{k,e}$:

$$\begin{aligned} \boldsymbol{f}_k(\boldsymbol{p}_k) &\approx \boldsymbol{f}_k(\boldsymbol{p}_{k,e}) + \boldsymbol{F}_k(\boldsymbol{p}_k - \boldsymbol{p}_{k,e}) = \boldsymbol{F}_k\boldsymbol{p}_k + \boldsymbol{a}_{k,e} \\ \boldsymbol{h}_k(\boldsymbol{p}_k) &\approx \boldsymbol{h}_k(\boldsymbol{p}_{k,e}) + \boldsymbol{H}_k(\boldsymbol{p}_k - \boldsymbol{p}_{k,e}) = \boldsymbol{H}_k\boldsymbol{p}_k + \boldsymbol{b}_{k,e} \end{aligned}$$

with $\boldsymbol{F}_k = [\partial \boldsymbol{p}_{k+1}/\partial \boldsymbol{p}_k)]_e$ and $\boldsymbol{H}_k = [\partial \boldsymbol{m}_k/\partial \boldsymbol{p}_k)]_e$ being the Jacobian matrices of derivatives at $\boldsymbol{p}_{k,e}$. The constant terms $\boldsymbol{a}_{k,e}$ and $\boldsymbol{b}_{k,e}$ can be transformed away, e.g., by redefining $\boldsymbol{p}_k \to \boldsymbol{p}_k - \boldsymbol{p}_{k,e}$ and $\boldsymbol{m}_k \to \boldsymbol{m}_k - \boldsymbol{h}_k(\boldsymbol{p}_{k,e})$, and are in the following omitted ("homogenization"). This yields

$$\boldsymbol{p}_k = \boldsymbol{F}_{k-1}\boldsymbol{p}_{k-1} + \boldsymbol{\omega}_k, \qquad \boldsymbol{m}_k = \boldsymbol{H}_k\boldsymbol{p}_k + \boldsymbol{\epsilon}_k$$

Each Kalman filter step consists of two basic operations: *Prediction* is an estimation of the track parameter vector at k, based upon the estimate at $k-1$, i.e., including only measurements $1, \ldots, k-1$. *Filtering* is an updated estimation at k, computed by a weighted mean of the predicted estimate and the measurements at k, i.e., now including measurements $1, \ldots, k$.

Only at the last reference surface n does the filtered estimate include all measurements $1, \ldots, n$. *Smoothing* is an update of the filtered estimate at any $k < n$ with the missing so far measurements $k+1, \ldots, n$. This can be achieved either iteratively ($k = n-1, \ldots, 1$) based upon the smoothed estimate at $k+1$; or equivalently by computing the weighted mean of the filtered estimate with the predicted one of a second filter running in the opposite direction.

Predictor formulae for estimate, residual, and their covariances (the lower index refers to the current reference surface, the upper index denotes the last measurements included; if equal, the upper index may be omitted; the subscript "back" refers to a filter running opposite) read

$$\begin{aligned} &\tilde{\boldsymbol{p}}_k^{k-1} = \boldsymbol{F}_{k-1}\tilde{\boldsymbol{p}}_{k-1} \\ &\operatorname{cov}\left(\tilde{\boldsymbol{p}}_k^{k-1}\right) \equiv \boldsymbol{C}_k^{k-1} = \boldsymbol{F}_{k-1}\boldsymbol{C}_{k-1}\boldsymbol{F}_{k-1}{}^T + \boldsymbol{Q}_k \\ &\boldsymbol{r}_k^{k-1} = \boldsymbol{m}_k - \boldsymbol{H}_k\tilde{\boldsymbol{p}}_k^{k-1}, \qquad \operatorname{cov}\left(\boldsymbol{r}_k^{k-1}\right) \equiv \boldsymbol{R}_k^{k-1} = \boldsymbol{V}_k + \boldsymbol{H}_k\boldsymbol{C}_k^{k-1}\boldsymbol{H}_k{}^T \end{aligned}$$

Filter formulae for estimate, residual, their covariances, and chi-squares:

$$\tilde{\boldsymbol{p}}_k = \boldsymbol{C}_k \left[\left(\boldsymbol{C}_k^{k-1} \right)^{-1} \tilde{\boldsymbol{p}}_k^{k-1} + \boldsymbol{H}_k{}^T \boldsymbol{G}_k \boldsymbol{m}_k \right]$$

$$\mathrm{cov}(\tilde{\boldsymbol{p}}_k) \equiv \boldsymbol{C}_k = \left[\left(\boldsymbol{C}_k^{k-1} \right)^{-1} + \boldsymbol{H}_k{}^T \boldsymbol{G}_k \boldsymbol{H}_k \right]^{-1}$$

$$\boldsymbol{r}_k = \boldsymbol{m}_k - \boldsymbol{H}_k \tilde{\boldsymbol{p}}_k, \qquad \mathrm{cov}(\boldsymbol{r}_k) \equiv \boldsymbol{R}_k = \boldsymbol{V}_k - \boldsymbol{H}_k \boldsymbol{C}_k \boldsymbol{H}_k{}^T$$

$$\chi^2_{k,F} = \boldsymbol{r}_k{}^T \boldsymbol{R}_k{}^{-1} \boldsymbol{r}_k \text{ (filtered)}, \qquad \chi^2_k = \chi^2_{k-1} + \chi^2_{k,F} \text{ (total when } k = n)$$

For starting up, set $\tilde{\boldsymbol{p}}_1^0 = \boldsymbol{p}_{1,e}$ (before its redefinition) and $\left(\boldsymbol{C}_1^0 \right)^{-1} = \mathrm{diag}(\zeta)$, with $\zeta = 0$ if $\dim(\boldsymbol{m}_1) > 5$, or $0 < \zeta \ll 1$ otherwise.

Smoother formulae for estimate, residual, their covariances, and chi-square:

$$\tilde{\boldsymbol{p}}_k^n = \boldsymbol{C}_k^n \left[(\boldsymbol{C}_k)^{-1} \tilde{\boldsymbol{p}}_k + \left(\boldsymbol{C}_k^{k+1} \right)^{-1}_{\text{back}} \tilde{\boldsymbol{p}}_{k,\text{back}}^{k+1} \right]$$

$$\mathrm{cov}\left(\tilde{\boldsymbol{p}}_k^n \right) \equiv \boldsymbol{C}_k^n = \left[(\boldsymbol{C}_k)^{-1} + \left(\boldsymbol{C}_k^{k+1} \right)^{-1}_{\text{back}} \right]^{-1}$$

$$\boldsymbol{r}_k^n = \boldsymbol{m}_k - \boldsymbol{H}_k \tilde{\boldsymbol{p}}_k^n, \qquad \mathrm{cov}\left(\boldsymbol{r}_k^n \right) \equiv \boldsymbol{R}_k^n = \boldsymbol{V}_k - \boldsymbol{H}_k \boldsymbol{C}_k^n \boldsymbol{H}_k{}^T$$

$$\chi^2_{k,S} = \boldsymbol{r}_k^{n\,T} \left(\boldsymbol{R}_k^n \right)^{-1} \boldsymbol{r}_k^n \quad \text{(smoothed)}$$

with the $\chi^2_{k,S}$ being not independent of each other, and $\sum_{k=1}^n \chi^2_{k,S} \neq \chi^2_n$.

Outliers are "bad measurements" $\boldsymbol{m}_k$, which may be mis-associated to this track, or have errors much larger than expected from $\boldsymbol{V}_k$, or are background noise. In case of no outliers, Gaussian errors well represented by $\boldsymbol{V}_k$, and good linear approximation, the smoothed $\chi^2_{k,S}$ are chi-square distributed with $\dim(\boldsymbol{m}_k)$ degrees of freedom. Thus, they are a practical test statistic for outlier removal. However, there is only little "robustness" against multiple outliers.

9.2 Kalman Vertex Fitting

The Kalman filter is used for geometrically fitting $n \geq 2$ tracks to a common vertex. Virtual measurements are the fitted track parameters $\boldsymbol{p}_k$, $k = 1, \ldots, n$, conveniently extrapolated inward, and assumed to be uncorrelated; thus the global covariance matrix is (5×5) block-diagonal. Aim is the estimation of a "state vector" composed of the vertex position $\boldsymbol{x}$, and of the momenta $\boldsymbol{q}_k$ of all tracks at the vertex.

Each *filter step* consists of adding a new track $\boldsymbol{p}_k$ to a vertex already fitted with $k-1$ tracks, by updating its position estimate $\boldsymbol{x}_{k-1} \to \boldsymbol{x}_k$, and estimating the track's $\boldsymbol{q}_k$ at the vertex. *Smoothing* is an update of the filtered estimates $\boldsymbol{q}_k$ for $k < n$, just using the final estimate of the vertex position $\boldsymbol{x}_n$.

There is no "process noise," thus the *prediction* is trivial, and the *system equations* are simply the identity. The *measurement equations* describe the dependence $(\boldsymbol{x}, \boldsymbol{q}_k) \to \boldsymbol{p}_k$ by deterministic functions $\boldsymbol{h}_k$ derived from the track model, plus random "measurement noise" $\boldsymbol{\epsilon}_k$ assumed to be unbiased and defined by the fitted tracks' covariance matrices $\boldsymbol{V}_k$:

$$\boldsymbol{x}_k(\boldsymbol{x}_{k-1}) = \boldsymbol{x}_{k-1} =: \boldsymbol{x}$$

$$\boldsymbol{p}_k = \boldsymbol{h}_k(\boldsymbol{x}, \boldsymbol{q}_k) + \boldsymbol{\epsilon}_k, \qquad \mathrm{cov}(\boldsymbol{\epsilon}_k) = \boldsymbol{V}_k = \boldsymbol{G}_k{}^{-1}$$

A *linearized track model* is obtained by approximating $\boldsymbol{h}_k$ by a first-order Taylor ansatz around some "expansion point" $(\boldsymbol{x}_e, \boldsymbol{q}_{k,e})$:

$$\begin{aligned}\boldsymbol{h}_k(\boldsymbol{x}, \boldsymbol{q}_k) &\approx \boldsymbol{h}_k(\boldsymbol{x}_e, \boldsymbol{q}_{k,e}) + \boldsymbol{A}_k(\boldsymbol{x} - \boldsymbol{x}_e) + \boldsymbol{B}_k(\boldsymbol{q}_k - \boldsymbol{q}_{k,e}) = \\ &= \boldsymbol{A}_k \boldsymbol{x} + \boldsymbol{B}_k \boldsymbol{q}_k + \boldsymbol{c}_{k,e}\end{aligned}$$

with $\boldsymbol{A}_k = [\partial \boldsymbol{p}_k / \partial \boldsymbol{x}]_e$ and $\boldsymbol{B}_k = [\partial \boldsymbol{p}_k / \partial \boldsymbol{q}_k]_e$ being the Jacobian matrices of derivatives at $(\boldsymbol{x}_e, \boldsymbol{q}_{k,e})$. The constants $\boldsymbol{c}_{k,e}$ can be transformed away by redefining $\boldsymbol{p}_k \to \boldsymbol{p}_k - \boldsymbol{c}_{k,e}$ and are in the following omitted ("homogenization"):

$$\boldsymbol{p}_k = \boldsymbol{A}_k \boldsymbol{x} + \boldsymbol{B}_k \boldsymbol{q}_k + \boldsymbol{\epsilon}_k$$

Filter formulae for estimate, residual, their covariances, and chi-squares:

$$\begin{aligned}
&\tilde{\boldsymbol{x}}_k = \boldsymbol{C}_k \left[{\boldsymbol{C}_{k-1}}^{-1} \tilde{\boldsymbol{x}}_{k-1} + {\boldsymbol{A}_k}^T {\boldsymbol{G}_k}^B \boldsymbol{p}_k \right], \\
&\qquad\qquad \text{with } {\boldsymbol{G}_k}^B = \boldsymbol{G}_k - \boldsymbol{G}_k \boldsymbol{B}_k \boldsymbol{W}_k {\boldsymbol{B}_k}^T \boldsymbol{G}_k \\
&\tilde{\boldsymbol{q}}_k = \boldsymbol{W}_k {\boldsymbol{B}_k}^T \boldsymbol{G}_k (\boldsymbol{p}_k - \boldsymbol{A}_k \tilde{\boldsymbol{x}}_k), \qquad \text{with } \boldsymbol{W}_k = \left({\boldsymbol{B}_k}^T \boldsymbol{G}_k \boldsymbol{B}_k \right)^{-1} \\
&\operatorname{cov}(\tilde{\boldsymbol{x}}_k) \equiv \boldsymbol{C}_k = \left({\boldsymbol{C}_{k-1}}^{-1} + {\boldsymbol{A}_k}^T {\boldsymbol{G}_k}^B \boldsymbol{A}_k \right)^{-1} \\
&\operatorname{cov}(\tilde{\boldsymbol{q}}_k) \equiv \boldsymbol{D}_k = \boldsymbol{W}_k + {\boldsymbol{E}_k}^T {\boldsymbol{C}_k}^{-1} \boldsymbol{E}_k \\
&\operatorname{cov}(\tilde{\boldsymbol{x}}_k, \tilde{\boldsymbol{q}}_k) \equiv \boldsymbol{E}_k = -\boldsymbol{C}_k {\boldsymbol{A}_k}^T \boldsymbol{G}_k \boldsymbol{B}_k \boldsymbol{W}_k \\
&\boldsymbol{r}_k = \boldsymbol{p}_k - \boldsymbol{A}_k \tilde{\boldsymbol{x}}_k - \boldsymbol{B}_k \tilde{\boldsymbol{q}}_k \\
&\operatorname{cov}(\boldsymbol{r}_k) \equiv \boldsymbol{R}_k = \boldsymbol{V}_k \left({\boldsymbol{G}_k}^B - {\boldsymbol{G}_k}^B \boldsymbol{A}_k \boldsymbol{C}_k {\boldsymbol{A}_k}^T {\boldsymbol{G}_k}^B \right) \boldsymbol{V}_k \\
&\chi^2_{k,F} = (\tilde{\boldsymbol{x}}_k - \tilde{\boldsymbol{x}}_{k-1})^T {\boldsymbol{C}_{k-1}}^{-1} (\tilde{\boldsymbol{x}}_k - \tilde{\boldsymbol{x}}_{k-1}) + {\boldsymbol{r}_k}^T \boldsymbol{G}_k \boldsymbol{r}_k \quad \text{(filtered)} \\
&\chi^2_k = \chi^2_{k-1} + \chi^2_{k,F} \quad \text{(total when } k = n\text{)}
\end{aligned}$$

If there exists prior information of the vertex position $\tilde{\boldsymbol{x}}_0$ and its covariance matrix $\boldsymbol{C}_0$ (e.g., from the beam interaction profile), use it as additional measurement. Otherwise, set $\tilde{\boldsymbol{x}}_0 = \boldsymbol{x}_e$ and $\boldsymbol{C}_0^{-1} = \operatorname{diag}(\zeta)$, with $0 < \zeta \ll 1$.

Smoother formulae for estimate, residual, their covariances, and chi-square:

$$\begin{aligned}
&\tilde{\boldsymbol{q}}^n_k = \boldsymbol{W}_k {\boldsymbol{B}_k}^T \boldsymbol{G}_k (\boldsymbol{p}_k - \boldsymbol{A}_k \tilde{\boldsymbol{x}}_n) \\
&\operatorname{cov}\left(\tilde{\boldsymbol{q}}^n_k\right) \equiv \boldsymbol{D}^n_k = \boldsymbol{W}_k + {\boldsymbol{E}^n_k}^T {\boldsymbol{C}_n}^{-1} \boldsymbol{E}^n_k \\
&\operatorname{cov}\left(\tilde{\boldsymbol{x}}_n, \tilde{\boldsymbol{q}}^n_k\right) \equiv \boldsymbol{E}^n_k = -\boldsymbol{C}_n {\boldsymbol{A}_k}^T \boldsymbol{G}_k \boldsymbol{B}_k \boldsymbol{W}_k \\
&\operatorname{cov}\left(\tilde{\boldsymbol{q}}^n_k, \tilde{\boldsymbol{q}}^n_j\right) = {\boldsymbol{E}^n_k}^T {\boldsymbol{C}_n}^{-1} \boldsymbol{E}^n_j \qquad (\text{for } k \neq j) \\
&\boldsymbol{r}^n_k = \boldsymbol{p}_k - \boldsymbol{A}_k \tilde{\boldsymbol{x}}_n - \boldsymbol{B}_k \tilde{\boldsymbol{q}}^n_k \\
&\operatorname{cov}\left(\boldsymbol{r}^n_k\right) \equiv \boldsymbol{R}^n_k = \boldsymbol{V}_k \left({\boldsymbol{G}_k}^B - {\boldsymbol{G}_k}^B \boldsymbol{A}_k \boldsymbol{C}_n {\boldsymbol{A}_k}^T {\boldsymbol{G}_k}^B \right) \boldsymbol{V}_k \\
&\chi^2_{k,S} = \left(\tilde{\boldsymbol{x}}_n - \tilde{\boldsymbol{x}}^{n*}_k\right)^T \left(\boldsymbol{C}^{n*}_k\right)^{-1} \left(\tilde{\boldsymbol{x}}_n - \tilde{\boldsymbol{x}}^{n*}_k\right) + {\boldsymbol{r}^n_k}^T \boldsymbol{G}_k \boldsymbol{r}^n_k \quad \text{(smoothed)}, \\
&\text{with } \tilde{\boldsymbol{x}}^{n*}_k = \boldsymbol{C}^{n*}_k \left[{\boldsymbol{C}_n}^{-1} \tilde{\boldsymbol{x}}_n - {\boldsymbol{A}_k}^T {\boldsymbol{G}_k}^B \boldsymbol{p}_k \right] \\
&\qquad \operatorname{cov}\left(\tilde{\boldsymbol{x}}^{n*}_k\right) \equiv \boldsymbol{C}^{n*}_k = \left({\boldsymbol{C}_n}^{-1} - {\boldsymbol{A}_k}^T {\boldsymbol{G}_k}^B \boldsymbol{A}_k \right)^{-1}
\end{aligned}$$

being the result of *removing* track k from the final vertex fit by an "inverse Kalman filter," formally replacing $\boldsymbol{G}_k \to -\boldsymbol{G}_k$.

In case of no outlier tracks, Gaussian errors well represented by $\boldsymbol{V}_k$, and good linear model, the smoothed $\chi^2_{k,S}$ are chi-square distributed with two degrees of freedom. Thus, they can be used as a symmetric *test for outliers*. However, if there are several outlier tracks (e.g., not belonging to this vertex), the estimates $\tilde{\boldsymbol{x}}_k^{n*}$ are biased by the remaining ones, and the power of the test decreases.

9.3 Robust Vertex Fitting

The *Adaptive Vertex Filter* is a nonlinear generalization of the Kalman filter, iteratively down-weighting the contribution of outlier tracks to the objective function ("soft assignment"). The extra weight $w_k \leq 1$ on each track's weight matrix $\boldsymbol{G}_k = \boldsymbol{V}_k{}^{-1}$ is calculated by a Fermi function of its smoothed chi-square $\chi^2_k \equiv \chi^2_{k,S}$ with a cutoff parameter χ^2_{cut} (Fig. 15).

An annealing schedule with decreasing "temperature" T is introduced in order to avoid local minima (*Deterministic Annealing Filter*, DAF):

$$w_k\left(\chi^2_k, T\right) = \frac{\mathrm{e}^{-\chi^2_k/2T}}{\mathrm{e}^{-\chi^2_k/2T} + \mathrm{e}^{-\chi^2_{\text{cut}}/2T}} = \frac{1}{1 + \mathrm{e}^{\left(\chi^2_k - \chi^2_{\text{cut}}\right)/2T}}$$

Iterations (index $i = 0, \ldots$ omitted above) should start with an a priori guess, e.g., $w_{k,0} = 1$, which is equivalent to the linear Kalman filter. For $i > 0$, $w_{k,i} = w_k\left(\chi^2_{k,i-1}, T_i\right)$ above, with $T_i \leq T_{i-1}$ defined by the annealing schedule. For $T \to 0$, the Fermi function approximates the Heaviside function, and the assignment turns into a "hard" one ($w_k = 1$ or 0 only).

The *Multi-Vertex Filter* is a generalized DAF, simultaneously fitting m vertices by "soft assignment" of each track to more than one vertex. The extra weights $w_{k\ell}$ of track k w.r.t. vertex ℓ in one iteration are

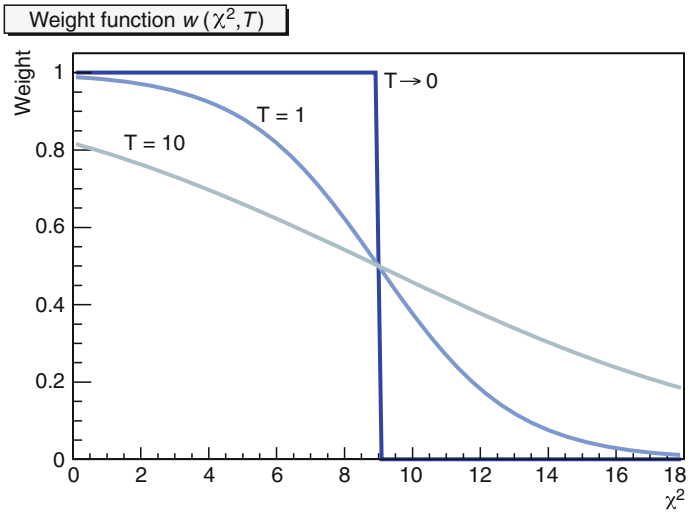


Fig. 15
Example of a Fermi function $w_k\left(\chi^2_k, T\right)$, with cutoff parameter $\chi^2_{\text{cut}} = 9$ and annealing temperatures $T = 10, 1, 0$

$$w_{k\ell}\left(\chi^2_{k\ell}, T\right) = \frac{\mathrm{e}^{-\chi^2_{k\ell}/2T}}{\sum_{\lambda=1}^{m} \mathrm{e}^{-\chi^2_{k\lambda}/2T} + \mathrm{e}^{-\chi^2_{\mathrm{cut}}/2T}}$$

9.4 Helix Coordinate Systems

The tracking detector layout is assumed to be approximately rotationally symmetric w.r.t. the z-axis, but not necessarily mirror symmetric w.r.t. the origin $z = 0$. The axes (x, y, z) define a right-handed orthogonal basis. By convention, the x-axis is chosen to be "horizontal," and the y-axis to point upward.

Surfaces are defined as either cylinders of radius R, or as planes either normal to the z-axis ("end caps"), or parallel to the z-axis ("prism"), or inclined w.r.t. the z-axis ("pyramid"). They may be real or virtual.

Besides Cartesian coordinates, cylindric coordinates, and spherical polar coordinates are defined for space points and/or momenta:

Space points $\mathbf{x} = [x, y, z]_{\mathrm{cart}} = [R, \Phi, z]_{\mathrm{cyl}}$:

$$\begin{aligned} x &= R \cdot \cos\Phi & R &= \sqrt{x^2 + y^2} \\ y &= R \cdot \sin\Phi & \Phi &= \arctan(y/x), \quad \text{azimuth angle } 0 \le \Phi < 2\pi \end{aligned}$$

Momenta $\mathbf{p} = [p_x, p_y, p_z]_{\mathrm{cart}} = [P, \vartheta, \varphi]_{\mathrm{sph}} = [p_T, \varphi, p_z]_{\mathrm{cyl}}$:

$$\begin{aligned} p_x &= P \cdot \sin\vartheta \cdot \cos\varphi & P &= \sqrt{p_x^2 + p_y^2 + p_z^2} = \sqrt{p_T^2 + p_z^2} \\ p_y &= P \cdot \sin\vartheta \cdot \sin\varphi & \vartheta &= \arccos(p_z/P), \quad \text{polar angle } 0 \le \vartheta \le \pi \\ p_T &= P \cdot \sin\vartheta & \lambda &\equiv \tfrac{\pi}{2} - \vartheta, \quad \text{dip angle } \tfrac{\pi}{2} \ge \lambda \ge -\tfrac{\pi}{2} \\ p_z &= P \cdot \cos\vartheta & \varphi &= \arctan(p_y/p_x), \quad \text{azimuth angle } 0 \le \varphi < 2\pi \end{aligned}$$

The *magnetic field* is assumed to be homogeneous and aligned parallel or antiparallel to the z-axis. It is defined by the flux density $\mathbf{B} = [0, 0, B_z]_{\mathrm{cart}}$. This implies a helix track model, with the helix axis being parallel to z.

The following units are most often used: [length] = m (or) cm, [angle] = rad, [momentum] = GeV/c, [B field] = T (Tesla), and [charge] = e (elementary charge). For a particle with momentum P and charge Q, the radius of the helix r_H and its conveniently signed inverse κ are

$$r_H = \frac{1}{K_{\mathrm{unit}}} \cdot \frac{P \cdot \sin\vartheta}{|\, Q \cdot B_z \,|} > 0, \qquad \kappa = -\operatorname{sign}(Q \cdot B_z) \,/\, r_H$$

with the unit-dependent constant (here shown for two units of length)

$$K_{\mathrm{unit}} = (10^{-9}\, \tfrac{c}{\mathrm{m/s}}) \cdot \tfrac{[\mathrm{length}]}{\mathrm{m}}\, \tfrac{\mathrm{GeV}/c}{\mathrm{T{\cdot}[length]}} \approx 0.29979\, \tfrac{\mathrm{GeV}/c}{\mathrm{T{\cdot}m}} = 0.0029979\, \tfrac{\mathrm{GeV}/c}{\mathrm{T{\cdot}cm}}$$

The sign convention corresponds to $\operatorname{sign}(\kappa) = \operatorname{sign}(\mathrm{d}\varphi/\mathrm{d}s) \equiv$ sense of rotation in the (x, y)-projection. Note that in the absence of matter, P and ϑ are constants of motion; in case of multiple scattering, only P remains constant.

The *helix equations* for a starting point $[x_S, y_S, z_S]$ and a starting azimuthal direction angle φ_S, as functions of the running parameter φ, are:

$$\begin{aligned} x(\varphi) &= x_S + (\sin\varphi - \sin\varphi_S) \,/\, \kappa \\ y(\varphi) &= y_S - (\cos\varphi - \cos\varphi_S) \,/\, \kappa \\ z(\varphi) &= z_S + \cot\vartheta \cdot (\varphi - \varphi_S) \,/\, \kappa \qquad \text{path length } s(\varphi) = (\varphi - \varphi_S) \,/\, (\kappa \cdot \sin\vartheta) \end{aligned}$$

A fitter's *internal track parameters* are often defined as follows:

$[R\Phi, z, \vartheta, \varphi - \Phi, \kappa]$	at $R = R_S$	in the radial ("barrel") region
$[x, y, \cot\vartheta, \varphi, \kappa]$	at $z = z_S$	in the forward/backward regions
$[x, y, \mathrm{d}x/\mathrm{d}z, \mathrm{d}y/\mathrm{d}z, Q/P]$	at $z = z_S$	in the extreme fwd/bkwd regions

Examples of popular alternative *external track parameters* are:

$[x, y, z, p_x, p_y, p_z]$	"6D Cartesian" (RAVE, (Waltenberger 2011)); note that the corresponding 6×6 covariance matrix is of rank 5 only
$[\delta_T, \delta_z, \cot\vartheta, \varphi_P, \kappa]$	"Perigee representation"

The *perigee point* $[x_P, y_P, z_P]$ of a helix track is defined, in the (x, y)-projection, as the "point of closest approach" (PCA) to a fixed pivot point $[x_0, y_0, z_0]$. The track parameters in perigee representation and usual convention are:

$\delta_T = \pm\sqrt{(x_P - x_0)^2 + (y_P - y_0)^2}$	projected distance between perigee and pivot points (transverse impact parameter), with + or − sign indicating the pivot point sitting to the left or to the right of the helix, respectively;
$\delta_z = z_P - z_0$	distance along z between perigee and pivot points;
$\cot\vartheta$	slope of the helix;
φ_P	azimuthal direction of the helix at the perigee point;
κ	inverse helix radius, with the sign defined as before.

Beware of subtle differences in the various conventions, e.g., for the units of length and momentum, the sign definitions for κ and δ_T, replacement of φ_P by the azimuthal position of the perigee point w.r.t. the pivot point, or an assumption about $\mathrm{sign}(B_z)$ with implicit consequences.

Disclaimer

The formulae of ❯ Sect. 9 are field proven. However, the authors deny any responsibility for possible damages caused by their usage.

Acknowledgment

The authors wish to thank *Rudolf Frühwirth* and *Meinhard Regler* (HEPHY, Vienna) for substantial contributions to ❯ Sects. 5.3, ❯ 6.1, and ❯ 7, for most valuable discussions, and for a careful reading of the manuscript.

References

Abashian A et al (2002) Nucl Instrum Methods A479:117, Belle Collaboration

Abramowitz M, Stegun IA (eds) (1965) Handbook of mathematical functions, section 25.5. Dover, Mineola

Abreu P et al (1996) Nucl Instrum Methods A378:57, DELPHI Collaboration

Achenbach P et al (2008) Nucl Instrum Methods A593:353

Adam W et al (2005) J Phys G: Nucl Part Phys 31:N9

Adam W et al (2009a) J Instrum 4:P06009

Adam W et al (2009b) J Instrum 4:T07001

Adinolfi M et al (2002) Nucl Instrum Methods A478:138

Adorisio C et al (2009) Nucl Instrum Methods A598:400
Agostinelli S et al (2003) Nucl Instrum Methods A506:250
Alfonsi M et al (2004) Nucl Instrum Methods A518:106
Altunbas C et al (2002) Nucl Instrum Methods A490:177
Anderson M (2003) Nucl Instrum Methods A499:659
Arneodo F et al (2003) Nucl Instrum Methods A498:292
Arrabito L et al (2007) J Instrum 2:P05004
Avery P (1999) Applied fitting theory VI: formulas for kinematic fitting. Report CBX 98 V-37. University of Florida, Gainesville
Bagliesi MG et al (2010) Nucl Instrum Methods A617:134
Beckmann M, List B, List J (2010) Nucl Instrum Methods A624:184
Beischer B et al (2010) Proceedings of the 12th Vienna conference on instrumentation. Nucl Instrum Methods A (to be published)
Beiser A (1952) Rev Mod Phys 24:273
Bethe H, Heitler W (1934) Proc R Soc Lond A 146:83
Blau M (1961/1963) In: Yuan LCL, Wu C-S (eds) Methods of experimental physics, vol 5, Nuclear physics. Academic, New York/London
Blobel V (2006) Nucl Instrum Methods A566:5
Blobel V (2007) Millepede II: the manual. University of Hamburg, Hamburg
Blobel V, Kleinwort C (2002) In: Proceedings of the conference on advanced statistical techniques in particle physics, Durham University, Durham
Brown DN et al (2009) Nucl Instrum Methods A603:467
Charpak G et al (1968) Nucl Instrum Methods 62:262
Chatrchyan S et al (CMS Collaboration) (2008) J Instrum 3:S08004
Colas P (2004) Nucl Instrum Methods A535:181
De Gerone M et al (2009) Nucl Instrum Methods A610:218
Derré J et al (2001) Nucl Instrum Methods A459:523
Doležal Z, Uno S (eds) (2010) The Belle II technical design report. KEK, Tsukuba
Dominguez A (2007) Nucl Instrum Methods A581:343
Eskut E et al (1997) Nucl Instrum Methods A401:7
Evans L, Bryant P (eds) (2008) J Instrum 3:S08001
Flanagan JW, Ohnishi Y (eds) (2004) Letter of intent for KEK Super B Factory: accelerator design. KEK report 2004-4, part 3. KEK, Tsukuba
Frühwirth R (1987) Nucl Instrum Methods A262:444
Frühwirth R, Kubinec P, Mitaroff W, Regler M (1996) Comput Phys Commun 96:189
Frühwirth R, Regler M (2001) Nucl Instrum Methods A456:369
Frühwirth R, Regler M, Bock RK, Grote H, Notz D (2000) (Eds. Regler M, Frühwirth R) Data analysis techniques for high-energy physics, 2nd edn. Cambridge University Press, Cambridge, UK
Frühwirth R, Strandlie A (1999) Comput Phys Commun 120:197
Frühwirth R, Waltenberger W (2004) CMS conference report CR 2004/062. CERN, Geneva
Giomataris Y et al (1996) Nucl Instrum Methods A376:29
Gluckstern RL (1963) Nucl Instrum Methods 24:381
Gorbunov S, Kisel I (2006) Nucl Instrum Methods A559:139
Hartmann F (2009) Evolution of silicon sensor technology in particle physics. Springer, Berlin
Highland V (1975) Nucl Instrum Methods 129:479
Höppner C, Neubert S, Ketzer B, Paul S (2010) Nucl Instrum Methods A620:518
Hough PVC (1959) In: Proceedings of the international conference on high energy accelerators and instrumentation, CERN, Geneva
Hübner K (2004) Phys Rep 403–404:177
Hyams B et al (1983) Nucl Instrum Methods 205:99
Ivaniouchenkov Y et al (1999) Nucl Instrum Methods A422:300
Jackson DJ (1997) Nucl Instrum Methods A388:247
Jacob MRM, Quercigh E (eds) (1997) Symposium on the CERN Omega Spectrometer: 25 Years of Physics. CERN-97-02. CERN, Geneva
Kartvelishvili V (2007) Nucl Phys B: Proc Suppl 172:208
Ketzer B (2002) Nucl Instrum Methods A494:142
Kim BJ et al (2003) Nucl Instrum Methods A497:450
Kisel I, Konotopskaya E, Kovalenko V (1997) Nucl Instrum Methods A389:167
Klingenberg R (2007) Nucl Instrum Methods A579:664
Kobayashi M (2007) Nucl Instrum Methods A581:265
Kurokawa S, Kikutani E (2003) Nucl Instrum Methods A499:1
Larsen DT (2010) Nucl Instrum Methods A617:35
Mankel R (1997) Nucl Instrum Methods A395:169
Morley A (2008) Nucl Instrum Methods A596:32
Moser F, Waltenberger W, Regler M, Mitaroff W (2008) In: Proceedings of the international linear collider workshop, University of Illinois, Chicago. arXiv:0901.4020
Oed A (1988) Nucl Instrum Methods A263:351
Ortner G, Stetter G (1928) Mitteilungen Inst. Radiumforschung, No. 228, Wien
Phinney N, Toge N, Walker N (eds) (2007) The international linear collider reference design report: accelerator. ILC report 2007-001, vol 3. DESY, Hamburg/FNAL, Batavia IL/KEK, Tsukuba

Qin ZH et al (2007) Nucl Instrum Methods A571:612
Regler M, Frühwirth R (2008) Nucl Instrum Methods A589:109
Regler M, Mitaroff W, Valentan M, Frühwirth R, Höfler R (2008a) Proceedings of the international conference on computing in high energy and nuclear physics (Victoria, BC 2007). J Phys: Conf Series 119:032034
Regler M, Valentan M, Frühwirth R (2008b) LiC detector toy user's guide, PUB-863/08. HEPHY, Vienna. http://stop.itp.tuwien.ac.at/websvn/
Reingamum M (1911) Phys Z 12:1076
Riegler W et al (2000) Nucl Instrum Methods A443:156
Sauli F (1997) Nucl Instrum Methods A386:531
Sauli F, Sharma A (1999) Annu Rev Nucl Part Sci 49:341
Speer T, Frühwirth R (2006) Comput Phys Commun 174:935
Stoeck H et al (ILD Concept Group) (2010) The International Large Detector letter of intent. DESY 2009-87/FNAL PUB-09-682-E/KEK 2009-6, revised
Strandlie A, Frühwirth R (2010) Rev Mod Phys 82:1419
Thomson M (2009) Nucl Instrum Methods A611:25
Titov M (2007) Nucl Instrum Methods A581:25
Urquijo P, Barberio E (2005) Belle note 756, KEK, Tsukuba
Valentan M, Regler M, Frühwirth R (2009) Nucl Instrum Methods A606:728
Waltenberger W (2008) CMS note 2008/033. CERN, Geneva
Waltenberger W (2011) RAVE – a detector-independent toolkit to reconstruct vertices. IEEE Trans Nucl Sci (accepted)
Waltenberger W, Frühwirth R, Vanlaer P (2007) J Phys G: Nucl Part Phys 34:N343
Widl E, Frühwirth R (2008) Proceedings of the international conference on computing in high energy and nuclear physics (Victoria, BC, 2007). J Phys: Conf Series 119:032038
Zerguerras T et al (2007) Nucl Instrum Methods A581:258

Further Reading

Frühwirth R, Regler M, Bock RK, Grote H, Notz D (2002) (Eds. Regler M, Frühwirth R) Data analysis techniques for high-energy physics, 2nd edn. Cambridge University Press, Cambridge, UK
Grupen C, Shwartz B (2008) Particle detectors, 2nd edn. Cambridge University Press, Cambridge, UK
Hartmann F (2009) Evolution of silicon sensor technology in particle physics. Springer, Berlin
Leroy C, Rancoita P-G (2009) Principles of radiation interaction in matter and detection, 2nd edn. World Scientific, Singapore
Strandlie A, Frühwirth R (2010) Track and vertex reconstruction: from classical to adaptive methods. Rev Mod Phys 82:1419

Software Portals (Selection)

CEDAR HepForge: http://www.hepforege.org/
CERN Physics Software: http://sftweb.cern.ch/sft/
DESY ILC Software: http://ilcsoft.desy.de/portal/
FreeHEP Software: http://www.freehep.org/
Open Directory: http://www.dmoz.org/Science/Physics/Particle/Software/
Scientific Computing in OO: http://www.oonumerics.org/
SLAC ILC Software: http://lcsim.org/software/

13 Photon Detectors

Peter Križan
University of Ljubljana, Ljubljana, Slovenia

1 Introduction ***298***
1.1 General Properties of Photon Detectors 298

2 Vacuum Photodetectors ***299***
2.1 Photomultiplier Tubes 299
2.2 Microchannel Plate Photomultiplier Tube 302
2.3 Hybrid Photodetectors 303

3 Gaseous Photon Detectors ***305***

4 Solid-State Photon Detectors ***306***

Acknowledgments ***310***

References ***310***

Further Reading ***311***

Suppliers of Technology ***311***

C. Grupen, I. Buvat (eds.), *Handbook of Particle Detection and Imaging*,
DOI 10.1007/978-3-642-13271-1_13, © Springer-Verlag Berlin Heidelberg 2012

Abstract: This chapter is intented as an overview of techniques for detection of photons, from the infrared to the ultraviolet and extreme ultraviolet. We discuss the vacuum photon detectors, gaseous photon detectors, and semiconductor sensors, as well as recent progress in novel photon detection methods.

1 Introduction

A large fraction of detectors in particle, nuclear, and astrophysics, as well as in medical imaging is based on the detection of photons in or near the visible range, 100 nm $< \lambda <$ 1,000 nm. This includes detection of scintillation and Cherenkov light as well as the light detected in astronomical observations.

In most photosensors, detection of photons proceeds in three steps. First, a primary photoelectron or electron–hole (e–h) pair is generated by an incident photon via the photoelectric or the photoconductive effect. By electron multiplication the number of electrons is increased to a detectable level, so that finally secondary electrons produce an electrical signal. The three steps are best illustrated on the example of the photomultiplier tube (PMT), the most common light sensor with a simple structure of electrodes enclosed in an evacuated glass vessel (*Fig. 1*). The photon hits a semitransparent photocathode, in which its energy is transferred to an electron. This photoelectron exits the photocathode and enters an electric field, which leads it to the first dynode. The dynodes serve as an electron multiplier chain; electrons are accelerated in the electric field, strike the next dynode, releasing more electrons. After several amplification stages, the swarm of electrons hits the last electrode, the anode, where a detectable current signal is produced.

1.1 General Properties of Photon Detectors

Photon detectors are charaterized by the following properties (Particle Data group 2009):

Quantum efficiency (QE or ϵ_q): the probability that the incident photon generates a photoelectron

Collection efficiency (CE or ϵ_c): the probability that the photoelectron starts the electron amplification

Gain (G): the number of electrons collected for each generated photoelectron

Dark current or dark noise: the electrical signal with no photon at the input

Dynamic range: the maximum signal available from the detector (usually expressed in units of the response to noise-equivalent power, or NEP, which is the optical input power that produces a signal-to-noise ratio of 1)

Time dependence of the response: this includes the transit time, which is the time between the arrival of the photon and the electrical pulse, and the transit time spread (TTS), which contributes to the pulse rise time and duration

Single-photon detection capability: important when measuring very-low-level light intensities

Rate capability: inversely proportional to the time needed, after the arrival of one photon, to get ready to receive the next (i.e., recovery time)

Stability: essential for long-term operation at elevated counting rates

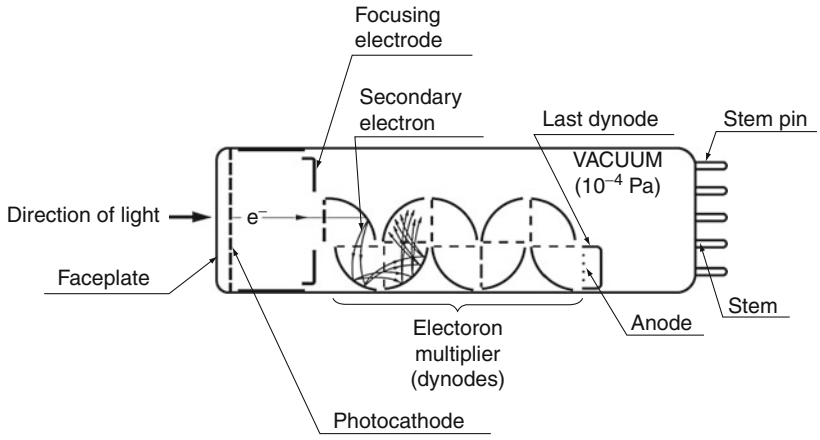


Fig. 1
Schematics of a transmission-type photomultiplier tube (Hamamatsu 2006)

Note that the term *photon detection effiency* (PDE) is often used for the combined probability to produce a photoelectron and to detect it (PDE = $\epsilon_q \epsilon_c$). Note also that producers often quote the cathode radiant sensitivity S_k, the ratio between the photocathode current to the incoming light power; S_k is related to the quantum efficiency through $S_k = \epsilon_q e/h\nu$.

2 Vacuum Photodetectors

In a vacuum photodetector, the photocathode and the electron multiplication stage are in vacuum, enclosed in a vessel made of various combinations of glass, ceramics, and metal. Vacuum photodetectors are of three types: photomultiplier tubes, microchannel plate photomultiplier tubes, and hybrid photodetectors.

2.1 Photomultiplier Tubes

Photomultiplier tubes (PMTs) (Fig. 1) have been up to very recently the most common photodetector in particle physics experiments and medical imaging (Arisaka 2000). In a transmission-type PMT, the photosensitive material (photocathode) is on the inside of a transparent window through which the photons enter (Fig. 1); in a reflection-type PMT, the photocathode material is deposited on a separate electrode inside the tube. The photosensitive material has a suitably low work function, such that when a photon hits the photocathode, an electron is produced in a photoelectric effect. This photoelectron is then accelerated and guided by the electric field to hit the next electrode (dynode), which is covered with a material suitable for high secondary emission. A single electron therefore produces a few (≈5) secondary electrons, and this process is repeated at subsequent dynodes, typically 10 altogether. By this electron multiplication process, 10^5–10^6 electrons are generated and collected at the anode. The total gain of a PMT depends on the applied high voltage U as $G = AU^{kn}$, where $k \approx 0.7$–0.8 (depending on the dynode material), n is the number of dynodes in the chain, and A is a constant (which also depends on n).

The sensitive wavelength range of the PMT is determined by the choice of the photocathode material. These materials are usually Cs- and Sb-based compounds such as CsI, CsTe, bialkali (SbRbCs, SbKCs), multialkali ($SbNa_2KCs$), as well as GaAs(Cs), GaAsP, etc. From the wavelength dependence of the quantum efficiency for these materials as displayed in *Fig. 2*, we observe that they cover a wide range of wavelengths, from the infrared (IR) to the extreme

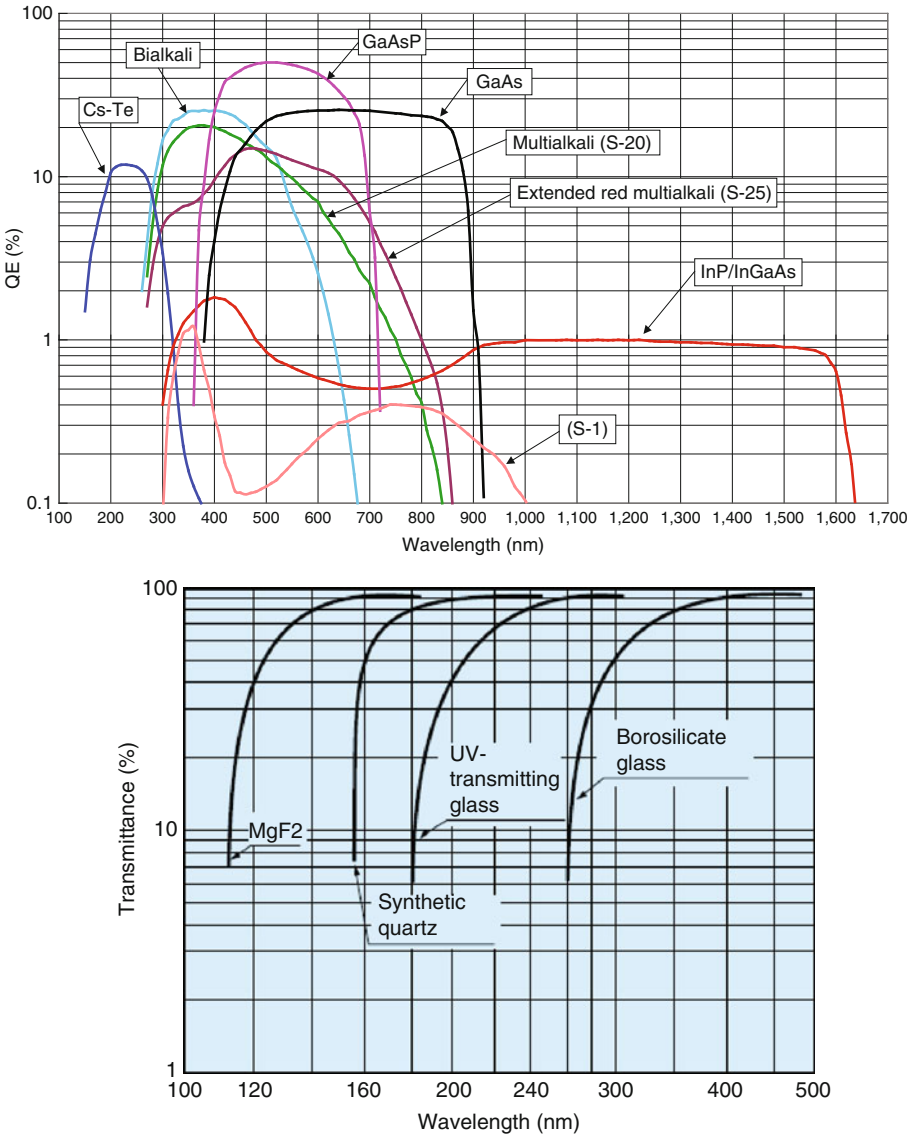


Fig. 2

Quantum efficiency for different photocathode materials (*top*), window trasmission curves (*bottom*) (Hamamatsu 2006)

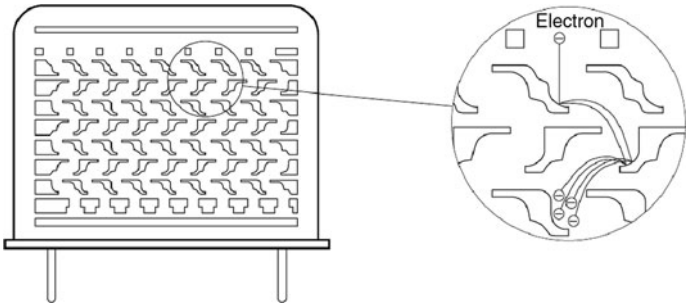


Fig. 3
Metal-foil dynode structure (Hamamatsu 2006)

ultraviolet (XUV) (Hamamatsu 2006). Note that recently new improved versions of bialkali photocathodes became available with a considerably higher quantum efficiency. They are known as super-bialkali (SBA) and ultra-bialkali (UBA) photocathodes, with peak efficiencies as high as 35% and 45%, respectively (Nakamura et al. 2010). The low-wavelength cutoff is usually given by the choice of the window material (*Fig. 2*). Common window materials are borosilicate glass for IR to near-UV, fused quartz and sapphire (Al_2O_3) for UV, and MgF_2 or LiF for XUV.

Typical dynode structures used in PMTs are box and grid (*Fig. 1*), circular cage, line focusing, venetian blind, fine mesh, and metal foil (*Fig. 3*). The choice of the structure depends on the use. While line focusing allows for a very fast response, the metal-foil structure is employed in position-sensitive PMTs with multichannel anode granularities as fine as $\approx 2 \times 2\,\mathrm{mm}^2$ (Hamamatsu Photonics). The transit time spread is typically a few hundred picoseconds for line-focusing, fine-mesh and metal-foil dynode structures, and a few nanoseconds for PMTs with box-and-grid, circular-cage, and venetian-blind dynodes.

Voltages are distributed to the dynodes by a resistive voltage divider chain so that the most negative high voltage is supplied to the photocathode. The divider chains sometimes include capacitors in the higher amplification stages, or even active elements, transistors, or diodes.

When a photomultiplier is used to detect very low light intensities, it is often advantageous to count single-photon pulses. Because of the nature of the secondary emission process, single-photoelectron pulses show large fluctuations. Some PMT types allow for an efficient single-photon detection, an example of which is shown in *Fig. 4*, while for some (like fine-mesh PMTs) no single-photon detection is possible because of a large variation in the gain.

PMTs are affected by the presence of magnetic fields. The magnetic field deflects electrons from their normal trajectories and causes orientation-dependent loss of efficiency and gain; these effects are in some cases observed even in the geomagnetic field. In particular PMTs with a long path from the photocathode to the first dynode are very vulnerable. A high-permeability metal shield is often necessary. However, PMTs with the fine-mesh dynode chain can be used even in a high magnetic field ($\approx$1 T) if the tube axis is aligned (or partially aligned) with the field, although the gain decreases considerably (Iijima et al. 1997).

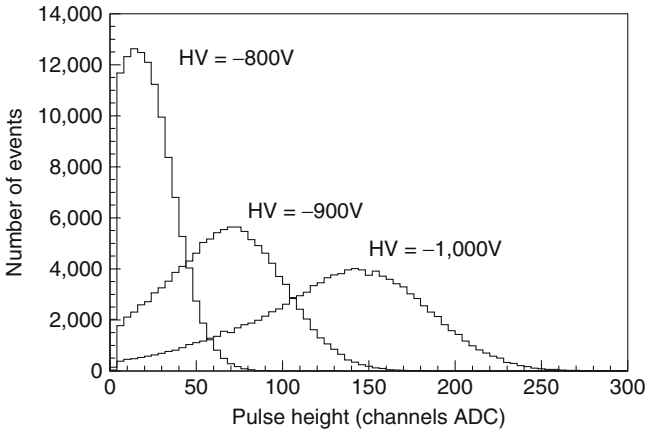


Fig. 4
Pulse-height distributions due to single photoelectrons as detected by a multi-anode R5900-M16 photomultiplier tube with three different cathode high voltages (Križan et al. 1997)

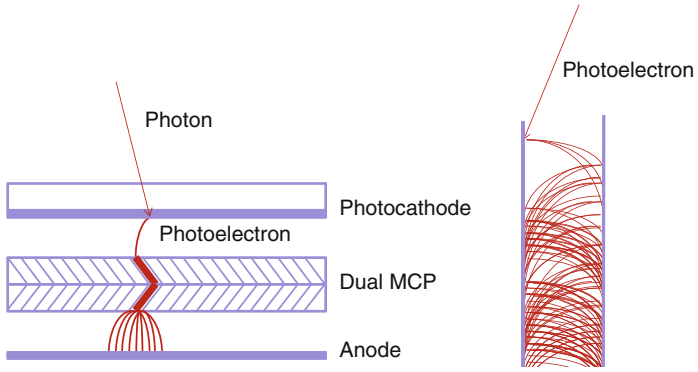


Fig. 5
Microchannel plate photomultiplier tube (*left*), electron multiplication in a channel (*right*) (Flyckt and Marmonier 2002)

2.2 Microchannel Plate Photomultiplier Tube

In a microchannel plate (MCP) photomultiplier tube, the discrete dynode chain is replaced by continuous multiplication in 6–20 μm-diameter cylindrical holes, or "channels" (*Fig. 5*). These channels are densely packed in a few mm thick glass plate. The inner channel walls are

processed to have proper electrical resistance and secondary-emission properties. The multiplication gain depends exponentially on the ratio of the channel length to its diameter. A typical value of this ratio is 40, and typical gains of a single microchannel plate are about 10^4. For higher gains, two or more microchannel plates are used in series.

MCP PMTs are thin: the gaps between the transmission-type photocathode and the microchannel plates, and between the microchannel plates and the anode plane are only a few mm thick. MCP PMTs offer good spatial resolution, have excellent time resolution (≈20 ps), and can tolerate random magnetic fields up to 0.1 T and axial fields exceeding 1 T. However, they suffer from a relatively long recovery time per channel and short lifetime. The performance of this sensor type may degrade due to residual gas in the channels, which gets ionized, and the feedback ions hit and gradually destroy the photocathode. Most modern MCP-based photodetectors consist of two microchannel plates with angled channels rotated with respect to each other producing a chevron (v-like) shape, as illustrated in *Fig. 5*. The angle between the channels reduces ion feedback in the device. In addition, in the last few years, further improvement was achieved in the lifetime of MCPs by improving the vacuum in the vessel and by a protective aluminum foil (Hamamatsu 2006; Flyckt and Marmonier 2002).

While MCPs have been widely employed as image intensifiers, they have not been used in particle physics or astrophysics. Recently, however, the need to perform Cherenkov imaging within magnetic spectrometers, combined with requests of excellent timing, has considerably increased the interest in this sensor type; a large-scale use is planned for the Cherenkov detectors of the Belle II and PANDA experiments (Abe et al. 2004; Foehl et al. 2008; Inami 2008).

2.3 Hybrid Photodetectors

Hybrid photon detectors (HPDs) combine the sensitivity of a vacuum PMT with the excellent spatial and energy resolutions of a Si sensor (Braem et al. 2004). The photoelectron is accelerated through a potential difference of ≈ 10–20 kV before it hits the silicon sensor, as shown in *Fig. 6*. The electric field is usually shaped in such a way that the entry window is demagnified onto the silicon sensor (*Fig. 6*); by this, the Si sensor can be kept smaller than the window, and the active area fraction can be very high. Because the silicon sensor is segmented, these devices are naturally of multichannel type, with a very flexible segmentation design.

The gain roughly equals the maximum number of e–h pairs that could be created from the kinetic energy of the accelerated electron: $G \approx e(U - U_{th})/w$, where U is the applied potential difference, $w \approx 3.7$ eV is the mean energy required to create an e–h pair in Si at room temperature, and eU_{th} is the energy lost by the photoelectron in the insensitive layer of the silicon; this relation is valid for U considerably higher than U_{th}. The variations of the gain are small since it is achieved in a single step, and are further reduced due to the Fano effect discussed in Chap. 17, "Gamma-Ray Detectors". The excellent single-photon resolution is displayed in *Fig. 7*.

The gain is smaller than in a typical PMT, and low-noise electronics must be used to read out HPDs when counting small numbers of photons. HPDs also require rather high biases, and do not function in a magnetic field. An exception is the proximity-focused device (with no demagnification of the entry window) in an axial field; the performance of such a

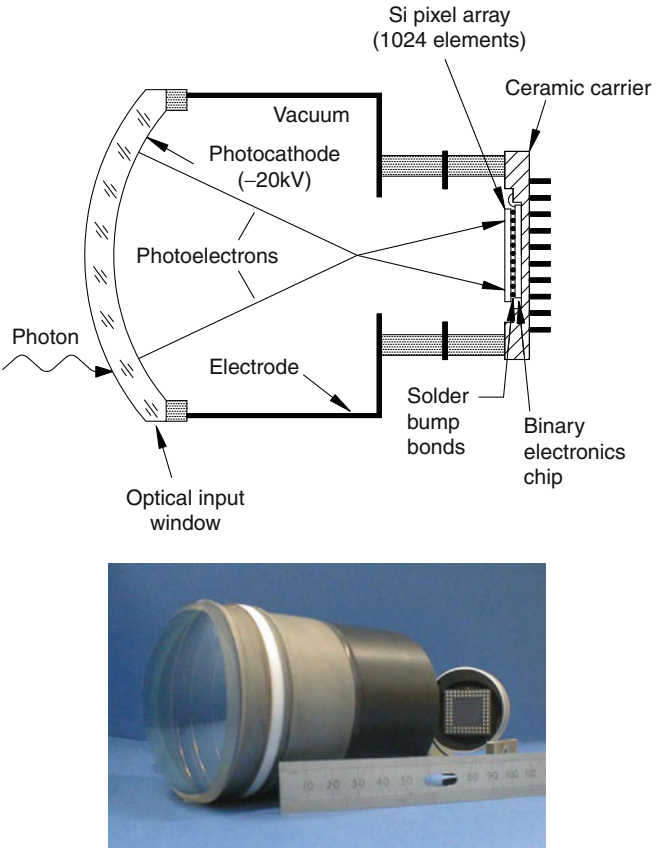


Fig. 6

Hybrid photon detector (HPD) of the LHCb experiment. The outer diameter of the detector is 17 cm (Eisenhardt 2006; Alves Jr et al. 2008). The large photocathode surface is imaged onto a much smaller Si sensor by a suitably shaped electric field

sensor is actually improved if operated in an axial magnetic field because of reduced photoelectron back-scattering effects and reduced influence of electric field inhomogeneities at the edges of the sensor. This sensor type also has excellent timing, with typical time resolution of ≈50 ps.

Large-scale applications include the CMS hadronic calorimeter and the RICH detector in LHCb. Large-size HPDs with sophisticated focusing are also considered for future very-large-scale water Cherenkov experiments.

In hybrid APDs (HAPDs) the silicon sensor is operated in the avalanche mode; with this additional multiplication step the gain is increased by a factor of ≈50. This allows for operation at a higher gain or at a lower supply voltage, but also degrades the signal-to-noise characteristics. A 144-channel proximity-focused HAPD has been developed for a RICH counter of the Belle II experiment (Abe et al. 2004; Nishida et al. 2009) with very good single-photon detection sensitivity.

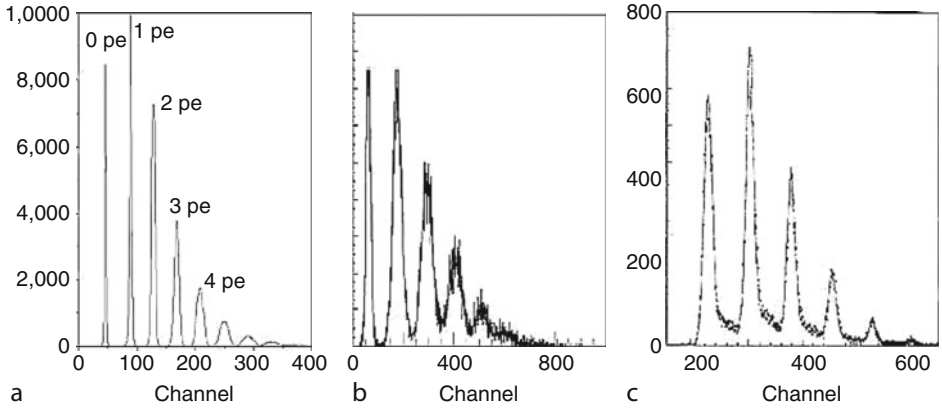


Fig. 7

Pulse-height spectrum of very-low-intensity light pulses recorded with a Geiger-mode APD (*left*), visual-light photon counter (VLPC, *middle*), and an HPD (Bross et al. 2002; Buzhan et al. 2003)

3 Gaseous Photon Detectors

In gaseous photodetectors, the electron multiplication step happens in an avalanche in the high-field region of a gaseous detector, in the same way as in gaseous tracking detectors such as multiwire proportional chambers, time projection chambers, micromesh gaseous detectors (Micromegas), or gas electron multipliers (GEM) (as discussed in ❯ Chap. 11, "Gaseous Detectors").

Photoelectrons are generated either on a photosensitive component of the gas mixture or on a solid photocathode material similarly as in a PMT (❯ *Fig. 8*). Since one of the cathodes of the gaseous photon detector can be structured in pads of few mm size, these devices can be employed as position-sensitive photon detectors. Just like tracking chambers, they can be made to cover large areas (several m^2). They can operate in high magnetic fields, and are relatively inexpensive. Their drawback is that they are only sensistive in the UV and XUV region (❯ *Fig. 8*).

As a solid photocathode material, ≈0.5 μm thick layer of CsI is commonly used since it is stable in gas mixtures typically employed in single-electron detection (e.g., CH_4, CF_4). Among photosensitive gas components, vapors of TMAE (tetrakis dimethyl-amine ethylene) or TEA (tri-ethyl-amine) have the highest thresholds at 230 nm (TMAE) and 165 nm (TEA), and can thus be used for UV and XUV photon detection (Arnold et al. 1992) (❯ *Fig. 8*).

In gaseous photon detectors, special care must be taken to suppress the photon-feedback process, that is, effects due to photons produced in an avalanche. It is also important to maintain high purity of the chamber gas since O_2 at concentrations exceeding a few ppm can significantly degrade the detection performance and long-term stability. Note also that TMAE and TEA are chemically agressive, so that special care has to be taken in the choice of materials for the chambers and supply tubing.

Most of the ring imaging Cherenkov (RICH) detectors of the first generation have used gaseous photon detectors for the detection of Cherenkov light. TEA was first used as the

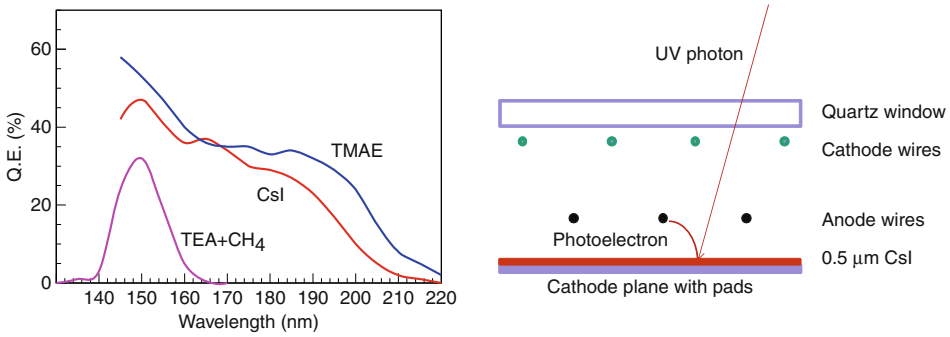


Fig. 8

Gaseous photon detectors: quantum efficiency for photosensitive substances used in gaseous photodetectors (*left*), typical wire-chamber geometry (*right*)

photosensitive component in the pioneering RICH experiment E605 (Mangeot et al. 1983). Later, it was considered as the RICH photon detector at B-factories, first as a proximity focusing RICH prototype (Seguinot et al. 1994), and then later successfully employed for the CLEO spectrometer (Artuso et al. 2005), with LiF as the solid radiator and CaF_2 as a UV-transparent wire-chamber window. TMAE was employed in the OMEGA, DELPHI, and SLD experiments (Apsimon et al. 1986; Arnold et al. 1988; Aston et al. 1989). The RICH counters of COMPASS, Hades, RICH at the JLAB-Hall A and ALICE experiments employ multiwire chambers with a solid CsI photocathode (Fabbietti et al. 2003; Albrecht et al. 2005; Iodice et al. 2005; Di Mauro et al. 2005). Robust gas-based photon detectors remain an attractive alternative for applications where large-area detectors are needed which have to operate in high magnetic fields. One of the recent developments is for the upgrades of RICH counters of the COMPASS and ALICE experiments. It uses the thick GEM, THGEM, as the electron amplification component; THGEM is made photosensitve by a CsI coating (Breskin et al. 2009).

An ideal photocathode should be sensitive for visual light rather than in the UV region, but the stability of such visual-light-sensitive photocathodes in a gas atmosphere has been a serious problem. Very recently, however, laboratory-produced K_2CsSb, Cs_3Sb, and Na_2KSb photocathodes have been found to be quite stable in a gas chamber (Chechik and Breskin 2008; Lyashenko et al. 2009). In an Ar/CH_4 (95/5) gas mixture, K_2CsSb photocathodes yielded quantum efficiency values above 30% at wavelengths between 360 nm and 400 nm.

4 Solid-State Photon Detectors

In a solid-state photodetector, production and detection of photoelectrons take place in the same thin material (*Fig. 9*). Solid-state photodetectors are a special sort of semiconductor detector, which is discussed in great detail in Chap. 16, "Semiconductor Counters".

In a silicon photodiode, photons with wavelengths shorter than about 1050 nm (i.e., with energies exceeding the bandgap of 1.12 eV) can create electron–hole pairs by the photoconductive effect. In its simplest form, the photodiode is a reverse-biased p–n junction (as shown in *Fig. 9*), and the electrons and holes are collected on the n and p sides, respectively. The same

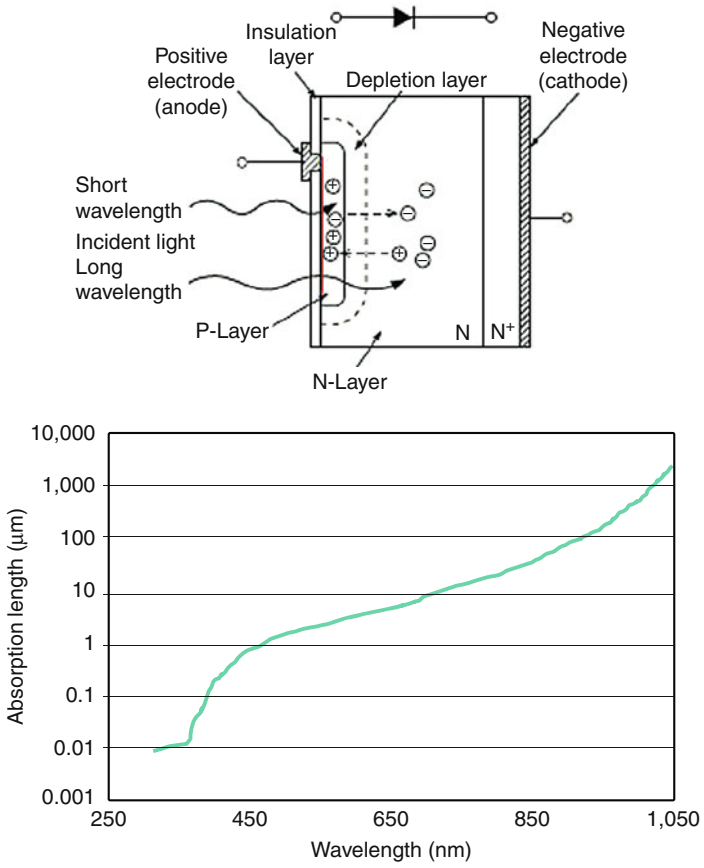


Fig. 9
Schematic view of a photodiode (*top*), photon absorption length in silicon at 77 K (*bottom*) (Renker and Lorenz 2009)

structure is widely used in high-energy physics as particle detectors and in a great number of applications. (Note that a solar cell is a large-area photodiode run with zero external bias.) In a PIN diode, intrinsic silicon is doped to create a p–i–n structure. The reverse bias increases the thickness of the depleted region; typically, the full depletion depth is about 100 µm. This has two benefits, decrease of electronics noise due to a lower capacitance, and an improved sensitivity at higher wavelengths where the absorption length is comparable to the thickness of the sensitive region (*Fig. 9*). The quantum efficiency can exceed 90%, but decreases toward longer wavelengths because of the increasing absorption length of light in silicon. The efficency is also reduced for low wavelengths when the photon absorption length becomes comparable to the thickness of the insensitive surface layer. Since there is no electron multiplication ($G = 1$), amplification of the induced signal is necessary. Low-light-level signal detection is limited to a few hundred photons even if slow low-noise amplifiers are used. This sensor type can therefore be used where enough light is available, typically for crystal scintillator readout like in the CLEO, L3, Belle, BaBar, and GLAST calorimeters.

Compared to traditional vacuum-based photodetectors, solid-state sensors have several advantages. They are compact and robust, do not require high voltages, and are insensitive to magnetic fields. Another advantage is the low cost because of the relatively cheap production methods. They can be pixelized and integrated into large systems. Arrays of tens of millions of pixels, photodioides at a few μm pitch, have to a large degree replaced the photographic film and plates in photography including the applications in astronomy. To read such large arrays, usually the charge-coupled device (CCD) scheme is employed by which the large number of pixels is read out by a much smaller number of electronic channels; after the exposure, the signal charge is transferred from pixel to pixel by using them as shift registers (Janesick 2001). The readout speed can be considerably increased in Active Pixel Sensor (APS) arrays with the first amplifying stage on each pixel.

Avalanche photodiodes (APDs) have a similar structure to regular photodiodes, but have a different doping profile and are operated with much higher reverse bias. This allows each photogenerated carrier to get multiplied in an avalanche, resulting in internal gain (typically $G = 10–200$) within the photodiode (Haitz et al. 1965; McIntyre 1966; Dautet et al. 1993; Perkin-Elmer Optoelectronics). As a result, a detectable electrical response can be obtained from low-intensity optical signals, as low as 10–20 photons. Well-designed APDs, such as those used in the crystal-based electromagnetic calorimeter of the CMS experiment, have achieved a photon detection efficiency of ≈0.7 with sub-ns response time (Deiters et al. 2000). The sensitive wavelength window and the gain depend on the semiconductor used. Stability and monitoring of the operating temperature are important for the linear-mode operation (e.g., when used in calorimeters), and cooling is often necessary.

Visible-light photon counter (VLPC) is a special kind of APD. By a very high donor concentration (As-doped Si) an impurity band is created 50 meV below the conduction band (Petrov et al. 1987; Atac and Petrov 1989; Atac et al. 1994). The device is charaterized by a high gain ($G \approx 5 \cdot 10^4$), high efficiency ($\epsilon_q \approx 0.9$), and very good single-photon sensitivity as illustrated in *Fig. 7*. The small gap makes the sensor, however, sensitive to infrared photons, and requires operation at cryogenic temperatures. The D0 detector at the Tevatron collider at Fermilab is the only experimental apparatus with a large-scale application of this sensor type; 86,000 VLPCs are employed to read out the optical signals from a scintillating-fiber tracker and scintillator-strip preshower detectors (Abachi et al. 1994).

Very-low-light-level detection (including single photons) is possible with yet another type of APD, operated in the limited Geiger mode (Buzhan et al. 2001, 2003a, b; Sadygov et al. 2003; Golovin and Saveliev 2004) with gains of $G \approx 10^6$. This device type is known as G-APD (Geiger mode APD), but other names are used as well (SiPM for silicon photomultiplier, PPD for pixelated photon detector). G-APDs typically cover an area of ≈ 1–10 mm^2 and consist of 100–1,000 APD cells (*Fig. 10*). All cells are read out by a single readout channel. Since each cell provides a binary output, the sensor output is a sum of cell outputs, proportional to the number of incoming photons as shown in *Fig. 7*. (Once the number of detected photons becomes comparable to the number of cells, a correction factor has to be applied to relate the numbers of detected and incoming photons.) G-APDs are also fast devices, with a time resolution of 100–200 ps for single photons. Another advantage for practical use is the moderate bias voltage of ≈50 V. However, the single-photoelectron noise of a G-APD is rather large, 0.1–1 MHz/mm^2 at room temperature.

G-APDs are particularly well-suited for detector systems where triggered pulses of several photons are expected over a small area, for example, fiber-guided scintillation light. Large-scale applications include the T2K experiment (Yokoyama et al. 2009) and the CALICE collaboration

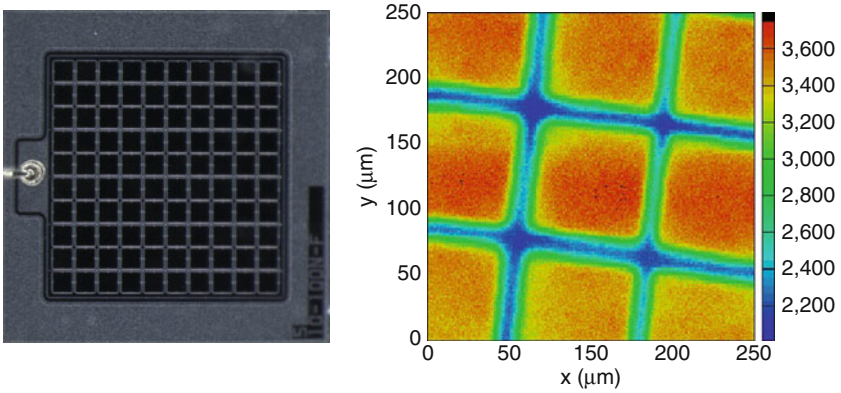


Fig. 10
Geiger-mode APDs: photograph of a 1 mm^2-sized sensor (*left*), sensitivity scan over 100 μm x 100 μm cells (*right*)

calorimeter prototype for the linear-collider detector (Danilov 2007). G-APDs are considered as an excellent alternative to the photomuliplier tubes in several other applications including medical imaging, in particular positron emission tomography (PET) (Chap. 38, "PET Imaging: Basics and New Trends").

G-APDs are in principle also a very promising candidate for detectors of Cherenkov photons in a RICH counter. However, due to the serious disadvantage of a very high rate of noise with a pulse-height distribution equal to the one for single photons, they have up to now not been used in ring imaging Cherenkov detectors, where single-photon detection is required at low noise. Because Cherenkov light is prompt, this problem could in principle be reduced by using a narrow time window (a few ns) for signal collection. To test such a device in a RICH counter, a study was carried out with cosmic rays, and led to the first detection of single Cherenkov photons with this sensor type (Korpar et al. 2008). Light concentrators, gathering light from a larger area and concentrating it onto the G-APD sensitive surface (*Fig. 11*), could increase the number of detected photons while conserving the dark count rate, and therefore improving the signal-to-noise ratio. A prototype with 64 sensors with light guides as shown in *Fig. 11* was indeed successfully employed in a pion test beam (Korpar et al. 2010).

When G-APDs are used in the hostile environments of high-energy physics experiments, radiation can produce defects in the silicon bulk or at the Si/SiO_2 interface (Chap. 16, "Semiconductor Counters"). As a result, some parameters of G-APDs may change during irradiation (Renker and Lorenz 2009). For low-level light detection, dark count rate is particularly harmful. Hadrons create defects in the bulk silicon, which act as generation centers, and the dark current, the dark count rate, and the after-pulsing probability will increase during an irradiation. Measurements done with G-APDs from different producers showed significantly increased dark currents and dark counts after irradiation with 10^{10} neutrons/cm^2 (the irradiation is normalized to an irradiation with 1 MeV neutrons, and known as neutron fluence). At neutron fluences exceeding 10^{11} neutrons/cm^2, single-photon counting becomes impossible.

We finally note that some hybrid devices try to combine the best features of different technologies, an example of which is a vacuum-based device where accelerated photoelectrons hit a fast-scintillating crystal; the resulting scintillation light is detected by a small phototube (as

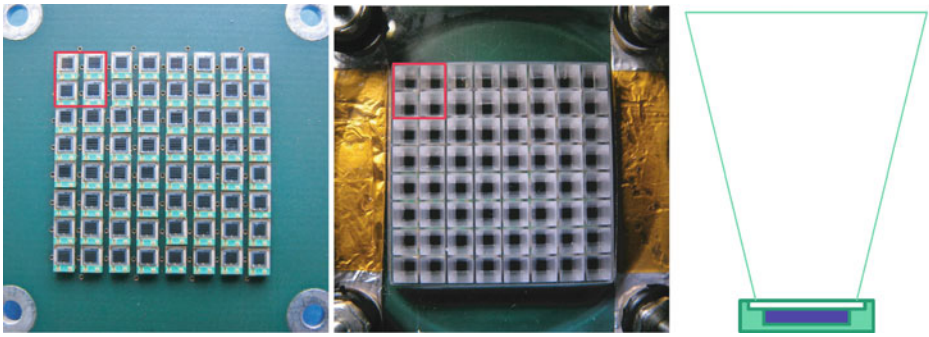


Fig. 11
The photon detector module consisting of 64 1 × 1 mm^2 G-APDs (Hamamatsu MPPC S10362-11-100P) without (*left*) and with (*middle*) the pyramidal light guides (*right*) on top of G-APD sensors (Korpar et al. 2010)

in the QUASAR-370 or XP2600 tubes) or with a G-APD. The advantage of such a sensor is that it can cover a large sensitive area without a complicated dynode structure of a photomultiplier tube. Such devices are considered as photon detectors for large-volume underwater neutrino telescopes.

Acknowledgments

The author wishes to thank A. Stanovnik, S. Korpar, and G. Kramberger for useful discussions. A. Stanovnik was also – as usual – kind enough to carefully read and comment the chapter.

References

Abachi S et al [D0 collaboration] (1994) Nucl Instrum Meth A 338:185

Abe K et al (2004) (edited by Hashimoto S, Hazumi M, Haba J, Flanagan JW and Ohnishi Y), Letter of Intent for KEK Super B Factory, KEK report 2004-04, http://belle.kek.jp/superb/

Albrecht E et al (2005) Nucl Instrum Meth A 553:215; Fabbietti L et al [HADES Collaboration] (2003) Nucl Instrum Meth A 502:256; Iodice M et al (2005) Nucl Instrum Meth A 553:231; Di Mauro A et al (2005) IEEE Trans Nucl Sci 52:972

Alves Jr AA et al [LHCb Collaboration] (2008) The LHCb detector at the LHC. J Instrum 3:S08005

Apsimon R et al (1986) IEEE Trans Nucl Sci 33:122; Arnold R et al (1988) Nucl Instrum Meth A 270:255–289; Aston D et al (1989) Nucl Instrum Meth A 283:582

Arisaka K (2000) Nucl Instrum Meth A 442:80

Arnold R et al (1992) Nucl Instrum Meth A 314:465

Artuso M et al (2005) Nucl Instrum Meth A 554:147

Braem A et al (2004) Nucl Instrum Meth A 518:574

Breskin A et al (2009) Nucl Instrum Meth A 598:107

Bross A et al (2002) Nucl Instrum Meth A 477:172

Buzhan P et al (2001) ICFA Instrum Bull 23:28; Buzhan P et al (2003) Nucl Instrum Meth A 504:48; Sadygov Z et al (2003) Nucl Instrum Meth A 504:301; Golovin V, Saveliev V (2004) Nucl Instrum Meth A 518:560

Chechik R, Breskin A (2008) Nucl Instrum Meth A 595:117

Danilov M [CALICE Collaboration] (2007) Nucl Instrum Meth A 582:451

Deiters K et al (2000) Nucl Instrum Meth A 442:193

Eisenhardt S [LHCb RICH Collaboration] (2006) Hybrid photon detectors for the LHCb RICH. Nucl Instrum Meth A 565:234
Flyckt SO, Marmonier C (2002) Photomultiplier tubes: principles and applications. Philips Photonics, Brive
Foehl K et al (2008) Nucl Instrum Meth A 595:88
Haitz R et al (1965) J Appl Phys 36:3123; McIntyre R (1966) IEEE Trans Electron Devices 13:164; Dautet H et al (1993) Applied Optics 32(21):3894; Perkin-Elmer Optoelectronics, Avalanche photodiodes: users guide
Hamamatsu Photonics, R5900-M64 data sheet, http://sales.hamamatsu.com/en/products/electron-tube-division/detectors/photomultiplier-tubes/part-r5900-00-m64.php. Accessed 27 February 2010
Hamamatsu (2006) Photomultiplier tubes, basics and applications, http://sales.hamamatsu.com/assets/applications/ETD/pmt_handbook_complete.pdf. Accessed 26 February 2010
Iijima T et al (1997) Nucl Instrum Meth A 387:64
Inami K (2008) Nucl Instrum Meth A 595:96
Janesick J (2001) Scientific charge-coupled devices. SPIE Press, Bellingham
Korpar S et al (2008) Nucl Instrum Meth A 594:13
Korpar S et al (2010) Nucl Instrum Meth A 613:195
Križan P et al (1997) Nucl Instrum Meth A 394:27
Lyashenko AV et al (2009) JINST 4:P07005
Mangeot P et al (1983) Nucl Instrum Meth A 216:79
Nakamura K et al (2010) Nucl Instrum Meth A 623:276
Nishida S et al (2009) Nucl Instrum Meth A 610:65
Particle Data Group (2009) Particle detectors for accelerators, http://pdg.lbl.gov/2009/reviews/rpp2009-rev-particle-detectors-accel.pdf. Accessed 26 February 2010
Petrov M, Stapelbroek M, Kleinhans W (1987) Appl Phys Lett 51:406; Atac M, Petrov M (1989) IEEE Trans Nucl Sci 36:163; Atac M et al (1994) Nucl Instrum Meth A 314:56
Renker D, Lorenz E (2009) J Instrum 4:P04004
Seguinot J et al (1994) Nucl Instrum Meth A 350:430
Yokoyama M et al (2009) Nucl Instrum Meth A 610:128

Further Reading

Flyckt SO, Marmonier C (2002) Photomultiplier tubes: principles and applications. Philips Photonics, Brive
Knoll GF (2000) Radiation detection and measurement. Wiley, New York
Krizan P (2009) Advances in particle-identification concepts. J Instrum 4:P11017
Leo WR (1994) Techniques for nuclear and particle physics experiments. Springer, Berlin
Photomultiplier technical papers from ET-Enterprises, http://www.etenterprises.com/technical-information/. Accessed 26 February 2010
Renker D, Lorenz E (2009) Advances in solid-state photon detectors. J Instrum 4:P04004
Rieke GH (2003) Detection of light, 2nd edn. Cambridge University Press, Cambridge

Suppliers of Technology

ET Enterprises. http://www.electrontubes.com/. Accessed 26 February 2010
Hamamatsu Photonics K.K. http://jp.hamamatsu.com. Accessed 26 February 2010
Photonique SA. http://www.photonique.ch/. Accessed 26 February 2010
Photonis Technologies S.A.S. http://www.photonis.com/. Accessed 26 February 2010
SensL. http://sensl.com/. Accessed 26 February 2010
Zecotek. http://www.zecotek.com/s. Accessed 26 February 2010

14 Neutrino Detectors

Franz von Feilitzsch · Jean-Côme Lanfranchi · Michael Wurm
Technische Universität München, Garching, Germany

1 Introduction . . . 314
1.1 Neutrino Sources . . . 315
1.2 Neutrino Properties . . . 316

2 Reactor Antineutrino Experiments . . . 317
2.1 The Reines–Cowan Experiment . . . 318
2.2 The Gösgen Experiment . . . 320
2.3 The KamLAND Experiment . . . 321

3 Solar-Neutrino Experiments . . . 322
3.1 Radiochemical Detectors . . . 323
3.1.1 Chlorine Experiment . . . 323
3.1.2 Gallium Experiments . . . 324
3.2 Water Čerenkov Detectors . . . 327
3.2.1 Kamiokande and Super-Kamiokande . . . 327
3.2.2 SNO . . . 328
3.3 The Borexino Experiment . . . 330

4 Neutrino Detectors at the GeV Range . . . 335
4.1 Super-Kamiokande . . . 336
4.2 MiniBooNE . . . 337
4.3 OPERA . . . 338
4.4 Further Neutrino Beam Detectors . . . 340

5 High-Energy Cosmic-Neutrino Detectors . . . 340
5.1 IceCube . . . 340
5.2 ANTARES . . . 341

6 Next-Generation Neutrino Observatories . . . 341
6.1 GLACIER . . . 342
6.2 Megaton Water Čerenkov Detectors . . . 342
6.3 LENA . . . 344

7 Conclusion . . . 345

References . . . 345

C. Grupen, I. Buvat (eds.), *Handbook of Particle Detection and Imaging*,
DOI 10.1007/978-3-642-13271-1_14, © Springer-Verlag Berlin Heidelberg 2012

Abstract: The neutrino was postulated by Wolfgang Pauli in the early 1930s, but could only be detected for the first time in the 1950s. Ever since scientists all around the world have worked on the detection and understanding of this particle which so scarcely interacts with matter. Depending on the origin and nature of the neutrino, various types of experiments have been developed and operated. In this entry, we will review neutrino detectors in terms of neutrino energy and associated detection technique as well as the scientific outcome of some selected examples. After a brief historical introduction, the detection of low-energy neutrinos originating from nuclear reactors or from the Earth is used to illustrate the principles and difficulties which are encountered in detecting neutrinos. In the context of solar neutrino spectroscopy, where the neutrino is used as a probe for astrophysics, three different types of neutrino detectors are presented – water Čerenkov, radiochemical, and liquid-scintillator detectors. Moving to higher neutrino energies, we discuss neutrinos produced by astrophysical sources and from accelerators. The entry concludes with an overview of a selection of future neutrino experiments and their scientific goals.

1 Introduction

Wolfgang Pauli invented neutrinos to save quantum statistics and energy conservation in nuclear β decay. At the same time he assumed that this particle would not be detectable in experiments as he invented it as an electrically neutral particle with practically no interaction with matter. The first experimental detection of neutrinos was achieved by Frederick Reines and Clyde Cowan only a quarter of a century later in 1956 (Cowan et al. 1956; Reines and Cowan 1956; Reines 1979). Since that time a continuously growing interest to explore the nature of this particle and to use it for the exploration of astrophysics led to numerous developments in the field of neutrino detection. In this chapter we try to give a review on neutrino detectors and the technology used. Due to the large number of neutrino detectors, which have been successfully operated we try to classify them in terms of neutrino energy and will describe outstanding detectors of those classes in more detail. Neutrinos are detected in a very large energy range starting from low-energy nuclear β decay up to the largest energies existing in cosmic rays, that is, $\sim 10^{20}$ eV. This leads to very different requirements in terms of detector technology. At low energies, due to the weak interaction cross section, the detectors have to allow for a high identification power against background arising from natural radioactivity and cosmic radiation. At energies which allow for deep inelastic scattering in nuclei, this is above 500–1,000 MeV, the interactions may lead to complex particle production which require good tracking and again a robust particle identification capability. At very high energies, which are no more achievable by laboratory-based neutrino sources but by cosmic rays, we expect a low neutrino flux. This requires very large detector masses and therefore technologies being applicable to very big detectors. After a short overview of possible neutrino sources and the relevant information on particle properties, we will describe neutrino detectors in the sequence of neutrino energies being detected.

1.1 Neutrino Sources

Figure 1 gives an overview of neutrino fluxes of different origins as a function of energy, whereas *Fig. 2* illustrates the energy dependence of neutrino cross sections.

Neutrinos are emitted by a variety of sources:

- Solar neutrinos from thermonuclear fusion processes inside the Sun
- Neutrinos from core-collapse Supernova explosions
- Neutrinos from neutron-rich isotopes inside nuclear reactors
- Geoneutrinos from the natural radioactivity of the Earth's interior
- Atmospheric neutrinos from the decay of cosmic muons
- Neutrinos from the pair annihilation of dark-matter particles
- Cosmic neutrinos from active galactic nuclei and gamma-ray bursts
- Accelerator-based neutrino beams

While artificial neutrino sources – nuclear power reactors and accelerator beams – are studied primarily to obtain information on neutrino particle properties, neutrinos from celestial bodies are used as probes in both astrophysics and geophysics. However, many of the most spectacular

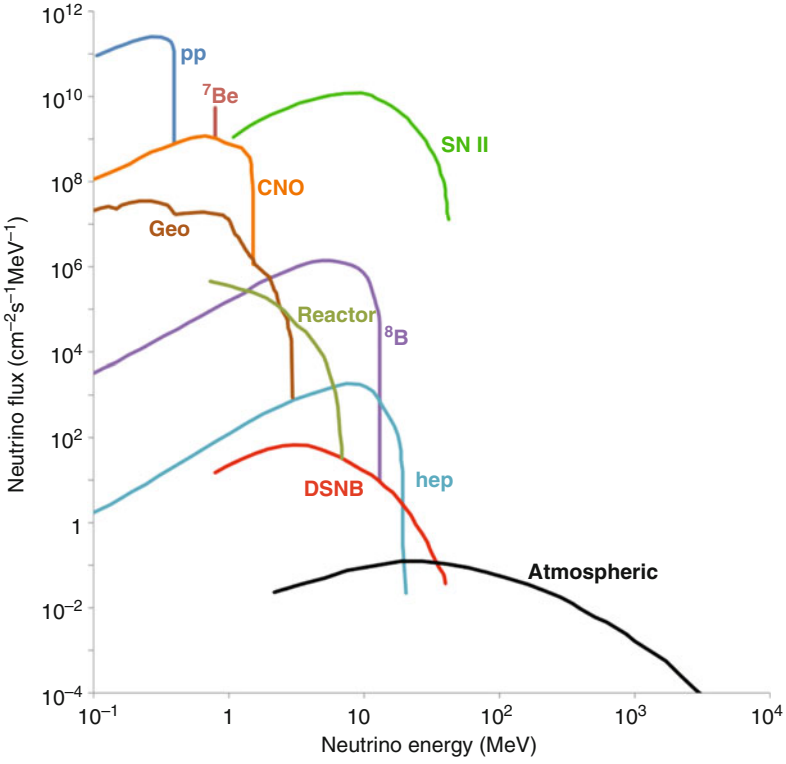


Fig. 1
Energy-dependent neutrino fluxes of different origins

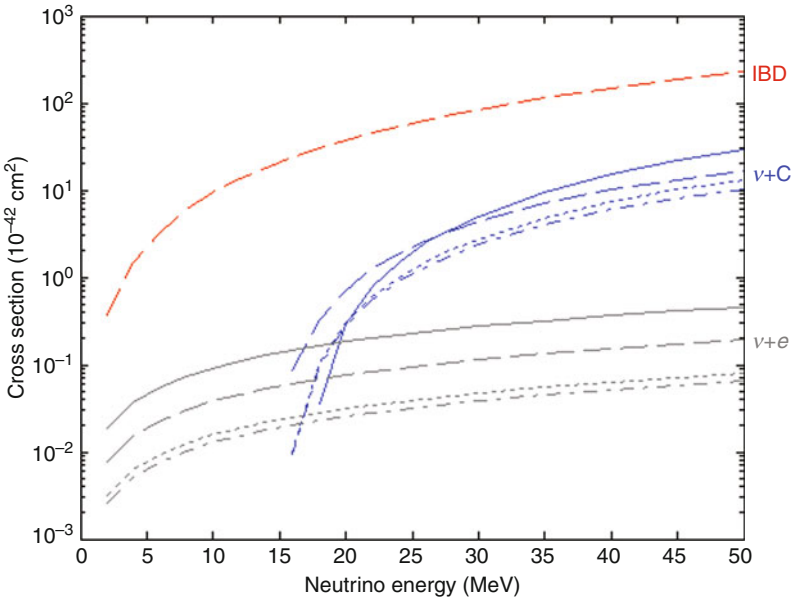


Fig. 2

Reaction cross sections for low-energetic neutrinos in a liquid-scintillator detector. The inverse beta decay (IBD) is the main detection channel for $\overline{\nu}_e$. The reactions on ^{12}C (ν + C) are depicted in *blue*, the elastic scattering on electrons (ν + *e*) in *grey*: *solid* – ν_e, *dashed* – $\overline{\nu}_e$, *dotted* – $\nu_{\mu,\tau}$, *dashes* and *dots* – $\overline{\nu}_{\mu,\tau}$

insights into the nature of the neutrino have been obtained by studying the neutrinos emitted by natural sources.

1.2 Neutrino Properties

The deciphering of the neutrino particle properties has always been closely linked to the detectors employed for their measurement. After establishing the existence of the neutrino by the Reines–Cowan experiment described in Sect. 2.1, solar radiochemical experiments returned a deficit in the measured neutrino flux that was incompatible with astrophysical expectations (Sect. 3.1). A similar deficit was observed with atmospheric neutrinos (Sect. 4). While in the beginning a variety of mechanisms was proposed to account for the measured deficit, neutrino oscillations emerged as the only explanation around the year 2000, supported mainly by the evidence of three experiments: the water Čerenkov detector Super-Kamiokande (Sect. 4.1), the heavy-water detector SNO (Sudbury Neutrino Observatory) (Sect. 3.2), and the liquid-scintillator detector KamLAND (Sect. 2.3).

In the standard model of particle physics, neutrinos are massless, electrically neutral particles that only interact weakly. There are three generations of particles: For instance, the electron e^- has two cousins, the muon μ^- and the tauon τ^-, that feature the exact same particle properties with exception of their larger masses. Accordingly, also the neutrino appears in three

kinds, commonly known as flavors, that each correspond to a charged lepton: the electron neutrino ν_e, the muon neutrino ν_μ, and the tau neutrino ν_τ. When reacting with matter via weak interactions, a ν_e will convert into an electron, but never into a μ or a τ. Also the antiparticles of the neutrino appear in three antiflavors, $\overline{\nu}_e$, $\overline{\nu}_\mu$, and $\overline{\nu}_\tau$, related to the antiparticle counterparts of the charged lepton, e^+, μ^+, and τ^+.

Neutrino oscillations become possible if the neutrino is not massless and if there is a mass difference for the three kinds of neutrinos, called mass eigenstates. However, quantum mechanics does not require these mass eigenstates to coincide with the flavor eigenstates that take part in the weak interaction. In fact, each flavor eigenstate is a mixture of mass eigenstates and vice versa. This mixture leads to the peculiar phenomenon known as neutrino oscillations: With a certain probability, depending on the ratio of the distance from the neutrino source to the detector and the neutrino energy, a neutrino emitted in electron flavor, ν_e, is detected as ν_μ or ν_τ. The maximum conversion factors are described by three neutrino mixing angles: θ_{12} describing solar-neutrino oscillations $\nu_e \rightarrow \nu_\mu$, θ_{23} observed in atmospheric neutrino oscillations, and θ_{13} unknown up to now. The distance at which the effect of oscillations becomes visible is known as the oscillation length, and depends on the mass-squared differences between the neutrino eigenstates, Δm^2_{21}, Δm^2_{32}, and Δm^2_{31}. Today, solar and atmospheric mixing angles are known to be large, leading to a dramatic change in the rates observed by the corresponding neutrino experiments. However, the third mixing angle, θ_{13}, is very small and is still searched for in reactor and neutrino beam experiments. Also the mass-squared differences are tiny, resulting in long oscillation lengths ranging from about 1 to several 1,000 km. However, their exact values are well determined by now, allowing to set up experiments at an optimal distance from the neutrino source at which the oscillation effects are most pronounced.

2 Reactor Antineutrino Experiments

The natural fluxes of low-energy neutrinos at energies around 1 MeV are far greater than the ones observed for atmospheric or cosmic neutrinos. On Earth, the highest natural flux is provided by the neutrinos from solar fusion. However, reactor antineutrinos provide locally even higher neutrino fluxes if the detector is not positioned too far from the reactor core. In the following, we will give an overview over the experimental techniques used for the detection of these neutrinos, starting from the discovery of the neutrino in the Savannah River reactor experiment.

The neutrinos emitted in the β decays of radioactive isotopes feature typical energies in the range of 0.1–10 MeV. A strong natural source of such neutrinos are the radioisotopes of the uranium and thorium decay chains that are embedded in the Earth's crust and mantle (see below). However, an even stronger source are the neutron-rich isotopes produced in nuclear fission reactors. A commercial nuclear power plant produces electron antineutrinos ($\overline{\nu}_e$) at a rate of ~10^{20} $\overline{\nu}_e$ per second. The $\overline{\nu}_e$'s emerge from the decay of neutrons that are part of the atomic nucleus of a neutron-rich fission product:

$$n \rightarrow p + e^- + \overline{\nu}_e \,, \tag{1}$$

where n is the neutron, p the proton, e^- the electron.

This source of neutrinos became available in the early 1950s with the rise of nuclear energy production. At that time, the only hint for neutrinos came from measurements of the electron energy emitted in β decays. The observation of an electron energy distribution had led Pauli to

the postulation of the neutrino. Now, F. Reines and C. L. Cowan had the goal to give final prove to the existence of this particle by an observation of the "free" neutrino at some distance from the neutrino source (Cowan et al. 1956).

2.1 The Reines–Cowan Experiment

Reines and Cowan set up their neutrino experiment first at the Hanford (Reines and Cowan 1953a, b) and later on at the Savannah River nuclear reactor (Cowan et al. 1956; Reines and Cowan 1956; Reines 1979). The experiment was designed to detect the low-energy $\overline{\nu}_e$ by an inversion of the process described by Eq. 1. This "inverse β decay" uses the capture of a $\overline{\nu}_e$ on a free proton,

$$p + \overline{\nu}_e \rightarrow n + e^+, \tag{2}$$

producing a neutron and a positron (e^+) in the end state. The coincident detection of the emerging positron and the neutron allows for a significant identification power and thus discrimination against natural radioactivity in the detector. A further advantage at that time was that the cross section of this reaction could be calculated accurately from standard weak interaction theory.

Originally, Reines and Cowan thought of using a nuclear explosion to yield a pulse of neutrinos intense enough to override the natural radioactive background. For some time this concept was under serious discussion, until it was realized that reaction (Eq. 2) bore a great rejection power so that the first tentative experiment was performed at the fission reactor at Hanford, Washington. After a few months of operation and after modifying the shielding of their detector several times, Cowan and Reines finally concluded that the reactor-independent background they were facing was overwhelming (Reines 1979). Thereafter, the experiment was dismantled. After a period of reflection and analysis of the data, they decided to place the next experiment underground, based on the conclusion that the background had to be of cosmic origin.

They performed their next attempt to detect the neutrino at the Savannah River Power Plant, South Carolina. There, again using the inverse-β-decay reaction, a detector was set up which could discriminate more selectively against reactor-independent as well as reactor-associated backgrounds. The new location was particularly well suited since the reactor had a comparably high power (700 MW) and a small physical size. In addition, a well-shielded place was available only 11 m away from the reactor core and protected against cosmic radiation by a "rock" overburden of ~12 m. The neutrino flux originating from this reactor at that distance was $\sim 1.2 \times 10^{13} \mathrm{cm}^{-2} \mathrm{s}^{-1}$.

The detection principle is illustrated in *Fig. 3*. A $\overline{\nu}_e$ originating from fission products in the reactor is incident on a water target that provides the hydrogen for the reaction of Eq. 2 and contains cadmium chloride as an additive. A positron and a neutron are produced in the water volume. The positron slows down and annihilates with an electron of the detector, emitting two 0.511 MeV gamma rays in opposite directions. The γ rays cross the water and are detected in coincidence by two large scintillation detectors on opposite sides of the target. The neutron is moderated in the water and finally captured by the cadmium, thus producing multiple gamma rays, which are also observed in coincidence by the two scintillation detectors. The cadmium is used for the neutron detection since it has a high neutron absorption cross section.

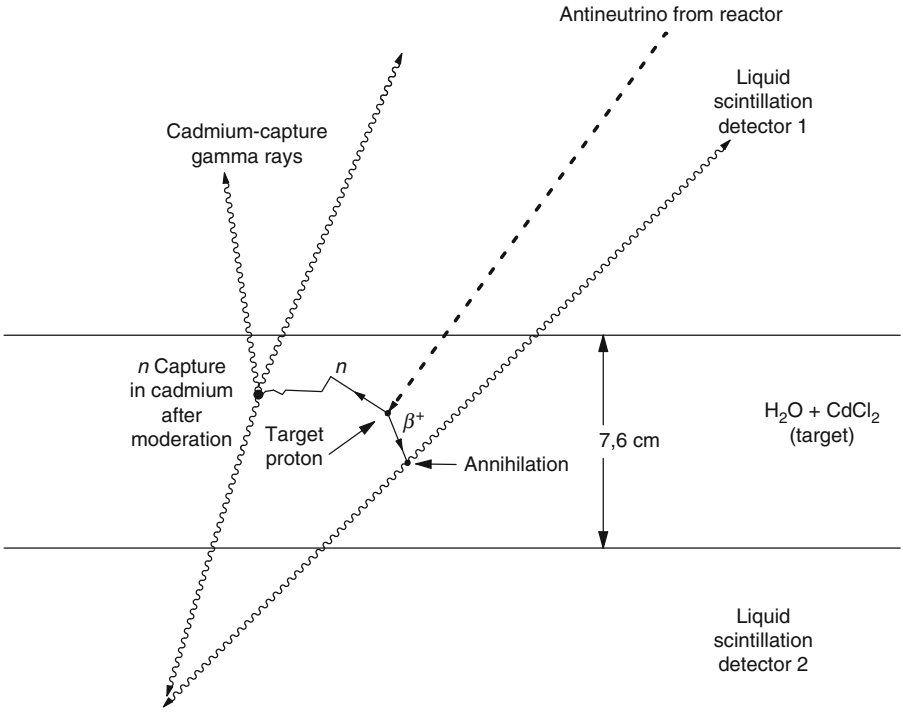


Fig. 3

Schematic of the neutrino detection principle used by C. Cowan and F. Reines. Electron antineutrinos interact with the water target containing cadmium chloride ($CdCl_2$). According to Eq. 2 a positron and a neutron are produced. The positron slows down and is annihilated with an electron thus producing two 0.511 MeV gammas, which are detected in coincidence by two scintillation detectors on opposite sides of the target. The neutron is moderated in the water and finally captured by the cadmium thus producing multiple gamma rays, which are observed in coincidence by the two scintillation detectors. The antineutrino signature consists of a delayed coincidence between the prompt gamma pulse produced by the e^+–e^- annihilation and gamma pulses produced microseconds later by the de-excitation of the cadmium

The signature of an $\overline{\nu}_e$ capture is therefore a delayed coincidence between the prompt gamma pulse produced by the e^+–e^- annihilation and gamma pulses produced microseconds later by the neutron capture on cadmium.

Even though the conditions were much more favorable at Savannah River, the whole experiment necessitated a running time of about 100 days over a period of roughly 1 year. In the end, Reines and Cowan succeeded in a definitive observation of the $\overline{\nu}_e$ in 1956 (Cowan et al. 1956; Reines and Cowan 1956). C. Cowan and F. Reines informed Wolfgang Pauli via telegram about their discovery and that the measured cross section agreed well with the theoretically expected one of $\sim 6 \times 10^{-44}\text{cm}^2$ (Fermi 1934). On receiving the news, he replied by telegram: "Thanks for message. Everything comes to him who knows how to wait. Pauli." In 1995, the Nobel Prize for physics was attributed to F. Reines (C. Cowan had died already in 1974).

Numerous questions that even today affect neutrino-detector generations later, are of the same kind as those faced by Cowan and Reines (Reines 1979): the transparency of the scintillator to transmit its own light for a few meters, the reflectivity of the inside of the target container, the chemical stability of the scintillator after addition of the neutron-capturing element, the behavior of the photomultiplier tubes (PMTs), and many more. The experimentalists also realized soon that their new detector designed to detect neutrinos had unusual properties concerning other particles such as neutrons and gammas, featuring detection efficiencies near 100%. They recognized that detectors of this type could be utilized to study diverse quantities such as neutron multiplicities in fission, muon capture, muon decay lifetimes, and natural radioactivity of the human body (Reines 1979).

2.2 The Gösgen Experiment

Even after their detection, neutrinos were still peculiar particles, inviting for a lot of speculation about properties not accounted for by the standard model of particle physics. The search for nonstandard effects picked up speed in the 1970s, when the solar-neutrino experiment conducted by Ray Davis in the Homestake mine showed a considerable deficit compared to the expected rate (see below). Again, nuclear reactors were exploited as an intense neutrino source, this time searching for neutrino oscillations that had been proposed by Bruno Pontecorvo. Compared to solar neutrinos, reactors provided the advantage that the $\overline{\nu}_e$ offered a clear detection signature and that the baseline to the neutrino source could be freely chosen.

In these experiments, the signature for neutrino oscillations was the disappearance of the $\overline{\nu}_e$ emitted by the nuclear reactor. The inverse β decay (Eq. 2) is only sensitive to antineutrinos of electron flavor; $\overline{\nu}_\mu$ and $\overline{\nu}_\tau$ are too low in energy to interact in the same way. Experimental technologies similar to the one of the Savannah River detector were employed, partly replacing the cadmium by different neutron-absorbing isotopes. For instance, ^{3}He was used in neutrino oscillation experiments performed at the research reactor of the Institut Laue Langevin (Grenoble, France) and at the nuclear power plant in Gösgen (Switzerland) (Zacek et al. 1986).

Figure 4 shows the detector setup used at the Gösgen reactor in Switzerland in the early 1980s. Here, the neutron emerging from the inverse beta reaction with an energy of several keV thermalizes within 10 μs in the liquid scintillator and diffuses into an adjacent wire chamber filled with ^{3}He gas. Capture of thermal neutrons in the ^{3}He gas yields a proton and a triton with a reaction Q value of 765 keV. The ^{3}He chambers are operated in the proportional region and charge division technique is applied to make counters position-sensitive. Again, the signature of a neutrino event is the coincident signal from the positron and the delayed gammas from the neutron capture. To further suppress accidental background, an additional position correlation of the positron and the neutron event is required. As depicted in *Fig. 4*, a total of four ^{3}He gas counters were mounted in between five liquid-scintillator target cell planes. In order to reduce cosmic-ray background, the detector was completely shielded by various passive shielding materials. An active muon veto consisting of additional liquid-scintillator cells surrounding the detector was operated in anti-coincidence. No evidence of neutrino oscillations could be found as the distance between reactor core and detector was not sufficient.

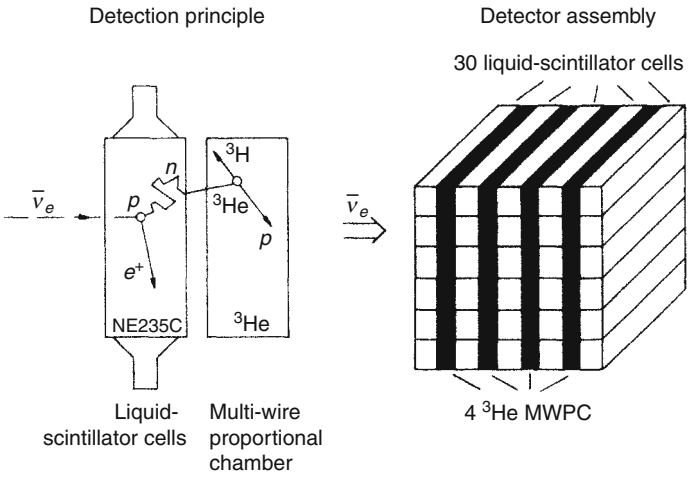


Fig. 4

Schematic of the Gösgen reactor experiment (Zacek 1986; Zacek et al. 1986). Neutrino detection principle (*left*). Realized neutrino detector (*right*). The central neutrino-detector unit consists of 30 liquid-scintillator cells arranged in five different planes for positron detection and four ^{3}He-filled multi-wire proportional chambers (MWPC) for neutron detection

2.3 The KamLAND Experiment

When it became clear that no evidence of oscillation could be found close to the reactor cores, experiments had to move further and further away from their neutrino source. This drastically increased the requirements in target mass and background rejection efficiency, as the observed neutrino flux and therefore the neutrino rate dropped with the square of the distance. Finally, a liquid-scintillator detector of ~1,000 t target mass was deployed in the Kamioka mine in Japan at a depth of 2,700 m water equivalent underground. Observing the $\overline{\nu}_e$ signals from a multitude of Japanese nuclear power plants at typical distances of ~180 km, the KamLAND (Kamioka Liquid-Scintillator Anti-Neutrino Detector) was finally able to provide convincing evidence of $\overline{\nu}_e$ oscillations.

Also in KamLAND, $\overline{\nu}_e$ were detected using the delayed coincidence signature of the inverse β decay. When passing the target volume, a $\overline{\nu}_e$ can be captured by a free proton of the hydrocarbon-based scintillator. The positron quickly deposits its energy and then annihilates. The remaining neutron thermalizes and is captured some 200 μs later on another hydrogen nucleus, yielding a deuteron and an associated 2.2 MeV de-excitation gamma. Positron and gamma signal again provide a clear signature for the detection of $\overline{\nu}_e$.

The energy of the emerging positron is closely related to the one of the incident $\overline{\nu}_e$. In KamLAND, this relation could be exploited to perform a spectrally resolved measurement of the reactor neutrinos. The most striking evidence for the observation of neutrino oscillations was therefore not only the deficit in the detected neutrino rate, but also the deformation of the observed neutrino energy spectrum.

KamLAND consists of an outer detector and an inner detector. The outer detector is a water Čerenkov detector equipped with ~200 PMTs acting as an active anti-coincidence veto for cosmic muons, that potentially produce background events in the central detector. The inner detector is enclosed by a metal sphere which is equipped with ~2,000 PMTs. A balloon suspended in the sphere contains the target scintillator (~1,000 t), which acts as both, the detection and target volume. Between the scintillator in the balloon and the PMTs plain mineral oil is used to shield against radioactivity originating from the surrounding rocks and the PMTs themselves. A similar experimental design was employed in the Borexino experiment (see below, Sect. 3.3).

The success of KamLAND depended primarily on the compliance of strict requirements for the radioactive contamination of all the utilized detector materials – natural radioactivity levels are orders of magnitude too large. Therefore, the detector had to be designed in a way that allowed to maintain the required radiopurity levels during the construction and operation phases. Otherwise, accidental pairing of uncorrelated single events would overwhelm the neutrino signal. However, the requirements are much less stringent than they would be for a non-coincident-type signal. Also, the large monolithic volume of liquid provides self-shielding capabilities, restricting most of the gamma-ray backgrounds caused by the other detector materials to the outer layers of the target volume. A serious background arises from radioactive processes that intrinsically produce a delayed coincidence signal similar to that of the inverse beta decay, and much effort was devoted to characterize and quantify events of that kind.

There is a further background at lower energies, originating from $\overline{\nu}_e$ produced by the β decays of natural radioactive isotopes embedded in the Earth's crust and mantle. These geoneutrinos originate mainly from decays of elements in the uranium and thorium chains. While originally a background to the reactor-neutrino analysis in the region below the endpoint of the geoneutrino spectrum at 2.6 MeV, the statistics gathered in KamLAND are now sufficient to give a positive evidence of the geoneutrino signal (Araki et al. 2005a). More recently, also the Borexino experiment performed a measurement of these neutrinos, increasing the significance of the result. While the statistics of the measurements is up to now not sufficient to deduce quantitative results on the Earth's radiogenic heat production (the statistics in Borexino is ten events in 2.5 years of measurement), future detectors will be able to perform much more accurate measurments of the geoneutrino flux and spectrum, shedding light on the otherwise unaccessible chemical composition and heat production of the inner layers of the Earth.

3 Solar-Neutrino Experiments

Another pioneering effort to detect low-energy neutrinos – this time from the Sun – was undertaken by Ray Davis in the early 1960s. A different detection approach is referred to as *radiochemical detection*. This technique was used at that time to allow for a disentanglement of neutrinos from natural radioactivity at sub-MeV energies as required for the detection of solar neutrinos.

Energy inside the Sun is emitted by the exothermal thermonuclear fusion of hydrogen to helium:

$$4p \rightarrow {}^4\mathrm{He} + 2e^+ + 2\nu_e\,. \tag{3}$$

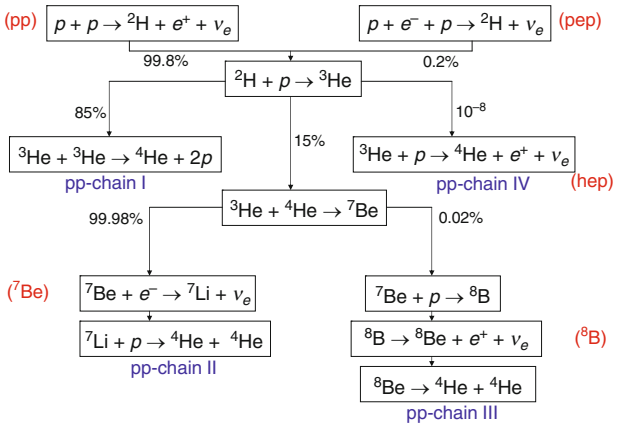


Fig. 5

Fusion branches in the Sun. The pp-fusion chain is responsible for ~98.5% of the energy production in the Sun as calculated by present standard solar models (SSMs) (Bahcall et al. 2005). The figure illustrates the various contributing reactions as well as their branching ratios; the names used to identify the neutrinos from the different reactions are marked in *red*

This fusion reaction takes place deep inside the Sun at temperatures of a few million degrees Kelvin. The energy released per ^{4}He fusion process is 26.73 MeV. From this total energy only about 2% are transferred onto the two ν_e. The remaining fraction of energy is set free in the form of thermal energy. Therefore, using the solar constant $S = 8.5 \times 10^{11}$ MeV cm^{-2} s^{-1} one can estimate the ν_e flux, Φ_ν, on Earth, which is then

$$\Phi_\nu = \frac{S}{13\ \mathrm{MeV}} = 6.5 \times 10^{10}\ \mathrm{cm}^{-2}\ \mathrm{s}^{-1}. \tag{4}$$

The energy production mechanism inside the Sun is divided into several sub-cycles, the pp, ^{7}Be, pep, ^{8}B, hep, and CNO cycle (Bethe–Weizsäcker cycle) as depicted in *Fig. 5*. For a more detailed and accurate description of the solar fusion processes so-called standard solar models (SSMs) have been developed mainly by Bahcall (Bahcall et al. 2005) and Turck-Chièze (Turck-Chièze 2001). These models provide predictions about the energy and abundance of neutrinos originating from different branches of the fusion cycles (Bahcall et al. 2005). *Figure 6* depicts the solar neutrino spectrum together with thresholds for different detectors.

3.1 Radiochemical Detectors

3.1.1 Chlorine Experiment

It was the goal of the Homestake Chlorine experiment, proposed by Ray Davis, to measure solar neutrinos for the first time (Davis 1964; Bahcall and Davis 1976). The detection reaction used is:

$$\nu_e + {}^{37}\mathrm{Cl} \rightarrow {}^{37}\mathrm{Ar} + e^-. \tag{5}$$

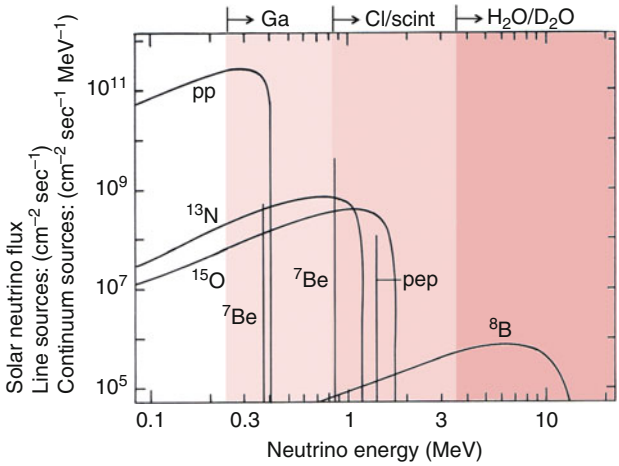


Fig. 6

The solar neutrino spectrum as predicted by the standard solar model (SSM) (Bahcall et al. 2005). The solar neutrino flux is plotted versus the neutrino energy in MeV. Depending on the fusion cycle in which the neutrino is generated, that is, two- or three-body process, the energy spectrum is discrete as in the case of ^{7}Be and pep and continuous in the case of pp, ^{8}B, and hep neutrinos. On the top *x* axis the energy detection threshold for different detector types is given

Solar neutrinos, ν_e's, weakly interact with a ^{37}Cl nucleus transforming one neutron into a proton thus producing ^{37}Ar. ^{37}Ar in turn is unstable and decays via electron capture (EC) back to ^{37}Cl with a half-life of ~35d,

$$^{37}\mathrm{Ar} + e^- \rightarrow {}^{37}\mathrm{Cl} + \nu_e \,. \tag{6}$$

The decay rate provides the rate of neutrino-produced ^{37}Ar atoms and thus the solar neutrino flux assuming the reaction cross section and the target mass is known. The back-decay of ^{37}Ar to ^{37}Cl yields an energy release of $E_{\mathrm{tot}} = 2.82$ keV via X-rays and Auger electrons. However, the results obtained by Davis et al. yielded only about one-third of the neutrino flux predicted by SSMs (Davis 1994, 1996; Cleveland et al. 1995). The reason for this was that radiochemical detectors in general are flavor specific, that is, the reaction is only valid for electron neutrinos, ν_e. At the time when Davis constructed his detector, neutrinos were still assumed to have no mass and could thus not oscillate from one flavor to another. Ray Davis was later awarded the Nobel Prize for physics in 2002 together with Masatoshi Koshiba "for pioneering contributions to astrophysics, in particular for the detection of cosmic neutrinos."

3.1.2 Gallium Experiments

Radiochemical gallium experiments are especially attractive since they provide information on the low-energy and predominant part of the solar neutrino spectrum, the pp neutrinos. At the beginning of the 1990s, two radiochemical gallium detectors each operated by an international collaboration, started solar-neutrino data taking: GALLEX in Gran Sasso, Italy, later

in 1998 to become GNO, the Gallium Neutrino Observatory and SAGE, the Soviet American Gallium Experiment in Bhaksan, Russia. In the following, we will discuss the GALLEX/GNO experiment.

GALLEX (Gallium Experiment) was a radiochemical detector located in Hall A of the Laboratori Nazionali del Gran Sasso (L.N.G.S.) with a rock overburden of ~3,600 m.w.e. The experiment measured solar neutrinos for ~12 years (1991–2003) thus monitoring one solar cycle (~11 y). Electron neutrinos with energies above 233 keV (see *Fig. 6*) are detected via the inverse-electron-capture reaction

$$\nu_e + {}^{71}\mathrm{Ga} \rightarrow {}^{71}\mathrm{Ge} + e^- \tag{7}$$

in a target consisting of 101 t of a $GaCl_3$ solution in water and HCl, containing 30.3 t of natural gallium. This amount corresponds to 10^{29} ^{71}Ga nuclei. Data taking was started by the international GALLEX collaboration in 1991, interrupted in 1997 for modifications in the data acquisition electronics and maintenance work concerning the chemical system and continued in May 1998 under the collaboration name GNO (Gallium Neutrino Observatory).

Figure 7 shows a schematic of the GALLEX/GNO experimental procedure.

The neutrino-produced ^{71}Ge is radioactive, and decays via electron capture (EC) to ^{71}Ga:

$$^{71}\mathrm{Ge} + e^- \rightarrow {}^{71}\mathrm{Ga} + \nu_e\,. \tag{8}$$

The half-life of this decay is of ~11.4 d, thus ^{71}Ge accumulates in the solution, reaching equilibrium when the number of ^{71}Ge atoms produced by neutrino interactions is just the same as the number of the decaying ones. When this equilibrium condition is reached, about a dozen ^{71}Ge atoms are present inside the chloride solution of GNO. The solar neutrino flux above threshold is then deduced from the number of ^{71}Ge atoms produced, using theoretically calculated cross sections as for example in (Bahcall et al. 2005). The ^{71}Ge atoms are identified by their decay after chemical extraction from the target. The experimental procedure for the measurement of the solar neutrino flux, referred to as a solar run (SR), is the following (see *Fig. 7*): The solution is exposed to solar neutrinos for about 4 weeks; at the end of this time ~10 ^{71}Ge nuclei are present in the solution, due to solar-neutrino interactions on ^{71}Ga. ^{71}Ge, present in the solution in the form of volatile $GeCl_4$, is chemically extracted (Anselmann et al. 1992) into water by pumping 2,000 m^3 of nitrogen through the target solution. The extracted ^{71}Ge is then converted into GeH_4 (germane gas), mixed with Xenon as counting gas and finally introduced into miniaturized proportional counters (MPCs). An efficiency of 95% is reached for the extraction of ^{71}Ga from the solution into the proportional counters where their decay is detected. Extraction and conversion efficiencies must be under constant control using nonradioactive germanium isotopes, such as ^{70}Ge, ^{72}Ge, ^{74}Ge, and ^{76}Ge alternately as carriers. ^{71}Ge electron-capture decay is observed for a period of 6 months, allowing the complete decay of ^{71}Ge and a good determination of the counter background. The intrinsic counter background is minimized by application of low-level-radioactivity technology in counter design and construction (Heusser 1995). The residual background is mostly rejected through application of amplitude and shape analysis to the recorded pulses. ^{71}Ge decays produce pulses originating from X-rays corresponding to an energy around 10.4 keV (K peak) or 1.3 keV (L peak) and Auger electrons. In the late measuring periods of GNO, the classical pulse-shape analysis is replaced by a neural-network analysis (Pandola et al. 2004). Counters are calibrated by an external Gd/Ce X-ray source, in order to carefully define amplitude and pulse-shape cuts with known efficiency for each measurement.

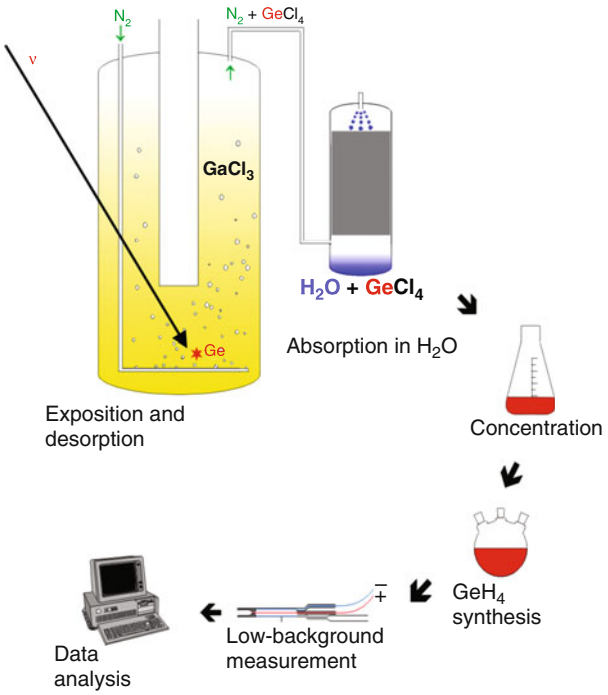


Fig. 7

Schematic illustration of a solar run (SR) in GNO: After an exposure of 3–4 weeks, the neutrino-produced ^{71}Ge is extracted by flushing the $GaCl_3$ solution with nitrogen and by absorption in water. After concentration of the obtained liquid and synthesis of GeH_4 (germane gas), the latter is mixed with xenon and filled into miniaturized proportional counters. The ^{71}Ge decay is then measured in a low-background environment, followed by the analysis of the data

The event amplitude and shape selection reduces the mean background rate to less than 0.1 counts per keV and day. The selected data are analyzed with a maximum-likelihood method to obtain the most probable number of ^{71}Ge nuclei at the beginning of counting, which (after correcting for counting, extraction, and filling efficiencies) gives the number of ^{71}Ge nuclei produced in the solution during the exposure and, therefore, the ^{71}Ge production rate. A correction is applied to account for contributions to the observed signal from processes other than solar-neutrino capture, producing ^{71}Ge as well (so-called side reactions), mainly due to interactions in the solution generated by high-energy muons from cosmic rays and by natural radioactivity. Another correction is made to account for background signals in the counter that can be misidentified as ^{71}Ge decays. The total correction amounts to typically a few percent of the signal (Altmann et al. 2005).

Not only did gallium experiments establish the presence of pp neutrinos (Anselmann et al. 1992) until then only predicted by standard solar models, but also a significant deficit, referred to as the solar neutrino puzzle, in the sub-MeV neutrino-induced rate (Anselmann et al. 1995; Hampel et al. 1996, 1999). At that time, this was the strongest indication for deviations of neutrino properties from the predictions of the standard model of particle physics, for example, for

neutrino transformations on the way between the solar core and the Earth, implying nonzero neutrino mass (Bahcall et al. 1998; Kirsten 1999; Berezinsky et al. 2000; Altmann et al. 2001). GNO, as well as similar measurements of the gallium experiment SAGE in Bhaksan, see Gavrin et al. (2003), have improved the quality of the data, added important restrictions on the presence of possible time variations in the integral solar neutrino flux, and substantially reduced the total error on the charged-current (CC) reaction rate for pp neutrinos as measured by the inverse beta decay on gallium (Altmann et al. 2005). Most important, together with the Cl experiment (Bahcall 2003) and the real-time measurements by Super-Kamiokande (Nakahata 2005) and SNO (Aharmim et al. 2005), it was established that neutrino flavor oscillations are by far the dominant mechanism responsible for the solar neutrino deficit (Wolfenstein 1978). Without radiochemical gallium detectors, the majority (~93%) of all solar neutrinos would still have remained unobserved.

GALLEX/GNO determined the bulk Ge-production rate with an accuracy of 5.5 SNU (1 SNU = 1 interaction per second per 10^{36} target nuclei). This is based on 123 solar runs (SRs), 65 from GALLEX and 58 from GNO. The results of GALLEX/GNO and SAGE and their precision will remain without competition from eventually upcoming low-threshold real-time experiments for many years to come. The gallium experiments have recorded a fundamental astrophysical quantity, the solar electron-neutrino flux at Earth, which after detailed measurements of the neutrino oscillation parameters allows to determine the solar neutrino luminosity. In the astrophysical context, the gallium results shed light also on the relative contributions of the PP-I, PP-II (see ❯ *Fig. 6*), and CNO cycles to the solar luminosity and on the agreement of the energy production derived from the photon and neutrino luminosities.

3.2 Water Čerenkov Detectors

3.2.1 Kamiokande and Super-Kamiokande

The Kamioka Nucleon Decay Experiment (Kamiokande) (Koshiba 1992) in the Kamioka mine (2,700 m.w.e.) 300 km west of Tokyo, Japan, was a water Čerenkov detector constructed in 1983 primarily aimed at the search for proton decay. Additionally, it was also capable of detecting solar ^{8}B neutrinos and supernova neutrinos via elastic neutrino–electron scattering (ES):

$$\nu + e^- \rightarrow \nu + e^- . \tag{9}$$

The Čerenkov light emitted by the recoil electron in water (fiducial volume, 680 t) was detected by a large number of PMTs surrounding the water target. In addition to radiochemical experiments, Čerenkov detectors offer the possibility of real-time detection at the cost of a higher energy threshold (see ❯ *Fig. 6*). The energy threshold (for recoil electrons) was ~9.3 MeV at first, and could later be improved to ~7.5 MeV. Nevertheless, the detector was only sensitive to the high-energy part of the solar neutrino spectrum (see ❯ *Fig. 6*), the ^{8}B and hep neutrinos. The neutrino flux measured was only ~50% of the value predicted by the SSM (Bahcall and Pinsonneault 1995) similar to the results of the ^{37}Cl experiment. However, in 1987 Kamiokande-II (Hirata et al. 1987) was able to detect neutrinos emitted by the supernova 1987A. A few of these events were also monitored by the IMB (Bionta et al. 1987) detector. In 2002, Masatoshi Koshiba, from the Kamiokande collaboration, was awarded the Nobel Prize together with R. Davis for pioneering contributions to astrophysics, in particular for the detection of cosmic neutrinos.

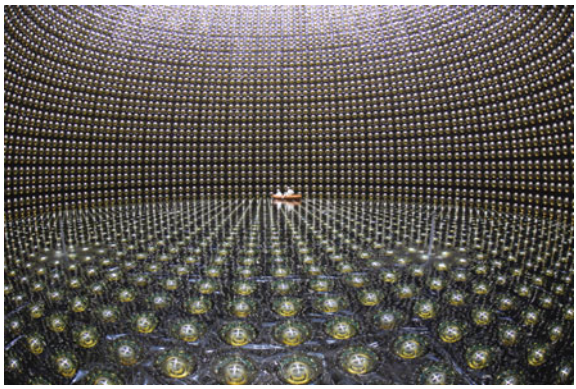


Fig. 8
Super-Kamiokande detector during water filling

In 1996, an enlarged version of the Kamioka water Čerenkov detector was built with a fiducial volume of 22,500 t and an energy threshold of ~5 MeV, Super-Kamiokande depicted in *Fig. 8*. Apart from the detection of the solar neutrinos, this detector was also conceived to measure atmospheric neutrinos, as well as to search for proton decay. Since this type of detectors exhibits a directional sensitivity, it proved that the neutrinos measured did indeed originate from the Sun. Again this detector measured only ~40.6% of the predicted solar-neutrino event rate. Further details are given in (Nakahata 2005).

3.2.2 SNO

The SNO (Sudbury Neutrino Observatory) detector depicted in *Fig. 9* located in the Creighton mine (6,000 m.w.e) close to Sudbury, Canada, is also a Čerenkov detector with an energy threshold of ~5 MeV (see *Fig. 6*). The target in this case is provided by ~1,000 t of heavy water (D_2O) contained in an acrylic vessel, monitored by ~9,500 eight-inch photomultipliers and surrounded by 1,700 t of ultrapure water. Neutrinos crossing the D_2O volume can be detected via three different reactions:

Charged-current **(CC)** reaction illustrated in *Fig. 10*:

$$\nu_e + d \rightarrow p + p + e^- ; \tag{10}$$

Neutral-current **(NC)** reaction illustrated in *Fig. 11*:

$$\nu_x + d \rightarrow p + n + \nu_x ; \tag{11}$$

Electron scattering **(ES)** illustrated in *Fig. 12*:

$$\nu_x + e^- \rightarrow e^- + \nu_x . \tag{12}$$

Flavor transitions from ν_e to ν_μ or ν_τ can be determined by comparing the interaction rates measured by the charged-current (CC) reaction and the neutral-current (NC) reaction.

Since only electron neutrinos are produced in the Sun, an excess in the neutral-current rate can only be attributed to a ν_μ or ν_τ component in the solar neutrino flux originating from



Fig. 9
Schematic of the SNO (Sudbury Neutrino Observatory) detector located at a depth of ~6,000 m.w.e in the Creighton mine, Canada. 9.500 eight-inch photomultiplier tubes (PMTs) are mounted on an 18 m sphere surrounding the heavy-water vessel. The PMTs are housed within "hex cells" which are bolted together into panels. The hex cells incorporate dielectric non-imaging reflectors, which increase the light acceptance of the PMTs. The heavy water is contained inside a 5-cm thick transparent acrylic vessel, submerged in light water. The water serves as a medium for absorption of gamma rays and neutrons that result from activity within the rock. The vessel is suspended from ropes. The vessel is a 12-m-diameter sphere with an 8-m tall neck section (2-m diameter) to provide access for filtering pipes or, for example, calibration devices

flavor transitions on their way to Earth. The electron scattering (ES) process allows a further measurement of neutrino interaction and a cross-check of the other two reactions, CC and NC, however, with less statistic significance. The results reported by the SNO collaboration are described in detail in Aharmim et al. (2005).

Results of the measurements by the SNO experiment are shown in *Fig. 13*. The diagram exhibits the flux of non-electron-flavor active neutrinos ($\nu_{\mu\tau}$) versus the flux of electron neutrinos (ν_e). The error ellipses shown are the 68%, 95%, and 99% joint probability contours for $\Phi_{\mu\tau}$ and Φ_e. The measured value of $\Phi_{\mu\tau}$ shows that ~2/3 of the electron neutrinos have undergone flavor transition.

The results include three different methods applied for the detection of neutrons from neutrino interactions (neutron absorption in D, NaCl, and in ^{3}He); these were successively used to reduce systematic uncertainties. For further details see Aharmim et al. (2005). The SNO experiment has hereby proven that part of the electron neutrinos produced inside the Sun alter their

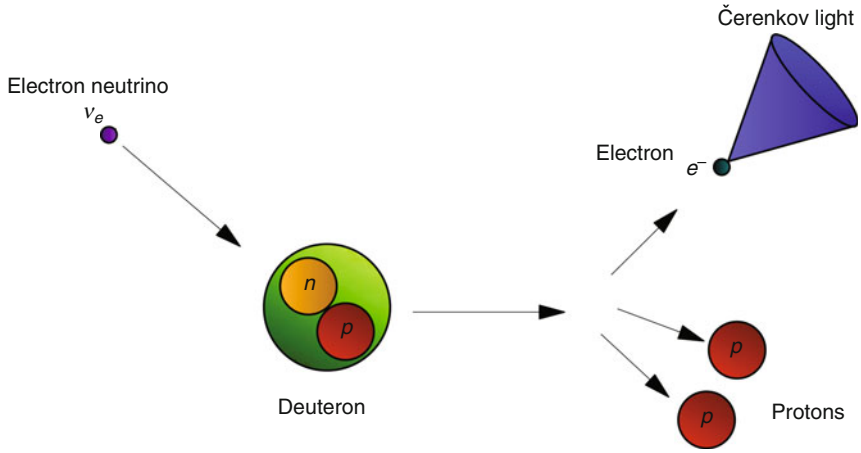


Fig. 10

Schematic of charged-current (CC) reaction. Via electroweak interaction (exchange of a *W* boson) the neutron in the deuterium is changed into a proton, and the neutrino into an electron. Due to the large energy of the incident neutrinos, the electron will be so energetic that it emits Čerenkov radiation, which can detected. The amount of light is proportional to the incident neutrino energy

flavor. Neutrino oscillations being the favored explanation for the neutrino deficit measured so far by every solar-neutrino experiment except for the NC reaction in SNO. Independently, Super-Kamiokande could prove strong evidence for neutrino oscillations measuring neutrinos produced in the atmosphere by cosmic rays (Fukuda et al. 1999). The idea of neutrino flavour oscillation was first proposed by Pontecorvo, Gribov, and Bilenky (Pontecorvo 1968; Gribov and Pontecorvo 1969; Bilenky and Pontecorvo 1978). The theoretical assumptions – involving nonstandard properties of the neutrinos – on which this theory is based are described in detail in Beuthe (2003), Naumov and Naumov (2010).

3.3 The Borexino Experiment

Water Čerenkov detectors have proven enormously successful in the detection of the solar neutrino spectrum above 5 MeV. At these comparatively high energies, the solar spectrum is dominated by ^{8}B neutrinos, while the majority of the solar neutrino flux lies below (see ❯ *Fig. 6*). Unfortunately, a further reduction of the detecion threshold in water Čerenkov detectors is extremely demanding. On the one hand, the amount of photons collected from neutrino interactions is about 4 per MeV of deposited energy in Super-Kamiokande (9 per MeV in SNO). At an energy of a few MeV, signals consisting of only 10–20 photoelectrons distributed over several thousand PMTs are hard to discern from instrumental backgrounds. Moreover, the natural radioactivity content of the detector materials begins to play a dominant role at energies below 3 MeV. At the time of writing, this proves to be an insurmountable obstacle for water

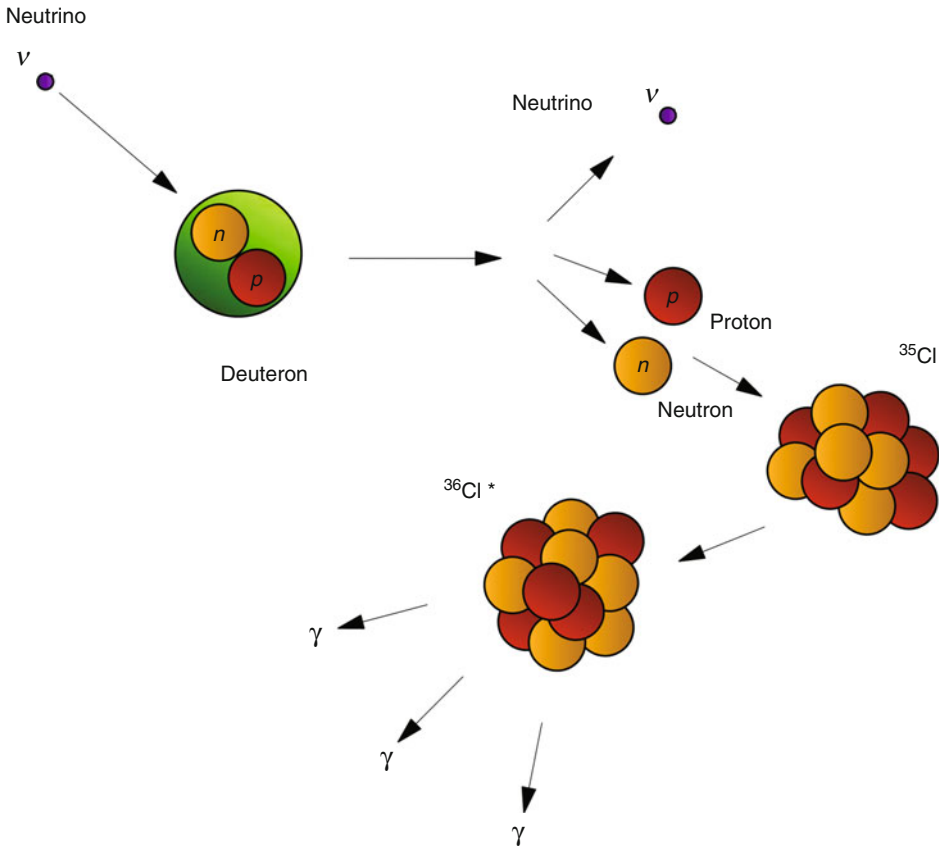


Fig. 11

Schematic of the neutral-current (NC) reaction. The weak interaction (exchange of a *Z* boson) causes the deuteron nucleus to break apart. The resulting free neutron is thermalized in the heavy water. The reaction can be observed due to gamma rays which are emitted when the neutron is finally captured by another nucleus. The gamma rays produce electrons which are detected via their Čerenkov light. The neutral-current reaction is equally sensitive to all three neutrino types; the detection efficiency, however, depends on the neutron capture efficiency and the resulting gamma cascade. Neutrons can hardly be captured directly on deuterium. Therefore, SNO – in a second phase of the experiment – enhanced the neutral-current detection by adding ^{35}Cl (in the form of NaCl) to the heavy water

Čerenkov detectors, limiting the most recent analysis of SNO data to energies above 3.5 MeV (Aharmim et al. 2010).

However, experiments based on liquid scintillator offer a viable alternative for the detection of low-energy neutrinos. Scintillation produces significantly more light than the Čerenkov effect, by about one to two orders of magnitude. Moreover, the organic liquids used as basis of the scintillators can be purified very efficiently of intrinsic contaminations with radioactive

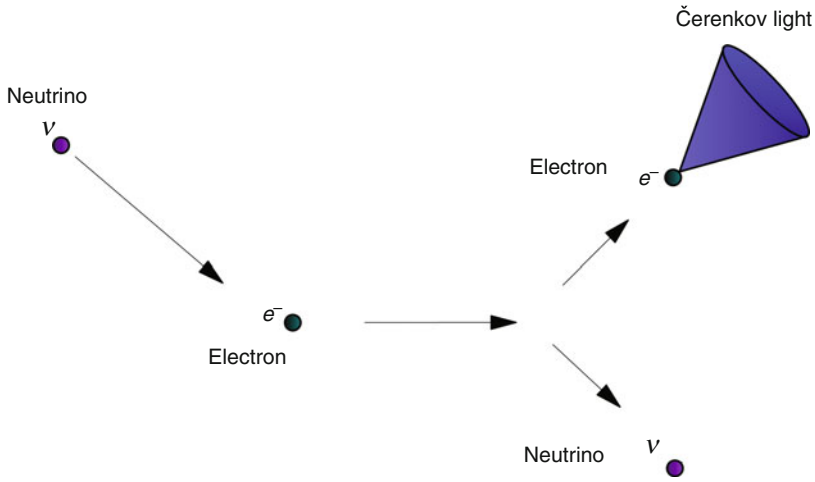


Fig. 12
Schematic of electron scattering (ES) reaction where the neutrino scatters off an electron. The reaction is sensitive to all neutrino flavors, the electron-neutrino contribution dominates by a factor of ~6. The final-state energy is shared between the electron and the neutrino, thus there is very little spectral information from this reaction. However, a good directional information is obtained

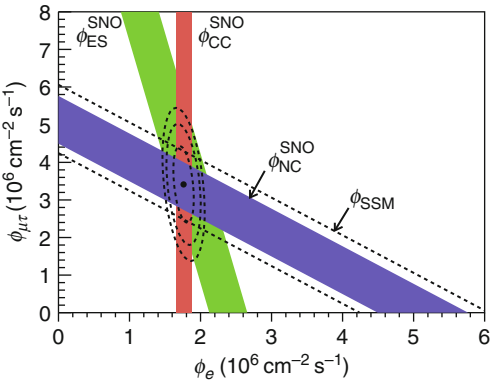


Fig. 13
Flux of μ and τ neutrinos versus flux of electron neutrinos. Charged-current (CC), neutral-current (NC), and electron scattering (ES) flux measurements are indicated by the filled bands (Aharmim et al. 2005). The total ^{8}B solar neutrino flux predicted by the SSM is shown as dashed lines, and that measured with the NC channel is shown as the solid band parallel to the model prediction. The intercepts of these bands with the axes represent the $\pm 1\sigma$ uncertainties. The nonzero value of $\Phi_{\mu\tau}$ provides strong evidence for neutrino flavor transformation. The point represents Φ_e from the CC flux and $\Phi_{\mu\tau}$ from the NC–CC difference with 68%, 95%, and 99% C.L. contours included

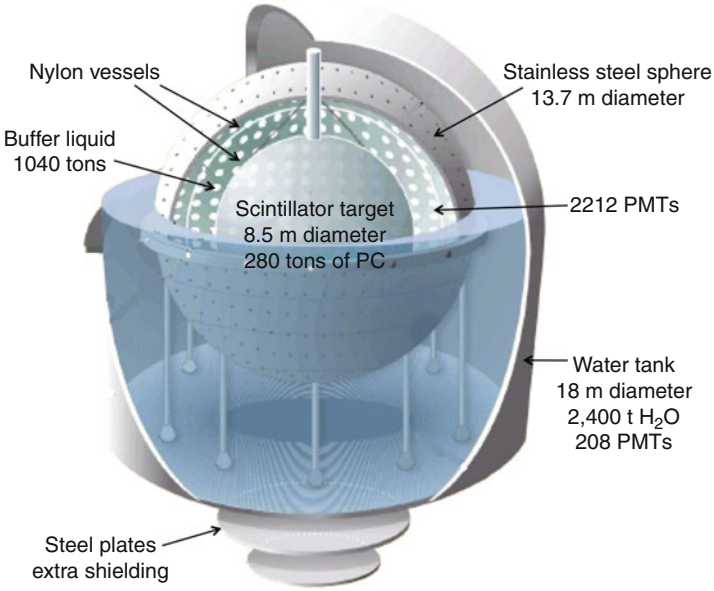


Fig. 14

Schematic of the Borexino detector. The experiment is contained in a stainless steel dome of ~18 m in diameter and consists of an outer and an inner detector. The outer detector serves as a shielding against external radioactivity and as an active water Čerenkov veto for cosmogenic muons. It is filled with 2,400 t of ultrapure water and is equipped with 208 PMTs. The inner detector of Borexino consists of a stainless steel sphere and two nested nylon vessels for radiopurity purposes. A total of 2212 PMTs are installed inside this sphere. It is filled with 1,040 t of shielding liquid outside and 280 t of liquid scintillator (with a fiducial volume of 100 t) inside the inner nylon vessel. The liquid scintillator is based on pseudocumene (1,2,4-trimethylbenzene $C_6H_3(CH_3)_3$) as it provides a high light yield and attenuation and scattering lengths suitable for this geometric configuration

isotopes, as distillation or column chromatography are purification techniques well-probed by (industrial) chemistry.

The liquid-scintillator experiment Borexino was designed for spectroscopy of low-energy solar neutrinos. In 2007, the detector started data-taking at the Gran Sasso national laboratories (LNGS) at a depth of 3,500 m.w.e. A sketch of the experimental setup is shown in *Fig. 14*. At the heart of the detector, ~280 t of liquid scintillator serve as the neutrino target. The scintillator is based on the organic solvent pseudocumene (PC), doped with a wavelength shifter at per mill concentration that allows the scintillation light to escape the reabsorption in PC by shifting the light to longer wavelengths. The target region is contained in an ultrathin spherical vessel of transparent nylon. It floats in an inactive buffer of pure PC held by a spherical steel tank of 13.7 m diameter. This steel sphere is again enclosed in a domed steel tank filled with ultrapure water, that serves both as shielding from external radioactivity and as an active veto for cosmic muons. It is equipped with 208 PMTs distributed on the sphere's outer surface and on the tank's floor to detect the Čerenkov light created by muons traversing the water.

Inside the steel sphere, the scintillation light of neutrino events is detected by 2212 PMTs mounted on the inner sphere surface. The individual multipliers are equipped with light-collecting conical mirrors around their photocathode that enhance the light detection area. Altogether, 30% of the sphere surface are photosensitive. The scintillator produces about 10,000 photons per MeV of deposited energy. Due to absorption of the light in the liquid, the limited wall coverage, and the photomultiplier detection efficiency of about 20%, merely 500 photoelectrons per MeV are registered at last. Nevertheless, the improvement in light yield compared to water Čerenkov detectors is substantial.

The primary objective of Borexino is the detection of solar ^{7}Be neutrinos at an energy of 866 keV (Back et al. 2008a). Like water Čerenkov detectors, the detection of solar neutrinos in Borexino relies on the elastic neutrino scattering off target electrons. However, as the signal of the recoil electron cannot be distinguished from those of the β and γ particles emitted in radioactive decays, purity of the target liquid is the main challenge of this experiment.

The expected signal induced by ^{7}Be neutrinos is of the order of 0.5 interactions per day and ton of detector material. To obtain a similar low rate of radioactive decays, the contamination with elements of the natural decay chains of ^{238}U and ^{232}Th must be below 10^{-17} gram per gram (g/g) scintillator in the target volume. For comparison, the natural uranium content of the Earth's crust is about 10^{-4} g/g. A reduction to the required level is only achievable by application of strict purity requirements during the construction of the detector, the transportation and the filling of the liquid, and the detector operation. In the end, radiopurity levels of $(1.6 \pm 0.1) \times 10^{-17}$ g/g for ^{238}U and $(6.8 \pm 1.5) \times 10^{-18}$ g/g for ^{232}Th were achieved (Back et al. 2008a). The remaining activity in the spectral region of the ^{7}Be neutrino signal, mainly caused by the radioactive noble gas ^{85}Kr dissolved in the scintillator, proved sufficiently low for detection (Back et al. 2008a).

Radiopurity requirements had not only to be set for the target scintillator, but also for the materials in contact with the liquid (mainly the nylon sphere, the surfaces of the liquid handling system, and the nitrogen used for flushing) and for the surrounding detector components. The latter, especially the radioactivity content of the photomultiplier glass bulbs, contributes to the background by emission of γ rays. A small fraction reaches the scintillation volume, creating electron-like signals. The necessity arises to limit the neutrino detection to the innermost 100 t of the scintillator volume, as it is shielded from the γ rays by the inactive buffer and the outer layers of the target liquid. This cut is possible as the spatial position of a neutrino or background event in the detector can be reconstructed by a comparison of the time-of-flight differences between the photons detected in individual PMTs. Dependent on event energy, accuracies on the level of centimeters can be reached. Finally, cosmic muons crossing the steel sphere near to its verge can produce low-light signals that can in principle be mistaken as neutrino signals. However, this background is effectively reduced by the passive shielding provided by the Gran Sasso mountain and the active suppression by the external muon veto. It is also possible to reconstruct the muon track by the light arrival-time patterns of the inner PMTs and the Čerenkov light cones detected in the water tank.

The science case of Borexino is not limited to ^{7}Be neutrinos – other contributions of the solar neutrino spectrum are within detection reach: The ^{8}B neutrino spectrum can be determined down to energies of ~3 MeV (Bellini et al. 2010b), and – though demanding – signals of pep, CNO, and even pp neutrinos are in principle accessible (see ❯ *Fig. 6*). The obstacles to be faced are cosmogenic backgrounds created by muon-induced spallation of carbon nuclei in the scintillator in the pep/CNO energy detection range, and the natural content of radioactive ^{14}C that largely covers the pp signal region. As described in ❯ Sect. 2.3, the detector is also sensitive

to geoneutrinos (Bellini et al. 2010a): Indeed, the location in central Italy is very favorable, as the main background to the signal – the flux of reactor neutrinos – is very low due to the distance to the next nuclear power plants in France and Germany.

4 Neutrino Detectors at the GeV Range

Solar and reactor neutrinos have proven very successful in the investigation of neutrino oscillations connected to the mixing angles θ_{12} and θ_{13} (commonly known as the "solar" and "reactor" angles). However, neutrino experiments at MeV energies are limited to disappearance searches of ν_e and $\overline{\nu}_e$, as the masses of μ and τ leptons potentially created by ν_μ and ν_τ are too large to be produced at these energies. However, neutrinos originating from cosmic μ decays in the Earth's atmosphere offer a natural source of ν_μ and $\overline{\nu}_\mu$ at GeV energies. The large energies are directly correlated to the oscillation length, which is of the order of hundreds to thousands of km. Dependent on whether atmospheric neutrinos are created in zenith, nadir, or at intermediate angles, oscillation baselines vary from 20 to 13,000 km, allowing for a broad range of oscillation parameters to be tested. In fact, the first evidence for neutrino oscillations of the type $\nu_\mu \to \nu_\tau$ was found in a disappearance search in the atmospheric neutrino signal of Super-Kamiokande (Sect. 3.2).

However, neutrinos at these energies also offer the possibility to search for the appearance of neutrinos of different flavor in an originally pure neutrino beam. This approach is broadly followed in accelerator-based neutrino experiments, in which the energy of the produced neutrinos as well as the oscillation baseline can be chosen artificially to obtain optimum sensitivity. Accelerators producing neutrino beams have been constructed at Fermilab in the USA, at JPARC laboratory in Japan, and at CERN in Europe. The typical layout of such a long-baseline experiment is shown in *Fig. 16*. Protons are accelerated in large quantities toward a solid target of light material, producing charged pions of large kinetic energy. These pions are selected and bent by a powerful, horn-shaped magnet to focus either a π^+ or π^- beam of defined energy into an evacuated decay tunnel. By the decay $\pi^+ \to \mu^+ \nu_\mu$, a collimated beam of ν_μ's is produced. The accompanying antimuons are stopped at the end of the decay pipe to impede a further decay of type $\mu^+ \to e^+ \nu_e \overline{\nu}_\mu$ that would contaminate the beam with unwanted neutrino flavors. Similarly, beams of $\overline{\nu}_\mu$'s can be produced by the decay of π^-.

The resulting neutrino beam travels for the length of the oscillation baseline through the Earth's matter, aimed at a "far" detector, in which the oscillated neutrino signal is recorded. Currently, baselines range from hundreds of meters to about 700 km. The far detector is sometimes accompanied by a smaller "near" detector close to the beam source in order to monitor flux and spectrum of the original beam. A comparison of "near" and "far" signal yields the information on neutrino oscillation effects. In the simplest case, the far detector searches for a disappearance of ν_μ's in the beam that can be explained by conversion to ν_e or ν_τ. However, there is also the possibility to search for the ν_e or even ν_τ by the production of electrons or tauons in the far detector.

In the following, we will describe three experiments in long-baseline neutrino oscillations, focussing on the far detectors and their techniques to identify neutrinos of different flavors: The Super-Kamiokande experiment, a large water Čerenkov detector that gave evidence for oscillations in atmospheric neutrinos and is currently also a far detector for a neutrino beam from JPARC in the T2K experiment (T2K collab. 2009). The MiniBooNE experiment is positioned

about 500 m down-beam of the Fermilab Booster, searching for nonstandard neutrino oscillations at very short baselines (MiniBoone collab. 2010). And third, the OPERA (Oscillation Project with Emulsion-tRacking Apparatus) experiment is a highly segmented sandwich detector, in which inactive sheets of heavy target material alternate with active sheets for particle detection, see OPERA collab. (2010).

4.1 Super-Kamiokande

The Super-Kamiokande detector is based on a large cylindrical water tank of 39 m diameter and 41 m in height, containing a total of about 50,000 t of ultrapure water. This volume is surrounded by 11,146 PMTs of 20 inch diameter each. The PMTs are placed at a distance of 70 cm, corresponding to a fraction of 40 % of the detector walls that is photosensitive. The detector is located 1,000 m underground (2,700 m of water equivalent) in the Kamioka mine in Gifu Prefecture, Japan. Data taking started in 1996, and has been going on to the present day, intermitted only by several breaks for detector upgrades.

Neutrinos are detected by the Čerenkov light of charged leptons (electrons or muons) that are produced in the interaction with the target material. At energies of several hundred MeVs, the dominant reaction occurs on single nucleons of the hydrogen and oxygen nuclei inside the target:

$$\begin{aligned} \nu_e + n &\to p + e^-, & \overline{\nu}_e + p &\to n + e^+, \\ \nu_\mu + n &\to p + \mu^-, & \overline{\nu}_\mu + p &\to n + \mu^+. \end{aligned}$$

While the recoil proton or neutron is below the Čerenkov threshold in water and remains therefore undetected, the emerging electron or muon will carry away most of the kinetic energy of the incident neutrino. Čerenkov light is emitted in a cone around the track of the recoiling electron or muon. This light cone is in many respects very similar to the Mach cone that is created by an airplane moving at supersonic speed through the Earth's atmosphere. Beginning from the interaction vertex in which it emerges close to vacuum light speed, the particle generates light until its velocity drops below the speed of light in water. The emitted light cone further propagates through the water bulk until it reaches the photoactive surface of PMTs. The ring-like image that is projected on this surface is evaluated by software analysis to obtain the particle's point of origin, its direction of flight, and its initial energy. In addition, the "fuzziness" of the ring can be used as a discrimination criterion for electrons and muons: Muons that are much heavier than the electrons of the water atoms in the target follow a straight track through the liquid. Electrons, however, will scatter repeatedly on target electrons, producing an electromagnetic shower of e^+e^- pairs and γ rays that is reflected in a washed-out Čerenkov ring. The resulting event signatures are shown in ❯ *Fig. 15*.

At higher neutrino energies, the analysis becomes more complicated due to the excitations of nuclear resonances in the neutrino interaction. By their decay, pions are created, that in turn decay themselves, either to $\mu^\pm$ in case of $\pi^\pm$ or to a pair of high-energetic gamma rays in case of π^0. All these particles may also lead to the emission of Čerenkov light and will cause additional rings overlaying the ring of the primary lepton. Moreover, neutrinos may also interact by weak neutral currents without creating a lepton in the final state. The methods applied to extract the signal events from these backgrounds is beyond the scope of this text. However,

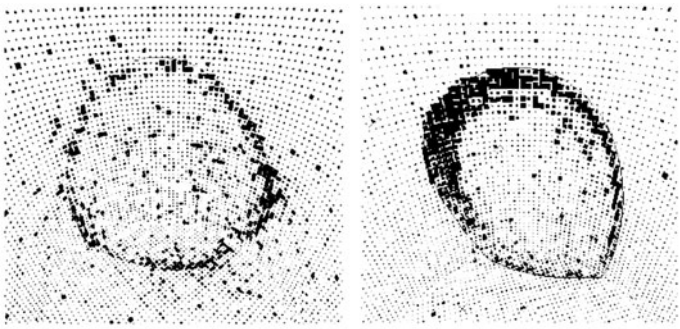


Fig. 15
The flavor recognition in the Super-Kamiokande relies on the "fuzziness" of the observed Čerenkov rings: In quasi-elastic scattering reactions, neutrinos create single electrons (*left*) and muons (*right*). As electrons scatter in the detector producing electromagnetic showers, the resulting ring is washed out, while a muon creates a clear signal

based on a distinct understanding of the reaction cross sections, event signatures, and discrimination efficiencies, it is possible to disentangle signal and background contributions, allowing for instance a statistic analysis of ν_τ appearance in atmospheric neutrinos (Hall 1999) or the search for ν_e appearance in the T2K neutrino beam experiment (T2K collab. 2009).

4.2 MiniBooNE

The Mini-Booster Neutrino Experiment (MiniBooNE) was built to probe an earlier result from the LSND (Liquid Scintillator Neutrino Detector) experiment (Hill 1995) at Los Alamos that indicated the oscillations of ν_μ's into sterile neutrinos. As the detector is very close to the beam source, beam intensity is very large. Therefore, MiniBooNE is also performing a precise measurement of neutrino interaction cross sections that will serve as input for the analysis of other experiments. First results obtained with this experiment can be found in (MiniBoone collab. 2010).

The Fermilab Booster produces a high-purity beam of ~0.7 GeV ν_μ by running protons of 8.8 GeV energy into a beryllium target of 71 cm length and 1 cm diameter (see *Fig. 16*). The protons arrive in spills of 4×10^{12} with a duration of 1.6 µs. As the produced neutrinos arrive in a sharply defined time window in the detector, background due to cosmic variation can be suppressed very effectively.

The detector is located 541 m from the beryllium beam target. The design is very similar to the one of KamLAND or Borexino: A large spherical volume with an inner radius of 610 cm is filled with 800 t of pure mineral oil (CH_2-based). The oil is monitored by 1280 PMTs and surrounded by a veto region viewed by 280 PMTs. Unlike in the low-energy experiments, the oil is only very weakly scintillating. The major part, 75%, of the light emitted from neutrino interactions originates from the Čerenkov effect.

To identify the signals of ν_e, ν_μ, and interaction vertices of different complexity, the pattern and the timing of the detected light is analyzed by neural networks. The analysis of $\nu_\mu \rightarrow \nu_e$

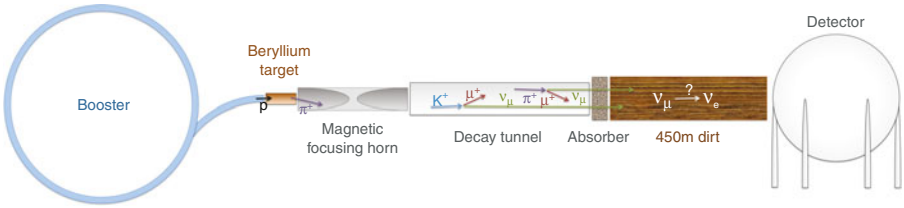


Fig. 16

Neutrino beamline to the MiniBooNE experiment: Protons are accelerated in the Fermilab Booster ring and are extracted to collide with a fixed beryllium target. The collisions of protons on light nuclei produce a beam mainly consisting of charged pions (and kaons). A magnetic focussing system ("magnetic horn") is used to culminate the π^+/K^+ beam and to reject other particles or mesons of negative charge. This beam enters into an evacuated decay tunnel in which the mesons decay to μ^+ and ν_μ. The remaining mesons and muons are absorbed in the beam dump ending the tunnel. The neutrinos (ν_μ) continue on through about 0.5 km of Earth matter, allowing for a flavor change due to neutrino oscillations. The MiniBooNE detector at the end of the beamline searches for the appearance of ν_e as well as the disappearance of ν_μ to probe oscillation models

oscillations relies on the isolation of the signals from electron appearance: The much larger sample of ν_μ events must be reliably rejected. Also neutral-current backgrounds due to interactions of neutrinos without a charged lepton in the end state play here an important role. Especially π^0 events may be mistaken as electrons: If one of the π^0-decay gamma quanta gains most of the kinetic energy while the other is nearly invisible, the single electromagnetic shower resembles an electron track from a ν_e event.

4.3 OPERA

In Europe, long-baseline projects currently focus on the ν_τ appearance in a ν_μ beam. For this purpose, the CNGS (CERN Neutrinos to Gran Sasso) beam at CERN was constructed. Its main physics objective is to prove explicitly the $\nu_\mu \rightarrow \nu_\tau$ nature of the atmospheric neutrino oscillation.

OPERA (Oscillation Project with Emulsion-tRacking Apparatus) (Agafonova et al. 2010) is an experiment aiming at detecting the appearance of ν_τ in an almost pure ν_μ beam through oscillation. ν_μ neutrinos are produced at CERN and pointed toward the Gran Sasso laboratory (LNGS). The distance between production and detection is ~730 km. The beam is optimized for the observation of ν_τ CC interactions, the average neutrino energy is ~17 GeV. The length of the baseline and the neutrino energy are therefore at a mismatch, so that the expected oscillation amplitude is small.

OPERA is a hybrid detector that associates nuclear emulsions to electronic detectors. The challenge of the OPERA experiment is the detection of the short-lived τ lepton produced in the charged-current interaction of ν_τ. This sets two requirements difficult to conciliate: a large target mass to collect enough statistics and a very high spatial resolution to observe the τ lepton, the decay length of which being in the millimeter range at the CNGS beam energy.

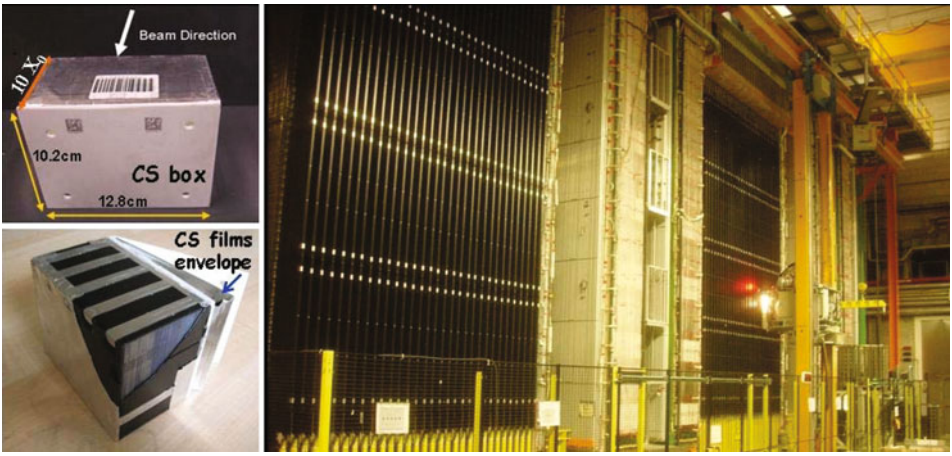


Fig. 17
The OPERA (Oscillation Project with Emulsion-tRacking Apparatus) experiment with two super modules (*right*). To reduce the emulsion scanning load, Changeable Sheets (CS) film interfaces have been used (*left*). They consist of tightly packed doublets of emulsion films glued to the downstream face of each brick. A target brick consists of 56 lead plates of 1 mm thickness interleaved with 57 emulsion films and weighs ~8.3 kg

In OPERA, neutrinos interact in a large-mass target made of lead plates interspaced with nuclear-emulsion films acting as high-accuracy tracking devices. This kind of detector is historically called Emulsion Cloud Chamber (ECC). It was successfully used to establish the first evidence for charm in cosmic ray interactions (Niu et al. 1971) and in the DONUT experiment for the first direct observation of the ν_τ (Kodama et al. 2001). OPERA (Acquafredda et al. 2009), which is depicted in *Fig. 17* is made of a veto plane followed by two identical super modules.

Each super module consists in turn of a target section of about 625 t made of 75,000 emulsion/lead ECC modules or bricks, of a scintillator target tracker detector to trigger the read out and localize neutrino interactions within the target. The super modules are followed by a muon spectrometer. A target brick consists of 56 lead plates of 1 mm thickness interleaved with 57 emulsion films and weighs 8.3 kg. Their thickness along the beam direction corresponds to about ten radiation lengths. In order to reduce the emulsion scanning load, Changeable Sheets (CS) film interfaces have been used. They consist of tightly packed doublets of emulsion films glued to the downstream face of each brick.

Charged particles from a neutrino interaction in a brick cross the CS films and produce signals in the target tracker that allow the corresponding brick to be identified and extracted by an automated system. Large ancillary facilities are used to bring bricks from the target up to the automatic scanning microscopes at LNGS and various laboratories in Europe and Japan. OPERA reported the first positive identification of a ν_τ appearance event in OPERA collab. (2010): The corresponding event is depicted in *Fig. 18*. The short track created by the τ from the neutrino interaction vertex to the particles decay is clearly visible. Extensive information on the OPERA detector and ancillary facilities are given in (Acquafredda et al. 2009). *Figure 17* shows a picture of the OPERA detector, as well as the submodules OPERA is made of.

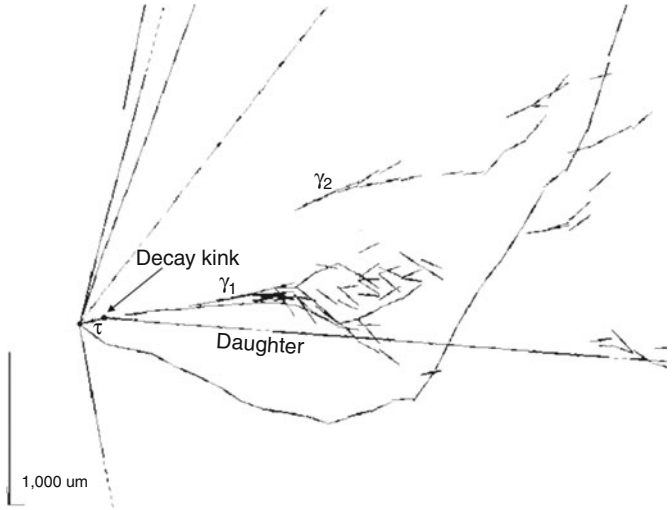


Fig. 18
Reconstruction of the first ν_τ appearance event in the OPERA experiment. The characteristic decay kink of the short-lived τ is clearly visible

4.4 Further Neutrino Beam Detectors

There is a variety of other neutrino beam detectors, based on radically different detection techniques, that we want to mention here only at a glance: The MINOS (Main Injector Neutrino Oscillation Search) experiment relies on a sandwich detector of alternating sheets of passive steel and active scintillator: The experiment explores the disappearance of ν_μ and $\overline{\nu}_\mu$ by the comparison of the signals in a near and a far detector, both similarly built but of different sizes (MINOS collab. 2010). The NOνA (NuMI Off-axis electron-Neutrino Appearance) detector currently under construction will be a fully active detector of $16 \times 16 \times 78\,\mathrm{m}^3$ in dimensions, highly segmented into individual detector cells of 3×6 cm thickness and 8 m in length, filled with liquid scintillator (Requejo et al. 2005). Both MINOS and NOνA are located in Minnesota, 735 km from the NuMI (Neutrinos at the Main Injector) beamline at the Fermilab in Illinois producing the neutrino beam. The ICARUS detector has started data taking in the Gran Sasso laboratory detecting neutrinos from the CNGS beam (Ankowski et al. 2010): It is the first large-scale (600 t) liquid-argon time-projection chamber, an experimental approach that is currently discussed also in the context of the large next-generation neutrino observatories presented in Sect. 6.

5 High-Energy Cosmic-Neutrino Detectors

5.1 IceCube

IceCube (Abbasi et al. 2010, 2011) the successor of the AMANDA (Antarctic Muon and Neutrino Detector Array) experiment (Ackermann et al. 2004), is a large-scale neutrino telescope at the South Pole which monitors neutrinos from the various astrophysical sources: events like

exploding stars, gamma-ray bursts, and cataclysmic phenomena involving black holes and neutron stars. IceCube has the potential to allow for an investigation of physical processes associated with the production of the most energetic particles in nature (GeV toTeV energies). The IceCube detector uses a cubic kilometer of ice to measure cosmic neutrinos as astronomical messengers from the universe. Due to their weak interaction, neutrinos are not affected by interstellar magnetic fields or by obscuring interstellar matter. Huge detectors are required to cope with the very low flux expected at these energies. To detect them a large volume of ice below the South Pole is utilized to monitor their interaction with the water molecules the ice is made of. Antarctic polar ice has exceptional intrinsic properties: It is transparent, and nearly free of intrinsic radioactivity; a mile below the surface, blue light travels a hundred meters or more through the otherwise dark ice. In the ultra-transparent ice just as in the water Čerenkov detectors described in Sect. 3.2, muons generate blue light that is detected by optical sensors (PMTs). The "trick" used in IceCube is that the muon preserves the direction of the incident neutrino thus allowing to trace back the neutrino to its cosmic source. However, muons produced in the Earth's atmosphere just above the detector also interact with the ice and create a huge background. To remove this undesired atmospheric muon contribution IceCube takes again advantage of the fact that neutrinos interact so weakly with matter. Since neutrinos can travel through the Earth unhindered, IceCube monitors the northern skies by "looking" through the Earth and thus using the planet as a cosmic ray shielding. "Real" cosmic neutrinos, which produce muons below the detector, are seen in IceCube as upcoming events.

5.2 ANTARES

Another approach for high-energy neutrino spectroscopy is pursued by the ANTARES experiment (Aguilar-Arevalo et al. 2010), which is presently the largest neutrino telescope operating in the northern hemisphere. In this case photomultiplier chains similar to those in AMANDA and IceCube are deployed deep underwater (South of France) to monitor the Čerenkov light emitted during interactions of high-energy neutrinos from various astrophysical sources.

6 Next-Generation Neutrino Observatories

Most of the current neutrino experiments rely on target masses of the order of kilotons to retrieve a sufficient rate of neutrino interactions and therefore statistical significance. The next generation of both low-energy astrophysical neutrino detectors and long-baseline neutrino beam experiments will have to increase their size considerably in order to improve their sensitivity substantially. Therefore, several detector options are discussed for sites in the USA, Japan, and Europe, varying the possible oscillation baselines to accelerator laboratories and the sensitivity to different sources of neutrinos. In the following, the three European projects GLACIER, MEMPHYS, and LENA are discussed, as they represent the largest variety of detector technologies (Autiero et al. 2007). However, very similar detectors are also discussed in the USA and in Japan.

There are also plans for a further development of the high-energy cosmic-neutrino detectors. While the DeepCore extension of IceCube that aims at lowering the neutrino detection threshold, the km3net collaboration promotes an underwater detector of $1\,km^3$ in size in the Mediterranean (www.km3net.org).

6.1 GLACIER

The GLACIER (Giant Liquid Argon Charge Imaging ExpeRiment) project foresees to apply the liquid-argon-chamber technique at a scale of 100 kt. Liquid Argon Time Projection Chambers (LArTPCs) allow for investigating the topology of interactions and decays of particles due to the bubble-chamber-like imaging performance. Like in gas-based time-projection chambers, electrons are created by ionizing particles inside the detector. A strong electric field of 100 kV/m is applied over the whole of the target volume to drift the electron cloud created along the particle track toward a two-dimensional detection array. The third dimension is provided by the relative drift-time differences of the ionization electrons, allowing to fully reconstruct the tracks with mm precision.

While a gas-filled detector of 100 kt would be of gigantic proportions, the density of liquid argon is sufficient to contain the volume in a cylindric cryogenic tank of 70 m diameter and 20 m in height. Cooling of the target will rely on argon boil-off. The electric field is applied in vertical direction, limiting the maximum drift length from the bottom of the detector to 20 m. The ionization signal will be complemented by scintillation light: Argon provides a high light yield even compared to organic liquid scintillators. The combination of both signals will provide high trigger efficiency, spatial and energy resolution.

A 16 kt LArTPC is also considered as a possible far detector for the American LBNE (Long Baseline Neutrino Experiment) project, investigating neutrino oscillation parameters by $\nu_\mu \rightarrow \nu_e$ appearance search. Also the physics program of GLACIER is clearly focused to a possible future long-baseline beam from CERN. The great strength of this detection technique in comparison to Čerenkov and scintillation detectors is the detailed resolution of GeV neutrino events, allowing for high beam energies and therefore very long beam baselines of more than 1,000 km.

The largest running LArTPC is ICARUS (Ankowski et al. 2010), which is operating successfully since summer 2010 at the LNGS underground laboratory, Italy. The detector principle is shown in ❯ *Fig. 19*. Since recently ICARUS is taking data utilizing the CERN–Gran Sasso neutrino beam (CNGS). For more detail we recommend (Ankowski et al. 2010).

6.2 Megaton Water Čerenkov Detectors

The extrapolation of the Super-Kamiokande experiment by an order of magnitude is currently discussed in Japan (Hyper-Kamiokande), the USA (LBNE), and in Europe (MEMPHYS). The aspired target masses range from 300 kt in the case of LBNE to more than 500 kt for Hyper-Kamiokande and MEMPHYS. However, the construction of the large underground caverns hosting the water volume sets limits to the maximum size of a single detector. Therefore, the experiment would consist of three egg-shaped subdetectors of smaller size, typically 50–60 m wide and 60–70 m in height (de Bellefon et al.). A recently proposed version considers two tanks with increased height.

In this context, also the doping of water with gadolinium is discussed. At low energies, antineutrino detection via the inverse beta decay in water Čerenkov detectors suffers from the invisibility of the neutron capture on hydrogen, as the 2.2 MeV gamma ray released is below the detection threshold. However, neutron capture on gadolinium produces a gamma cascade

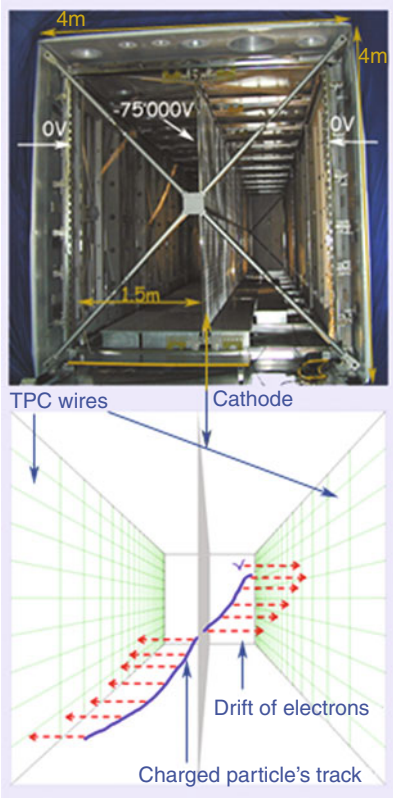


Fig. 19

Schematic and working principle of the TP600 modules of ICARUS. The charged particles pass inside the volume, where a uniform electric field is applied and produce ion–electron pairs. A fraction of them, depending on the field intensity and on the density of ion pairs, will not recombine and will immediately start to drift parallel to the field in opposite directions. Only the motion of the much faster electrons induces a current on a number of parallel wire planes located at the end of the sensitive volume

with a total energy of 8.8 MeV, a signal that produces sufficient light to be detected. Like in a liquid-scintillator detector, the coincidence of positron and neutron signal can be used to reject a variety of background events. Laboratory-scale tests have succeeded and gadolinium doping is also discussed for the currently running Super-Kamiokande detector.

The main advantage of a megaton water detector is its sheer target size: While lacking vertex reconstruction capabilities in comparison to LArTPCs at high energies and energy resolution and background discrimination capabilities compared to liquid-scintillator detectors at low energies, the enormous statistics can compensate for the poorer detector performance.

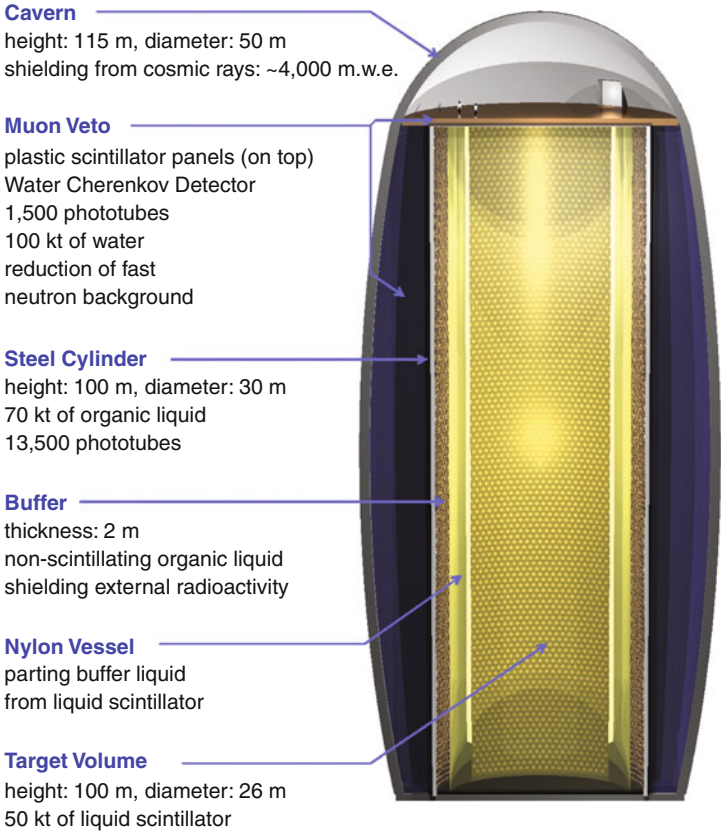


Fig. 20
Schematic of the future liquid-scintillator detector LENA for low-energy neutrino astronomy

6.3 LENA

The LENA (Low Energy Neutrino Astronomy) project is discussed as a next-generation liquid-scintillator detector (Wurm et al. 2010). The present design shown in *Fig. 20* foresees a cylindrical detector of about 100 m length and 30 m diameter. An inside part of 13 m radius containing 50 kt of liquid scintillator is separated from a non-scintillating buffer region by a nylon barrier. A steel or concrete tank separates this inner detector from an outer water tank that is used both for shielding and as an active muon veto. This onion-shell structure was already applied in KamLAND and Borexino and is a prerequisite for the favorable background conditions achieved in these experiments.

About 45,000 PMTs of 20 cm diameter each will be mounted to the internal walls. The photosensitive area will be further enhanced by conic mirrors (Winston cones) mounted to the individual PMTs, corresponding to an effective surface coverage of about 30%. In this way at least 200 photoelectrons (pe) per MeV will be detected. The limiting factor for the effective yields is not the light production, but the transparency of the organic liquid regarding the scintillation light. However, organic liquids featuring attenuation lengths of 10–20 m are available, in principle allowing for a yield of 450 pe/MeV, comparable to the situation in Borexino.

Due to the low detection threshold inherent to the liquid-scintillation technique, LENA will be the most versatile detector at low energies, providing high-statistics measurements of sub-MeV solar neutrinos, geoneutrinos, and the rare $\overline{\nu}_e$ signals of the cosmic Supernova neutrino background.

A liquid scintillator might also provide the possibility of particle tracking at GeV energies. In general, light emission in liquid scintillators is isotropic, impeding a directional reconstruction at MeV energies where particle tracks are short compared to the spacial resolution of the order of several centimeters. However, particle tracks become reconstructable when they exceed this length scale: The resulting light front is deformed from spherical to a conic shape very similar to the Čerenkov cone. This aspect is currently being studied by European and American groups, showing first promising results (Peltoniemi 2009).

7 Conclusion

For more than 50 years now successful neutrino detectors have been operated and have obtained results with profound scientific discoveries. It may be expected, in particular with a view to next-generation large-volume neutrino detectors that this field of research will develop into high-statistic-precision measurements. With a neutrino beam from an accelerator there may be a new window opening for further evaluation of the phenomenon of neutrino oscillation and a better understanding of the physics involved.

References

Abbasi R et al (2010) Search for muon neutrinos from gamma-ray bursts with the icecube neutrino telescope. Astrophys J 710:346

Abbasi R et al (2011) Measurement of the atmospheric neutrino energy spectrum from 100 GeV to 400 TeV with icecube. Phys Rev D 83: 012001

Ackermann M et al (2004) Search for neutrino-induced cascades with AMANDA. Astropart Phys 22:127

Acquafredda R et al (2009) The OPERA experiment in the CERN to gran sasso neutrino beam. JINST 4, P04018

Adamson P et al (2010) Neutrino and antineutrino inclusive charged-current cross section measurements with the MINOS near detector. Phys Rev D 81:072002

Agafonova N et al (2010) Observation of a first $\nu\tau$ candidate event in the OPERA experiment in the CNGS beam. Phys Lett B 691:138

Aguilar-Arevalo AA et al (2010) Event excess in the MiniBooNE search for $\bar{\nu}\mu \rightarrow \bar{\nu}e$ oscillations. Phys Rev Lett 105:181801

Aharmim B et al (2005) Electron energy spectra, fluxes, and day-night asymmetries of 8B solar neutrinos from measurements with NaCl dissolved in the heavy-water detector at the sudbury neutrino observatory. Phys Rev C 72:055502

Aharmim B et al (2010) Low-energy-threshold analysis of the Phase I and Phase II data sets of the sudbury neutrino observatory. Phys Rev C 81:055504

Ahn MH et al (2006) Measurement of neutrino oscillation by the K2K experiment. Phys Rev D 74:072003

Altmann M et al (2001) Solar nutrinos. Rep Prog Phys 64:97

Altmann M et al (2005) Complete results for five years of GNO solar neutrino observations. Phys Lett B 616:174

Ankowski A et al (2010) Energy reconstruction of electromagnetic showers from decays with the ICARUS T600 liquid argon TPC. Acta Phys Pol B41:103

Anselmann P et al (1992) Solar neutrinos observed by GALLEX at gran sasso. Phys Lett B 285:376

Anselmann P et al (1995) GALLEX solar neutrino observations: Complete results for GALLEX II. Phys Lett B 357:237

Araki T et al (2005a) Experimental investigation of geologically produced electron antineutrinos with KamLAND. Nature 436:499

Araki T et al (2005b) Measurement of neutrino oscillation with KamLAND: evidence of spectral distortion. Phys Rev Lett 94:081801

Arpesella C et al (2008b) Direct measurement of the 7Be solar neutrino flux with 192 days of borexino data. Phys Rev Lett 101:091302

Autiero L et al (2007) Large undergroung liquid based detectors for astro-particle physics in Europe: Scientific case and prospects. JCAP 0711:011

Back H et al (2006) CNO and pep neutrino spectroscopy in borexino: measurement of the deep-underground production of cosmogenic 11C in an organic liquid scintillator. Phys Rev C 74:045805

Back H et al (2008a) First real time detection of 7Be solar neutrinos by borexino. Phys Lett B 658:101

Bahcall J et al (1998) Where do we stand with solar neutrino oscillations? Phys Rev D 58:096016

Bahcall JN (2003) Solar models: An historical overview. Nucl Phys B (Proc Suppl) 118:77

Bahcall JN et al (2005) New solar opacities, abundances, helioseismology, and neutrino fluxes. Astrophys J 621:L85

Bahcall JN, Davis R Jr (1976) Solar neutrinos: a scientific puzzle. Science 191:264

Bahcall JN, Pinsonneault MH (1995) Solar models with helium and heavy-element diffusion. Rev Mod Phys 67:781

Bellini G et al (2010a) Observation of geo-neutrinos. Phys Lett B 687:299

Bellini G et al (2010b) Measurement of the solar 8B neutrino rate with a liquid scintillator target and 3 MeV energy threshold in the Borexino detector. Phys Rev D 82:033006

Berezinsky V et al (2000) Vacuum oscillations and excess of high energy solar neutrino events observed in superkamiokande. Astropart Phys 12:299

Beuthe M (2003) Oscillations of neutrinos and mesons in quantum field theory. Phys Rep 375:105

Bilenky SM, Pontecorvo B (1978) Lepton mixing and neutrino oscillations. Phys Rep 41:225

Bionta RM et al (1987) Observation of a neutrino burst in coincidence with supernova 1987A in the large magellanic cloud. Phys Rev Lett B 58:1494

Cleveland BT et al (1995) Update on the measurement of the solar neutrino flux with the homestake chlorine detector. Nucl Phys B (Proc Suppl) 38:47

Cowan CL et al (1956) Detection of the free neutrino: a confirmation. Science 124:103

Davis R (1994) A review of the homestake solar neutrino experiment. Part Nucl Phys 32:13

Davis R (1996) A review of measurements of the solar neutrino flux and their variation. Nucl Phys B (Proc Suppl) 48:284

Davis R Jr (1964) Solar neutrinos. II. Experimental. Phys Rev Lett 12:303

de Bellefon A et al MEMPHYS:A large scale water cerenkov detector at Fréjus. hep-ex/0607026

Eisele F (1986) High energy neutrino interactions. Rep Prog Phys 49:233

Fermi E A (1934) An attempt of a theory of beta radiation. Z Phys 88:161

Fiorentini G et al (2007) Geo-neutrinos and earth's interior. Phys Rept 453:117

Fukuda Y et al (1999) Constraints on neutrino oscillation parameters from the measurement of day-night solar neutrino fluxes at super-kamiokande. Phys Rev Lett 82:1810

Fukuda Y et al (2002) Determination of solar neutrino oscillation parameters using 1496 days of super-kamiokande-I data. Lett B 539:179

Gavrin VN et al (2003) Measurement of the solar neutrino capture rate in sage. Nucl Phys B (Proc Suppl) 118:39

Gribov VN, Pontecorvo B (1969) Neutrino astronomy and lepton charge. Phys Lett B 28:493

Hall LJ (1999) Phys Lett B463:241

Hampel W et al (1996) GALLEX solar neutrino observations: results for GALLEX III. Phys Lett B 388:384

Hampel W et al (1999) GALLEX solar neutrino observations: results for GALLEX IV. Phys Lett B 477:127

Heusser G (1995) Low radioactivity background techniques. Annu Rev Nucl Part Sci 45:543

Hill J (1995) An alternative analysis of the LSND neutrino oscillation search data on $\bar{\nu}\mu \rightarrow \bar{\nu}e$. Phys Rev Lett 75:2654

Hirata K et al (1987) Observation of a neutrino burst from the supernova SN1987A. Phys Rev Lett 58:1490

Kirsten T (1999) Solar neutrino experiments: results and implications. Rev Mod Phys 71:1213

Kodama K et al (2001) Observation of tau neutrino interactions. Phys Lett B 504:218

Koshiba M (1992) Observational neutrino astrophysics. Phys Rep 220:229

MiniBoone collab. (2010) Event excess in the MiniBooNE search for $\bar{\nu}\mu \rightarrow \bar{\nu}e$ oscillations. arXiv:1007.1150v2

MINOS collab. (2010) Phys Rev Lett 105:151601

Nakahata M (2005) Super-Kamiokande's solar neutrino results. Nucl Phys B (Proc Suppl) 143:13

Naumov DV, Naumov VA (2010) A diagrammatic treatment of neutrino oscillations. J Phys G: Nucl Part Phys 37:105014

Niu K et al (1971) A possible decay in flight of a new type particle. Prog Theor Phys 46:1644

OPERA collab. (2010) Observation of a first ν_τ candidate in the OPERA experiment in the CNGS beam. Phys Lett B691:138

Pandola L et al (2004) Neural network pulse shape analysis for proportional counters events. Nucl Instr Meth A 522:521

Peltoniemi J (2009) Liquid scintillator as tracking detector for high-energy events. arXiv: 0909.4974

Pontecorvo B (1968) Neutrino experiments and the problem of conservation of leptonic charge. Sov Phys JETP 26:984

Presani E (2009) Antares completed: First selected results. Nucl Phys B (Proc Suppl) 188:270

Reines F (1979) The early days of experimental neutrino physics. Science 203:11

Reines F, Cowan CL (1956) The neutrino. Nature 178:446

Reines F, Cowan CL Jr (1953a) A proposed experiment to detect the free neutrino. Phys Rev 92:492

Reines F, Cowan CL Jr (1953b) Detection of the free neutrino. Phys Rev 92:830

Requejo OM et al (2005) Super-NOVA: A long-baseline neutrino experiment with two off-axis detectors. Phys Rev D72:053002

T2K collab. (2009) The T2K experiment at J-parc. arXiv:0910.4211

Turck-Chièze S (2001) Solar neutrino emission deduced from a seismic model. Astro J 555:69

Wolfenstein L (1978) Neutrino oscillations in matter. Phys Rev D 17:2369

Wurm M et al (2010) The physics potential of the LENA detector. Acta Phys Pol B41:1749

www.km3net.org

Zacek G et al (1986) Neutrino-oscillation experiments at the Gösgen nuclear power reactor. Phys Rev D 34:9

Zacek G (1986) Ph.D. thesis, Technische Universität München

15 Scintillation Counters

Zane W. Bell
Oak Ridge National Laboratory, Oak Ridge, TN, USA

1 Introduction 350

2 Characteristics of Scintillators 351
2.1 Interaction of Radiation with Scintillators 351
2.2 Processes Governing the Generation and Decay of Light Pulses 355
2.3 Resolution 358
2.4 Considerations in Matching Scintillators to Photosensors 361

3 Scintillators 364
3.1 Inorganic Crystals 364
3.2 Organic Scintillators 369

4 Conclusions 373

5 Cross-References 373

References 373

Further Reading 374

C. Grupen, I. Buvat (eds.), *Handbook of Particle Detection and Imaging*,
DOI 10.1007/978-3-642-13271-1_15, © Springer-Verlag Berlin Heidelberg 2012

Abstract: Scintillators find wide use in radiation detection as the detecting medium for gamma/X-rays, and charged and neutral particles. Since the first notice in 1895 by Roentgen of the production of light by X-rays on a barium platinocyanide screen, and Thomas Edison's work over the following 2 years resulting in the discovery of calcium tungstate as a superior fluoroscopy screen, much research and experimentation have been undertaken to discover and elucidate the properties of new scintillators. Scintillators with high density and high atomic number are prized for the detection of gamma rays above 1 MeV; lower atomic number, lower-density materials find use for detecting beta particles and heavy charged particles; hydrogenous scintillators find use in fast-neutron detection; and boron-, lithium-, and gadolinium-containing scintillators are used for slow-neutron detection. This chapter provides the practitioner with an overview of the general characteristics of scintillators, including the variation of probability of interaction with density and atomic number, the characteristics of the light pulse, a list and characteristics of commonly available scintillators and their approximate cost, and recommendations regarding the choice of material for a few specific applications. This chapter does not pretend to present an exhaustive list of scintillators and applications.

1 Introduction

Scintillators have been used by the radiation detection community since the development of radiation generators and the discovery of radioactivity. Among the conditions of employment at an early nuclear physics laboratory was the requirement of good eyesight. Laboratory assistants were expected to examine zinc-sulfide screens under magnifiers to count the number of blue-green flashes they observed when the screens were exposed to alpha particles. Prospective employees sat in a dark room until their eyes acclimated to the darkness and were then presented with screens and sources to determine their visual acuity. Those failing this practicum were encouraged to seek other careers. Today, of course, electronic means of detecting the light exist, and students and researchers with relatively poor eyesight are free to pursue careers in nuclear physics, high energy physics, medical physics, health physics, and other areas making use of radiation detection. The radiation detection community continues to make heavy use of scintillators; even zinc sulfide continues to be used.

In the early days, the discovery of scintillators was often accidental. Roentgen found barium platinocyanide while investigating the effects of different materials on the absorption of X-rays. This material glowed when exposed to a nearby X-ray tube. Thomas Edison systematically examined crystal and mineral collections to find those that fluoresced when irradiated. Many of the first materials to be evaluated were those that were known to fluoresce under ultraviolet light (carbon arcs lights and gas discharge tubes had been developed by 1880). Most materials discovered in these ways had long decay times and were visible to the naked eye. At the time, these were important characteristics because electronic photosensors did not become available until the 1930s. The first application of photomultiplier tubes to scintillation counting was reported by Curran and Baker in a confidential report related to the war effort in 1944, but not published until 1948.

By 1950, a number of scintillators (NaI:Tl was shown to scintillate in 1948, Hofstadter (1948)) had been demonstrated, and the materials that have been examined since then number in the thousands. Of those thousands, however, approximately 20 are in commercial production and

only a few have light yields greater than NaI:Tl, the most recent of those being the lanthanum halides, which are commercially available, and SrI_2:Eu (Cherepy et al. 2009), which is not commercially available as of 2010. Research on the physical processes responsible for scintillation is ongoing as is research to discover new, brighter, and more linear scintillators.

2 Characteristics of Scintillators

In this section, a general description of the properties of scintillators is given. The interaction of gamma rays, neutrons, and charged particles with scintillators, and the effects of atomic number and mass density are briefly described. The processes governing the generation of light are enumerated and characteristics of the decay of the light pulse are discussed. The section concludes with a brief discussion of the considerations in play when matching a scintillator to a photosensor.

2.1 Interaction of Radiation with Scintillators

Scintillators depend on the production of charged particles within the scintillating medium to generate excitation resulting in the emission of light. Consequently, they detect radiation capable of producing ionization. Incident electromagnetic radiation typically ionizes atomic electrons or creates electron–positron pairs, which, in turn, ionize other atoms as they lose energy. Thus, gamma and X-rays are detected by a two-step process (here, step refers to an event leading to energy transferred to the scintillator) in which a scattering or photoelectric event must occur first, and the resulting charged particle (or particles, if there are multiple scatterings) then ionizes the scintillator. Light is produced from excitation of the scintillator occurring within microns of the track of the charged particles. On the other hand, incident electrons, positrons, and other charged particles interact directly with electrons in the scintillator, resulting in a one-step process. Neutrons interact only through nuclear interactions and produce gamma rays by capture and inelastic scattering (three steps: capture/scattering, gamma produces electrons, electrons ionize scintillator), charged particles (two steps: neutron scatters a proton or knocks out a proton or alpha particle, charged particle ionizes the scintillator), or, if sufficiently energetic, additional neutrons (at least three steps: (n, xn) first, then neutrons participate in capture, inelastic scattering, and charged-particle reactions).

The above description applies to what might be called "pure" materials. When a material comprises an inert medium containing particles of scintillator (such as a composite of micro- or nanocrystals of a scintillating compound suspended in a plastic or otherwise non-scintillating material) or a solution of scintillating molecules in an inert carrier (such as diphenyl oxazole dissolved in toluene or polystyrene), an additional step in the production of scintillation light almost always occurs: The inert medium that can make up more than 90% of the total will be ionized and that energy must migrate through the inert component to the scintillating component. This step can be important when the light yield of a composite is of prime importance because electrons lose energy traversing inert medium, which leaves less energy for excitation of the scintillating component.

The ranges of ionizing radiation differ widely. Between 800 keV and approximately 6 MeV, the mean free path of electromagnetic radiation ranges from 2 to 15 cm. Electrons and positrons with energy of about 1 MeV have ranges that are approximately 1 mm in most scintillators, while

heavy charged particles with energies of a few MeV have a range typically less than 15 μm. Neutrons interact via a series of scatterings with nuclei, and since the cross sections (away from resonances) are typically in the range of 1–10 barns (excluding, for the present, possibly high thermal capture cross sections), they tend to travel ten or more centimeters in many scintillators. These rough order-of-magnitude figures for mean free path are important to the understanding of the operation of scintillators because they imply that light is generated in close proximity to the track of incident charged particles, while it might be generated at multiple, widely scattered points throughout the scintillator in the case of incident neutral radiation.

Incident charged particles (α, $\beta^{-,+}$, $\mu^{-,+}$, $\pi^{-,+}$, etc.) ionize the medium by scattering electrons (the dominant interaction at lower energies) and participating in nuclear reactions. The energy loss due to the collisions of charged particles of mass M with atomic electrons can be estimated with the well-known Bethe formula

$$\frac{\mathrm{d}E}{\mathrm{d}x} = -\frac{4\pi z^2 e^4}{m_\mathrm{e}\beta^2} nZ \left[\ln\left(\frac{2m_\mathrm{e}\beta^2}{I\left(1-\beta^2\right)}\right) - \beta^2\right], \tag{1}$$

where ze is the charge of the incident particle, n is the number density of atoms in the medium, Z is the atomic number of the medium, m_e is the electron rest mass, β is v/c, and I is the average excitation potential of the medium. Equation 1 is in units of energy per unit length and is valid for $m_\mathrm{e} \ll M$ and $\gamma \approx 1$. Equation 1 accounts only for collisions between the incident particles and electrons and does not include shell or density corrections. It is not correct for incident electrons and positrons. However, a similar expression was derived by Bethe and is used in Monte Carlo codes such as Geant4 and MCNP5 for electrons,

$$\frac{\mathrm{d}E}{\mathrm{d}x} = -\frac{4\pi e^4}{m_\mathrm{e}\beta^2} nZ \left[\ln\left(\frac{m_\mathrm{e}\beta^2 T}{2I^2\left(1-\beta^2\right)}\right) - \ln(2)\left(2\sqrt{1-\beta^2} - 1 + \beta^2\right) + \left(1-\beta^2\right) + \frac{1}{8}\left(1-\sqrt{1-\beta^2}\right)^2\right], \tag{2}$$

and for positrons,

$$\frac{\mathrm{d}E}{\mathrm{d}x} = -\frac{4\pi e^4}{m_\mathrm{e}\beta^2} nZ \left[\ln\left(\frac{m_\mathrm{e}\beta^2 T}{2I^2\left(1-\beta^2\right)}\right) - \ln\left((\gamma-1)^2\right) - (\gamma-1)^2 \left(3 + \frac{3}{2}\frac{\gamma-1}{\gamma+1} - \frac{1-\frac{1}{3}(\gamma-1)^2}{(\gamma+1)^2} - \frac{(\gamma-1)}{(\gamma+1)^3}\left(1-\frac{(\gamma-1)^3}{6}\right)\right)\right], \tag{3}$$

where $\gamma^2 = 1/(1-\beta^2)$ and T is the kinetic energy of the incident particle.

The stopping power and CSDA (continuous slowing down approximation) range for electrons in CsI ($4.5\ \mathrm{g/cm^3}$) as computed by MCNP5 is shown in *Fig. 1*. Below approximately 12 MeV, collisional losses account for most of the stopping power. However, bremsstrahlung becomes dominant above that, increasing proportionally to the energy. The CSDA range is obtained by integrating Eq. 2 and does not account for deviations from straight-line motion caused by scattering.

At energies below the critical energy E_c, the energy at which ionization and radiative losses are equal, ionization dominates. The critical energy is ~6 MeV in uranium, varies approximately as $610\ \mathrm{MeV}/(Z+1.24)$ in solids and $710\ \mathrm{MeV}/(Z+0.92)$ in gases. For applications involving natural radioactivity and electrons under ~10 MeV, bremsstrahlung is not an important energy-loss process. At higher energies, and in applications requiring precise knowledge of the track of an electron or positron, bremsstrahlung must be considered because the mean free path of photons, possibly being much larger than that of charged particles, implies that the process may

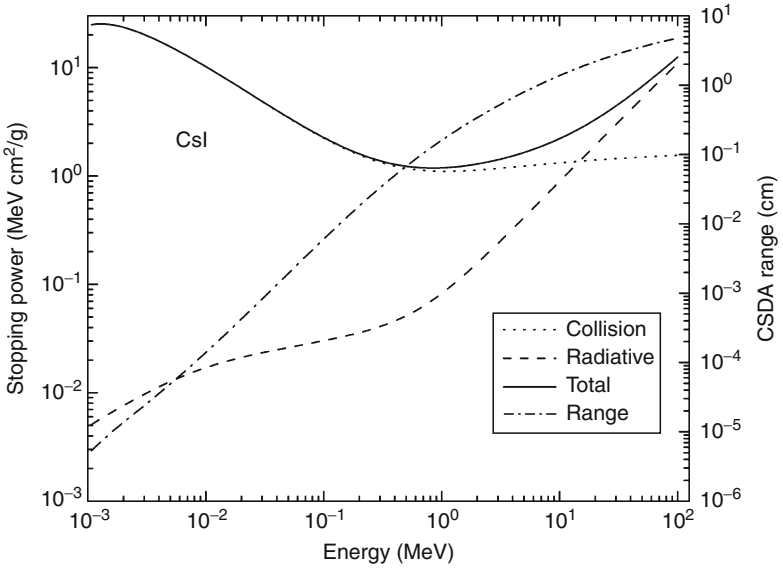


Fig. 1
Stopping powers and range for electrons in CsI

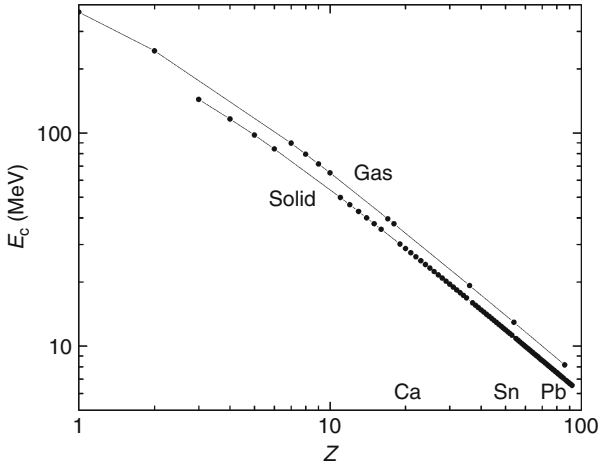


Fig. 2
Critical energies for electrons

deposit energy relatively far from the actual electron or positron track. The calculated critical energy as a function of atomic number is shown in *Fig. 2*.

Since bremsstrahlung dominates at energies above the critical energy, and the stopping power of this process increases approximately proportionally to the electron's energy, the energy of an electron after traversing a distance x (expressed in g/cm^2) is conveniently expressed by

$$E = E_0 \, \mathrm{e}^{-\frac{x}{X_0}} ,$$

where X_0 is the radiation length (usually measured in g/cm^2), and E_0 is the initial energy of the electron. This characteristic distance is the same as the 7/9 of the mean free path for pair production by high-energy photons. The radiation length has been parameterized in terms of atomic number and mass as

$$X_0 = \frac{716.4\,A}{Z(Z+1)\ln\left(\frac{287}{\sqrt{Z}}\right)}\,(\mathrm{g/cm^2})$$

and is the appropriate parameter to consider when selecting a scintillator for electrons above the critical energy.

Also of interest to designers of detectors for high energy physics experiments is the Molière radius, which is the transverse distance containing 90% of the energy of an electromagnetic shower. It is given by $R_M = 0.03475\,X_0(Z + 1.24)$, for solids. A small Molière radius, like a small radiation length, is desirable to minimize the size of the scintillator. Calculated radiation lengths and Molière radii for the elements are shown in ❯ *Fig. 3*.

Both curves were calculated as if all elements are solids; the reader should recalculate the points for gases. Note that the scale for the Molière radius is linear, while the scale for the radiation length is logarithmic.

Photons interact with matter via atomic photoelectric absorption, coherent scattering from atoms (Rayleigh scattering), Compton scattering, and pair production. Each interaction is most important in a different energy regime with photoelectric absorption dominating below ~500 keV, Compton scattering dominating between ~800 keV and 3 MeV, and pair production dominating above ~10 MeV. In units of g/cm^2 (i.e., divide by the density to obtain the mean free path in centimeters), the mean free path of photons in a number of scintillators is shown in ❯ *Fig. 4*.

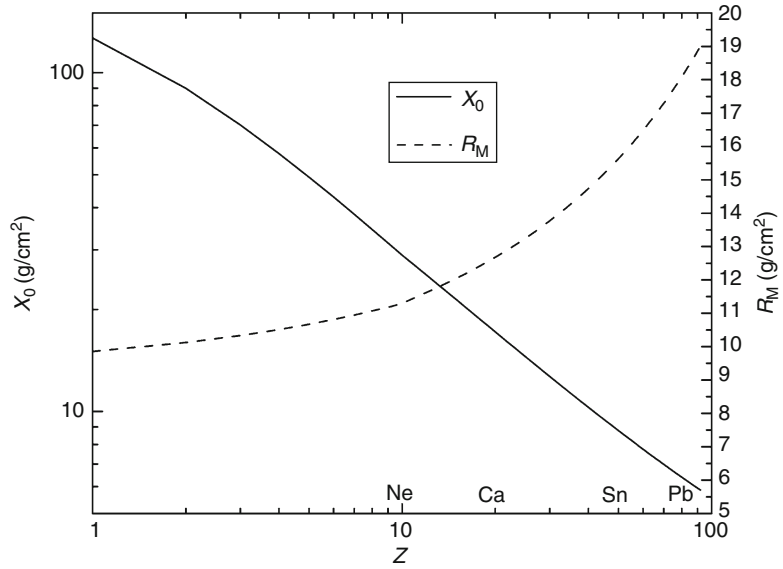


Fig. 3
Radiation lengths and Molière radii for elements

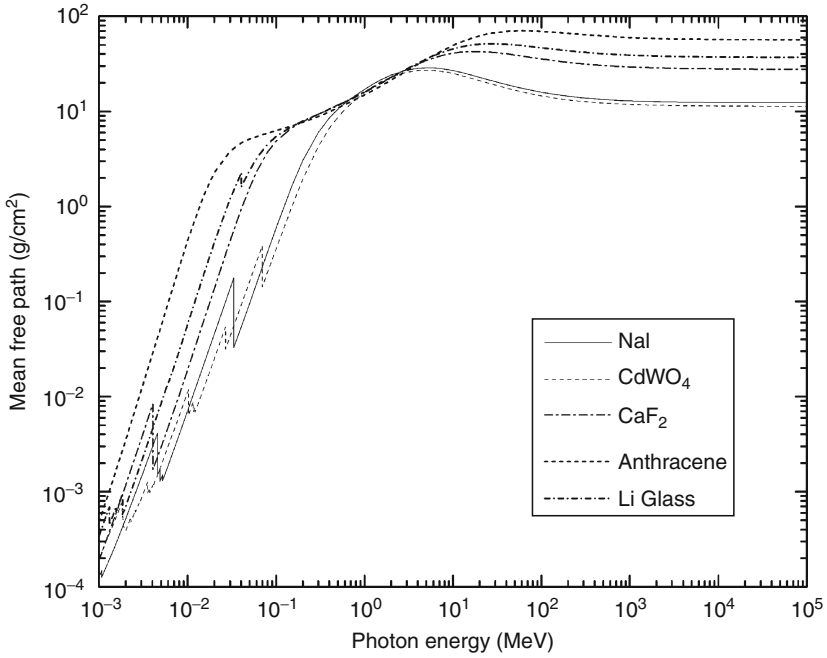


Fig. 4
Photon mean free path for selected scintillators

A few features of the curves are noteworthy. The K, L, M, etc., edges account for the jagged shapes below 100 keV. The interaction cross section falls until the photon becomes sufficiently energetic to eject an atomic electron from its shell, at which point, the mean free path drops abruptly. As the photon energy continues to rise, the photoelectric absorption cross section falls approximately like $E^{-3.5}$, while the Compton cross section takes over. Between 800 keV and ~3 MeV, attenuation is governed primarily by Compton scattering. Since the Compton cross section per atom is proportional to Z, and the number density of atoms is proportional to ρ/Z to the extent that atomic mass is proportional to atomic number, the linear attenuation coefficient is approximately independent of Z. This is seen between ~600 keV and 3 MeV in *Fig. 4*, where all the curves come together. An energy dependence remains, however, because the Compton cross section has an $\ln(E)/E$ dependence. The curves take on the characteristics of the pair-production cross section and flatten at high energies.

2.2 Processes Governing the Generation and Decay of Light Pulses

We reviewed in the previous section how various particles interact with materials. That discussion covered only the deposition of energy in the scintillator, but not the processes by which that energy is converted into light. That topic is the subject of the present section.

In all cases (crystalline and noncrystalline scintillators), the incident particle first transfers energy via collisions with electrons. These electrons typically have energy well in excess of their binding energy and are termed "primary" electrons. Primary electrons leave behind primary holes; both primary electrons and holes are hot in the sense that the temperature corresponding to their average energies is well above that of the remainder of the scintillator. The next step is the "cooling" or relaxation of these primary electrons and holes by additional collisions with atomic electrons, producing secondary electrons, holes, X-rays, plasmons, and other electronic excitations. The third step is the cooling of the secondary excitations until further ionization is not possible, resulting in a number of electron–hole pairs. The time required to reach this level of thermalization is between 1 and 100 fs. Strictly speaking, electron–hole pairs and band gap are applicable to crystals; however, the concept of a band gap applies to individual organic scintillator molecules, which, because of the presence of delocalized electrons (π electrons) confined to cyclic or polycyclic structures, also exhibit a band structure.

Once the energy of the electrons and holes is insufficient to produce further ionization, they begin to interact with the scintillator medium via electron–phonon relaxation (essentially coupling electron and hole motion to vibrations of the lattice or molecules). This process lowers the energies of the electron–hole pairs to the band-gap energy, E_g. The number of pairs surviving to this stage is given by E_0/kE_g, where k varies between 1.5 and 2.0 for ionic crystals, 3–4 for covalently bonded materials, and 1–2 for noble-gas scintillators. The last two steps of the scintillation process are the transfer of energy from thermalized electron–hole pairs to luminescence centers (typically found in inorganic scintillators), and the de-excitation of the excited luminescence centers.

Often the energy efficiency, ε, of a scintillator is described as the product of the efficiency for conversion of incident energy to electron–hole pairs ($1/k$ in the previous discussion), the efficiency, S, of the transfer of energy to luminescent centers (typically found in inorganic scintillators), and the efficiency, Q, with which the excited luminescent center generates light. The energy efficiency and the light yield, L, are related by

$$\varepsilon = \frac{\langle h\nu \rangle \, n_{\mathrm{ph}}}{E_0} = \langle h\nu \rangle \, L \,,$$

where $\langle h\nu \rangle$ is the average energy of emitted scintillation photons and n_{ph} is the number of scintillation photons emitted. For most purposes, the average energy of the scintillation photons can be taken to be the energy of the peak of the emission spectrum. Assuming that each electron–hole pair produces $S \cdot Q$ scintillation photons, and knowing the number of electron–hole pairs produced (E_0/kE_g), the energy efficiency and light yield can be written as

$$\begin{aligned} \varepsilon &= \frac{\langle h\nu \rangle \, SQ}{kE_g} \,, \\ L &= \frac{SQ}{kE_g} \,. \end{aligned} \tag{4}$$

Since SQ can be at most unity, with the band gap expressed in eV, L has an upper bound of about $660/E_g$ photons/keV for ionic crystals, $330/E_g$ photons/keV for covalent materials, and $1/E_g$ photons/keV for noble gases. Thus, CsI:Tl and NaI:Tl should produce no more than about 110 photons/keV. In actuality, under gamma-ray or electron irradiation at energies between 500 keV and 1 MeV, CsI:Tl produces approximately 65 photons/keV, and NaI:Tl about 40 photons/keV, implying their energy efficiencies are 14% and 12%, respectively. The energy efficiencies do not track the light yields because the wavelengths of peak emission for the two scintillators

are different. The energy efficiency of $CdWO_4$, which has a spectrum similar to CsI:Tl, but a slightly smaller band gap, is 5.3%, implying the transfer and transport of energy in that intrinsic (requiring no activator) scintillator crystal is somewhat poorer than in CsI:Tl.

The time characteristics of the light pulse from a scintillator are determined by the kinetics of the energy transfer, the index of refraction, n, and the oscillator strength of the scintillating transition. The decay time constant of systems with a single type of luminescent center is inversely proportional to the square of the energy of the transition, and also approximately proportional to $10^{-(n/1.2)}$ for n in the range 1.4–2.5. Although the implication is that higher index of refraction can lead to faster scintillators, the section on matching scintillators to photosensors, below, shows that high indices of refraction lead to light trapping and inefficient transfer of light to the photosensor.

Some scintillators have multiple components in their light pulse, and this is a consequence of multiple light-emitting states. Notable among these scintillators are those exhibiting pulse shapes that vary according to the identity of the ionizing radiation. BC501A (an organic liquid scintillator manufactured by Saint-Gobain Crystals) has three components in its light pulse, with time constants of 3, 30, and 270 ns. The proportion of light emitted in the 270 ns component is measurably larger under heavy-ion irradiation (most work has been done with protons scattered by fast neutrons) than under electron or gamma irradiation. The phenomenon occurs because in organic scintillators, ionizing particles produce both spin singlet and triplet excitations of the phenyl rings, and transitions between the different spin states are severely suppressed by dipole selection rules. However, collisions between molecules can effect such transitions, and the increased density of excited molecules produced by heavy particles enhances singlet-to-triplet transitions, while triplet states cannot decay promptly to the singlet ground state and must wait until a second collision transfers energy back to an excited singlet state.

Of recent interest has been the elpasolite material Cs_2LiYCl_6:Ce and its variants. This crystal exhibits 1, 50, and 1,000 ns decay components, but the 1 ns component disappears when irradiated by neutrons. Neutrons participate in the $^6Li(n, \alpha)t$ reaction and so produce a 2 MeV α and a 2.7 MeV triton. The 1 ns component has been attributed to core–valence luminescence (excitation of a core-electron to the conduction band of the crystal with a valence-band electron subsequently filling the hole in the core), but no theory has yet been proposed to adequately explain why this process is suppressed by high ionization density. Perhaps, the high number of ionized electrons enhances the probability that the core holes are filled by them non-radiatively.

Much of the previous discussion has revolved around inorganic crystal scintillators. Organic scintillators differ from inorganic crystals in that the periodic structure established by the crystal lattice is established by the organic molecules' polycyclic structure and delocalized electrons. It is not necessary that organic molecules be arranged as crystals; the process is at the molecular level instead. This property enables the fabrication of organic-liquid (toluene, xylene, and mineral-oil-based) and polymer scintillators (polyvinyltoluene, polystyrene, polysiloxane-based), as well as organic crystals (stilbene, anthracene). The generation of light in these materials does not depend on the presence of an activator. Rather, by adjusting the number and arrangement of phenyl rings, the molecule's structure generates a set of energy levels whose transitions have wavelengths typically ranging from 360 to 500 nm.

When organic scintillators are dissolved at the level of 1–10% by weight in a solid polymer or liquid solvent containing a large number of phenyl rings, the composite material continues to exhibit the characteristics of the pure scintillator. This is because of two mechanisms. Most ionization is of the solvent, and benzene-like solvents will fluoresce with the emission of UV

light with a time constant of about 16 ns. However, energy transport by resonant dipole–dipole interactions between nearby phenyl rings, first proposed by Förster in 1948, is an efficient transporter of energy over distances of 0.1 nm, which is approximately the average distance between molecules in a scintillating solution (also termed as "cocktail"). There is no charge transported by this process, which occurs in about 1 ns. The decay time constant of the cocktail is significantly shortened by the action of resonant interactions. Organic scintillators typically exhibit multiple decay time constants, with the shortest being that of the scintillating solute and of the order of 3–5 ns.

2.3 Resolution

A scintillator's resolution is most often quoted as the ratio of the full width at half-maximum (FWHM) of the peak in the pulse-height spectrum due to the 662 keV gamma ray from ^{137}Cs to the centroid of the peak. That is,

$$R = \frac{\mathrm{FWHM(keV)}}{662\ \mathrm{keV}}\,.$$

In some applications, though, the resolution is quoted at a different energy, usually in the range of interest to the application. Resolution is a function of the number of electrons delivered by the photosensor to the readout and conversion electronics (and, therefore, a function of deposited energy) and the noise in the electronics. We will concentrate on the delivered electrons in this section and include electronics noise as a single term that adds in quadrature. Electronics noise can be measured by injecting a precision pulser into the stream of analog data.

The number of electrons delivered to the electronics is a stochastic variable that depends on several statistically independent (or nearly so) factors whose variances add in quadrature. The number of electrons, N_e, delivered to the electronics is given by

$$N_\mathrm{e} = N_\mathrm{ph} \cdot \varepsilon_\mathrm{t} \cdot G\,,$$

where N_ph is the number of scintillation photons, ε_t is the efficiency for creating photoelectrons in the readout device, and G is the gain of the readout. A silicon PIN photodiode has a gain of unity; a photomultiplier can have a gain of 10^7. In the case of a photomultiplier, the gain is, theoretically, the product of the gains of each stage, not all of which need to be equal. However, for convenience, G will be taken to be

$$G = \prod_{i=1}^{N} g_i = g^N.$$

Although in practice the exponent is not the number of stages, but usually closer to $N - 2$, the value N will be retained below. In addition, it is convenient to set all the g_i equal. Energy dependence enters via the number of scintillation photons.

If the amplitude of the light pulse is sufficiently small, the readout device remains linear and does not saturate, making these processes statistically independent. Consequently, the relative variance associated with N_e is then given (to first order) by

$$R^2 = \left(\frac{\sigma_{N_\mathrm{e}}}{N_\mathrm{e}}\right)^2 = \left(\frac{\sigma_{N_\mathrm{ph}}}{N_\mathrm{ph}}\right)^2 + \left(\frac{\sigma_{\varepsilon_\mathrm{t}}}{\varepsilon_\mathrm{t}}\right)^2 + \left(N\frac{\sigma_g}{g}\right)^2. \qquad (5)$$

The mean number of scintillation photons in Eq. 5 depends on the band-gap energy, the efficiency of the transfer of energy to luminescent centers, and the quantum efficiency of the luminescent centers. The variance of the number of scintillation photons comes from the (Poisson) statistics associated with the number of scintillation photons, intrinsic non-proportionality of the scintillator, and nonuniformities in the material. The latter is caused, for example, by inhomogeneous distribution of activators (the Tl present in NaI:Tl or the mixing of organic fluors in plastics, for examples), defects in the physical structure of the scintillator (lattice vacancies, inclusions, incomplete polymerization of plastics, and phase separations, for example), and inhomogeneous distribution of unintentional impurities.

Non-proportional response is an intrinsic property of the scintillator and is caused by the energy dependence of the probability of the excitation of luminescent states by electrons. This means that the number of electron–hole pairs created per unit deposited energy, taken to be a constant in the previous section, is actually a function of energy. Birks, noting that the response of scintillators decreased with increasing ionization density along a particle's track, proposed that the light yield per unit path length can be parameterized by

$$\frac{\mathrm{d}L}{\mathrm{d}E} = \frac{L_0}{1 + k_\mathrm{B}\frac{\mathrm{d}E}{\mathrm{d}x}},$$

where $\mathrm{d}E/\mathrm{d}x$ is given by Eqs. 1–3, and L_0 (related to the constant k in Eq. 4) and k_B are constants that depend on the scintillator. This equation predicts that for large stopping power, L will be inversely proportional to the range of the radiation in the scintillator, which is not proportional to the energy of the electrons. For small stopping power, such as for electrons between ~400 keV and ~3 MeV (see *Fig. 1* for the stopping power of CsI:Tl), L will be proportional to the energy of the radiation. It has been observed, however, that for electrons, although the light yield is decidedly nonlinear, it does not follow the Birks equation.

In the case of NaI:Tl, when the non-proportionality is defined to be the ratio of the light yield at energy E to that at 480 keV (the energy of Compton electrons generated by ^{137}Cs gamma rays scattered 180°) the ratio rises to 1.2 near 15 keV. This means that a 15 keV electron produces

$$L(15\,\mathrm{keV}) = \frac{1.20 \times L(480\,\mathrm{keV})}{32},$$

that is, 20% more light than expected from the ratio of energies. Since all light production in scintillators is caused by secondary electrons, and these secondary electrons have a continuum of energies, the distribution of electron energies is important. *Figure 5* shows Hull's measurements of the relative light yield of NaI:Tl crystals. The curve rises monotonically to a maximum near 15 keV and appears smooth. This is in disagreement with Dorenbos et al. (1995), who has reviewed measurements of the relative light yield of NaI:Tl, CsI:Tl, CsI:Na, BGO, LSO, CaF_2:Eu, CWO, GSO, YAP, LuAG, and K_2LaCl_5 and shows that the relative light yield exhibits local minima near the K edges of the constituents. Dorenbos also shows that some crystals are less than proportional at lower energies (the relative-light-yield curve is never greater than unity). The effect of variable relative light yield is to increase the variance of N_ph above what would be expected from Poisson statistics.

The second term in Eq. 5 accounts for light trapping and collection, and is derived from an average over all paths to the exit window from points of scintillations and over all points of scintillation. It accounts for variation in reflectance of materials surrounding the scintillator and the coupling between the scintillator and the entrance window of the photosensor. It also accounts for trapping within the window of the photosensor. The last term in Eq. 5 accounts

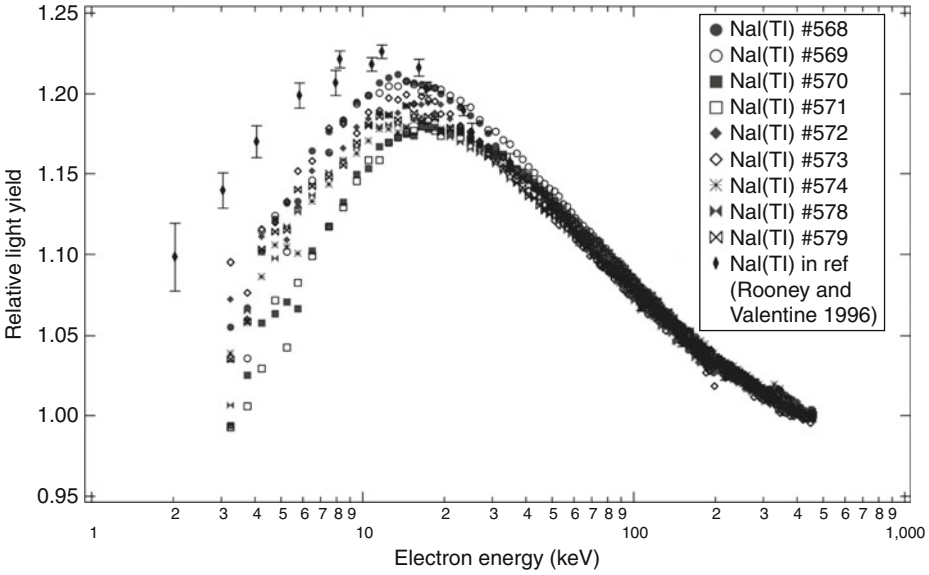


Fig. 5
Relative light yield of NaI:Tl samples. (From Hull et al. 2009; Original figure © 2008 IEEE, reprinted with permission)

for the fluctuations in the gain of the photosensor. In a typical photomultiplier, each dynode contributes a factor of ~3.5; in a silicon PIN photodiode, the total gain, G, is between ~0.35 and 1, depending on the spectrum of the light and the spectral response of the diode.

Curves similar to *Fig. 5* have been measured for many scintillators. For energies above ~100 keV, the variation in light yield is small, and the resolution is dominated by terms proportional to the number of scintillation photons, which is approximately proportional to the deposited energy. Practically, this leads to the following relationship between resolution and incident photon energy:

$$R^2 = a + bE\,,$$

where a represents the contribution of the electronics (usually small) and b depends on the scintillator and construction of the detector.

Dorenbos et al. (1995) provides a description of the effects of statistics on resolution and provides an expression for the best resolution expected from a scintillator, given the light yield. *Figure 6* shows the locus points for commercially available scintillators and their relation to the best theoretical resolution obtainable for their light yield.

Obviously, almost all scintillators lie far from the theoretical line. Notable exceptions are $LaBr_3$:Ce and YAP:Ce ($YAlO_3$:Ce). The poorer resolution point is reported by Rodnyi (1997); however, Kapusta et al. (1999) reports a remarkable 4.8% resolution for YAP, which is its theoretical best resolution. Further improvements in the performance of YAP will require improvements in photosensors and electronics. $LaBr_3$:Ce, which is close to its theoretical limit, has found wide use in the gamma spectroscopy community because of its much higher stopping power.

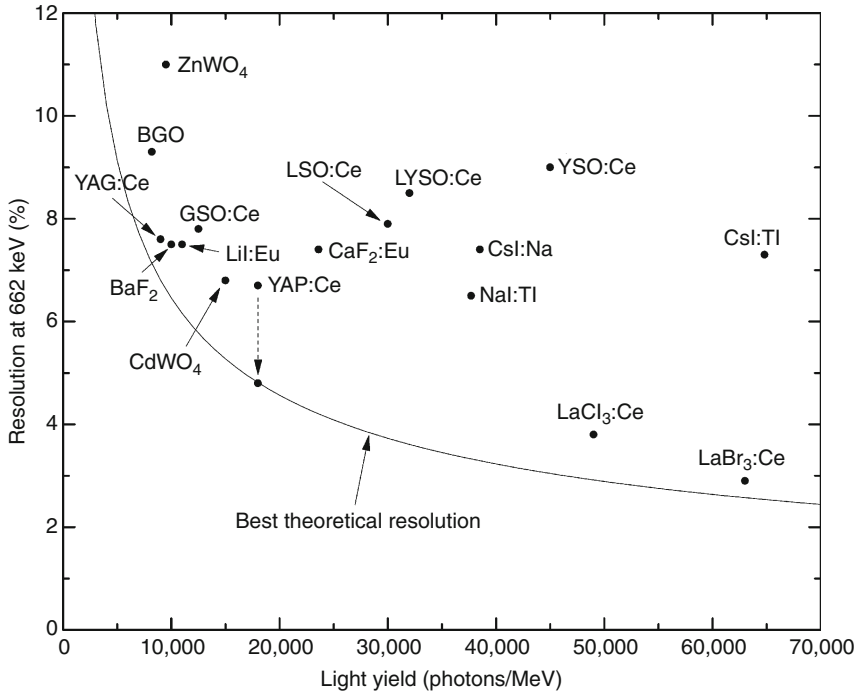


Fig. 6

Best theoretical resolution and observed resolution of scintillators. A photomultiplier quantum efficiency of 20% was used in the theoretical calculation

2.4 Considerations in Matching Scintillators to Photosensors

All scintillators rely on the energy bands of the constituent materials having transitions with energies ranging from approximately 2–6 eV. These transitions result in the emission of light with wavelengths between 620 and 200 nm, which is well matched to the sensitivities of photomultiplier tubes and photoconductors. The relationship between transition energy and emission wavelength is given by

$$\lambda(\text{nm}) = \frac{1240}{E(\text{eV})}, \tag{6}$$

where λ is the wavelength in nm and E is the energy in eV. Equation 6 is of importance in the selection of the scintillator and readout device because it is necessary to match the scintillator's emission band to the sensitive band of the readout to maximize the number of photoelectrons or electron–hole pairs. A photosensor's sensitivity, $S(\lambda)$, measured in mA/W, gives the current liberated per unit incident power at a specified wavelength. The quantum efficiency, $QE(\lambda)$, gives the average number of photoelectrons generated per photon absorbed by the photosensor, and is easily found from Eq. 6 to be

$$QE(\lambda) = \frac{1.24}{\lambda} S(\lambda). \tag{7}$$

Manufacturers often quote the sensitivity of a photosensor at the wavelength of the peak of S; the designer must pay careful attention to this because the response of semiconductors is often heavily biased toward longer wavelengths. For example, the manufacturer of one photodiode quotes S to be 660 mA/W at 960 nm, making the quantum efficiency at this wavelength approximately 0.85. Examination of the device's data sheet, however, provides the information that S (420 nm), a common wavelength of maximum emission of blue-emitting scintillators, is 200 mA/W, making $QE\,(420\,\text{nm}) = 0.65$. If the designer uses the values at 960 nm to estimate the response at 420 nm, the contribution to the resolution from charge-generation statistics would be underestimated by 70%. Ideally, the response of the photosensor closely matches the emission spectrum of the scintillator.

It is also important to match the scintillator's index of refraction to that of the entrance window of the photosensor. If there is a significant difference between them, then much of the light can be trapped within the scintillator resulting in loss of light, an artificially lengthened light pulse (because of multiple bounces within the scintillator volume), or both. For normal incidence, the fraction of light energy reflected (R) from, and transmitted through (T) a dielectric interface is given by

$$R = \left(\frac{n_1 - n_2}{n_1 + n_2}\right)^2,$$

$$T = \frac{4n_1 n_2}{(n_1 + n_2)^2},$$

where n_1 and n_2 are the indices of refraction of the two media. For the case of light from a CsI crystal normally incident on a glass entrance window, $n_{\text{CsI}} = 1.79$, $n_{\text{glass}} = 1.52$, and $R = 0.0067$. However, for $CdWO_4$, $n_{\text{CWO}} = 2.25$ and $R = 0.037$. The reflection coefficient at

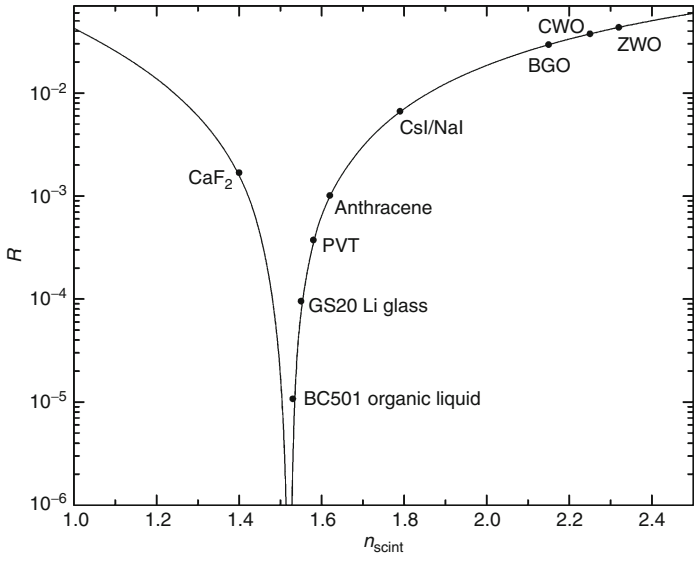


Fig. 7
Reflection coefficient at normal incidence

normal incidence for borosilicate-glass–scintillator interfaces is shown in *Fig. 7* for a continuum of indices of refraction, and the positions of representative, commercially available, scintillators are indicated.

Although the reflection coefficient at normal incidence does not seem large, the mismatch can have a significant effect on light collection. The critical angle, the angle at which light is totally internally reflected when going from a material of high index of refraction to one of lower index of refraction n, given by $\theta_C = \sin^{-1}(n/n_{scint})$, determines the "cone of acceptance" (similar in concept to the cone of acceptance of optical fibers) of the interface. Scintillation light incident on the scintillator–entrance-window interface outside the critical angle must be reflected from the interface. This lengthens the path of the light within the scintillator and increases the probability of self-absorption (caused by the overlap of the emission and absorption bands of the scintillator) by trapping light within the scintillator, and, in the case of large blocks of scintillator, can increase the decay time of the light pulse. The latter is generally important only with scintillators larger than about 30 cm and having decay times of a few nanoseconds, such as in $10 \times 10 \times 40\ cm^3$ polyvinyltoluene-based plastics (PVT) used in portal monitors. Every meter increase of the optical path length increases the apparent decay time by approximately 3 ns.

Figure 8 shows the critical angle, in degrees, calculated for a scintillator–glass interface with light entering the glass from the scintillator (left vertical axis), and the fraction of light striking the interface outside the cone of acceptance (right vertical axis). The calculation related

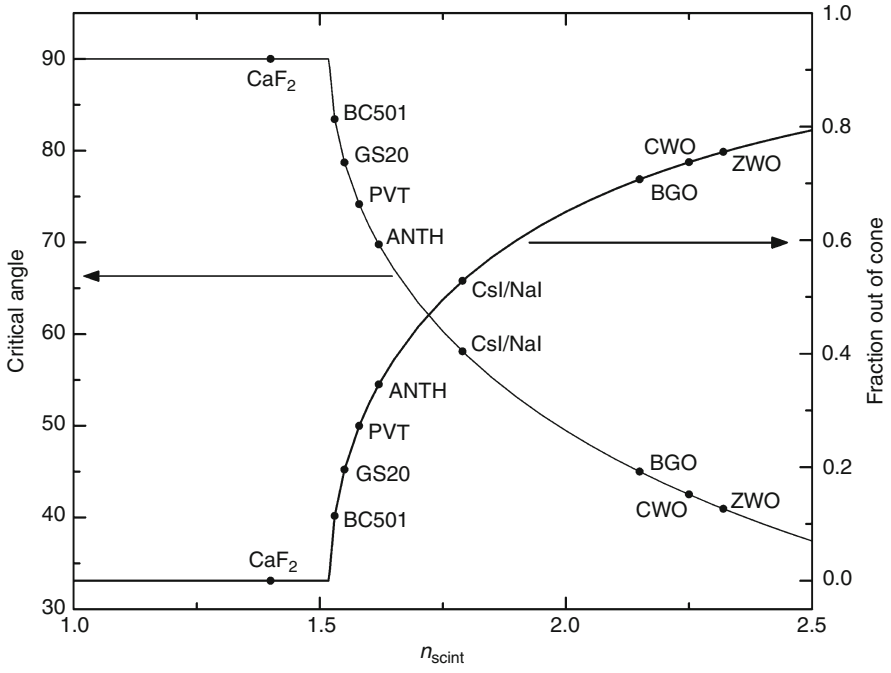


Fig. 8

Critical angle and fraction of scintillation photons emitted outside the cone of acceptance for light exiting a scintillator and entering a borosilicate-glass layer with index of refraction 1.52

to the cone of acceptance is done for an isotropic emitter above an infinite plane interface, which implies that the solid angle subtended by the interface relative to any point of scintillation is 2π.

When the scintillator's index of refraction is less than 1.52, the critical angle is 90°. Consequently, some light from $CaF_2 (n = 1.44)$ can enter a photomultiplier's faceplate regardless of the angle of incidence. However, as the index of refraction rises to that of CsI or NaI, even though only 0.7% of the light is reflected at normal incidence, approximately half the light incident on the interface is totally internally reflected; the remainder has a transmission coefficient that depends on the angle of incidence. The situation rapidly deteriorates as the index of refraction exceeds 2.0. For this reason, diffuse reflectors and index-matching couplants (such as glycerin and silicone greases) are generally used. The reader is encouraged to consult textbooks on electromagnetic theory for the formulae pertaining to reflection and transmission coefficients as a function of index of refraction and angle of incidence.

Photomultiplier tubes are available with different entrance windows. The most common (and least expensive) is the borosilicate-glass faceplate; this case has been considered above. Manufacturers, however, list MgF_2 as an optional faceplate material, especially for ultraviolet-sensitive photomultipliers. This material is not hygroscopic and has an index of refraction of approximately 1.38 in the visible region, rising to 1.78 at 114 nm. Fused quartz is also used as a faceplate; its index of refraction is 1.46 in the visible region, rising to 1.52 near 230 nm. Sapphire (Al_2O_3) is available in a more limited number of models of photomultiplier tubes, mostly for solar-blind and aerospace applications. The index of refraction of these faceplates is approximately 1.78 in the visible region, rising to 1.83 near 263 nm. It is a birefringent, non-isotropic material with coefficient of thermal expansion $4.3 \times 10^{-6}/°C$ perpendicular to and $5.4 \times 10^{-6}/°C$ parallel to the c-axis, respectively. This is in contrast to the coefficient of thermal expansion of borosilicate glass (an isotropic material) of $3.2 \times 10^{-6}/°C$ and implies that care must be taken when bonding these materials to avoid cracking due to thermal stresses during normal use. Sapphire–metal brazes are more commonly found than sapphire–glass bonds.

3 Scintillators

It is likely that thousands of materials have been tested for scintillation since the discovery of X-rays. Papers by Derenzo et al. (1991), Moses et al. (1997), and van Eijk (2001) mention nearly 600 materials, and the book by Shinonoya and Yen (1999) mentions over 1,000 materials evaluated as scintillators. Of all these, only three organic crystals and fewer than 20 inorganic crystals are commonly available. In this section, the properties and representative applications of commercially available materials are described.

3.1 Inorganic Crystals

Data in ❯ *Table 1* are compiled from the Review of Particle Physics (Amsler et al. 2008, 2009), Wilkinson (2004), Rodnyi (1997), Knoll (2000), and the web sites of Hilger Crystals, Applied Scintillation Technologies, Hitachi Chemical Company, and Saint-Gobain Crystals, Inc.

Table 1

Properties of common inorganic scintillators

Material	ρ (g/cm^3)	Emission maximum (nm)	Decay time constant	Index of refraction[a]	Light yield (ph/MeV)	Hygroscopic?
NaI:Tl	3.67	415	230 ns	1.85	37,700	Yes
CsI:Tl	4.51	550	600, 3,400 ns	1.79	64,800	No
CsI:Na	4.51	420	630 ns	1.84	38,500	Slightly
CaF_2:Eu	3.18	435	840 ns	1.47	23,600	No
^{6}LiI:Eu	4.08	470	1,400 ns	1.96	11,000	Yes
^{6}Li glass	2.6	390–430	60 ns	1.56	2,000	No
BaF_2	4.88	315/220	0.63 μs/0.8 ns	1.50/1.54	10,000/1,400	No
YAP:Ce	5.55	350	27 ns	1.94	18,000	No
YAG:Ce	4.57	550	70, 300 ns	1.82	19,700	No
LSO:Ce	7.40	420	40 ns	1.82	30,000	No
LYSO:Ce	7.10	420	40 ns	1.81	32,000	No
YSO:Ce	4.55	420	37 ns, 82 ns	1.80	45,000	No
GSO:Ce	6.71	440	56 ns, 600 ns	1.85	12,500	No
BGO	7.13	480	0.3 μs	2.15	8,200	No
$CdWO_4$	7.90	470, 540	20 μs, 5 μs	2.3	15,000	No
$PbWO_4$	8.28	420, 425	10 ns, 30 ns	2.20	100, 31	No
$ZnWO_4$	7.62	490	20 μs	2.32	9,500	No
$LaBr_3$:Ce	5.08	380	16 ns	1.9	63,000[b]	Yes
$LaCl_3$:Ce	3.85	350	28 ns	1.9	49,000	Yes

[a]Index of refraction at the wavelength of maximum emission
[b]Reports in the literature are as high as 75,000 ph/MeV

The data in *Table 2* are compiled from the Review of Particle Physics (Amsler et al. 2008, 2009), Wilkinson (2004), Rodnyi (1997), and the web sites of Hilger Crystals, Applied Scintillation Technologies, Hitachi Chemical, Bright Crystals Technology, and Saint-Gobain Crystals, Inc. The reader should take pricing information below as an approximate guide and must contact a vendor with specific requirements to obtain an accurate quote.

The inorganic scintillator crystals find application in many areas of radiation detection. NaI:Tl, although one of the first discovered scintillators, has been the workhorse of the industry. This is because of its low cost, high brightness, and moderate energy resolution of about 6%. It can be grown in large ingots, routinely seen as single $10 \times 10 \times 40\,\text{cm}^3$ crystals in detectors used for homeland security applications. This scintillator, CsI:Tl, and CsI:Na are used for general gamma-ray spectroscopy. In thin wafers, NaI:Tl is used in α/β probes in health physics instruments.

Both CsI:Tl and CsI:Na are sensitive to fast and slow neutrons. NaI:Tl is mainly used for fast neutrons (see *Table 2*). Thermal neutrons are captured by iodine, which results in the production of ^{128}I, with a half-life of 25 min. This isotope is a β–γ emitter that, when in a scintillator, produces an unexpected background that builds up according to $1 - 2^{-t/25}$, with the irradiation time, t, in minutes. The cross section for thermal neutron capture in Na is sufficiently small that a large ambient flux is required before its activity is noticed. However, when Na is

Table 2
Additional properties of common inorganic scintillators

Material	$\frac{1}{L}\frac{dL}{dT}$ [a]	Neutron sensitive?	Radiation hardness (Gy)[b]	Radiation length (cm)	Price (US$/cm^3)[c]
NaI:Tl	−0.2	Yes (F)[d]	10	2.6	$6
CsI:Tl	~0 (fast)	Yes (F, S)[e]	10	1.86	$4
CsI:Na	0.39	Yes (F, S)	10	1.86	$4
CaF_2:Eu	−0.33	No		3.50	$20
^{6}LiI:Eu		Yes (F, S)		2.55	~$100
^{6}Li glass		Yes (S)		7.09	~$1, 500
BaF_2	−1.3 (slow)	No	$>10^5$	2.03	$15
YAP:Ce		No	10^4	2.67	$100
YAG:Ce	−0.27	No		3.5	$90
LSO:Ce		Yes (S)	10^{4-5}	1.14	$60
LYSO:Ce	−0.2	Yes (S)	10^{4-5}	1.15	$70
YSO:Ce		No	10^4	2.75	~$85–100
GSO:Ce	−0.1	Yes (S)	10^6	1.38	
BGO	−0.9	No	100–1,000	1.11	$35
$CdWO_4$	−0.1	Yes (S)	10–1,000	1.06	$60
$PbWO_4$	−2.7	Yes (S)	$>10^5$	0.89	$5–6[f]
$ZnWO_4$	−1.2	Yes (S)		1.10	$40
$LaBr_3$:Ce	~0	Yes (S)	10^5	1.88	~$500
$LaCl_3$:Ce	0.7	Yes (S)	10^5	3.12	~$500

[a] Percent/°C at 20 °C. Light yield of all crystals eventually falls with increasing temperature
[b] Dose causing significant loss of transmission at the peak emission wavelength
[c] Unless otherwise noted, pricing estimates are courtesy of James Telfer of Hilger Crystals
[d] F, sensitive to fast neutrons by (n, 2n) and X-rays by (γ, n) reactions
[e] S, sensitive to slow neutrons by generating prompt-capture gamma rays and/or generating a radioactive product subsequent to capture
[f] Approximate pricing provided by Alexander Gektin of the Institute for Single Crystals, Kharkiv, Ukraine. Dr. Gektin notes that the price of $PbWO_4$ may fluctuate significantly as the demands of the high energy physics community rise and fall

activated, it produces a continuum with an endpoint of 5.5 MeV and a half-life of 15 h. The activation of Cs by thermal neutrons is not usually a problem because ^{134}Cs has a half-life of 2 years, making it unlikely that sufficient amounts will be made in most situations.

Both these crystals are sensitive to neutrons with energy above 14 MeV because of (n, xn) reactions. In the cases of Cs and I, the products of (n, 2n) reactions, ^{132}Cs and ^{126}I, are readily made by a D–T generator. *Figure 9* shows a spectrum obtained with an HPGe detector from a 10 cm diameter, 2 cm thick CsI:Tl sample exposed to a D–T generator. The sample was placed 5 cm from the anode of the generator (in the air), which was operated at approximately 10^7 n/s for 3 h. After irradiation, the sample was placed 10 cm from the front face of the HPGe detector and counted for 10 min.

The spectrum shows three prominent peaks from four gamma rays from ^{132}Cs and ^{126}I; two of the gamma rays are sufficiently closely spaced that they were not resolved by the detector. The remaining unattributed gamma rays are from natural ^{214}Bi, ^{214}Pb, ^{208}Tl, positron annihilation, ^{228}Ac, and other decay products from the ^{232}Th and ^{238}U decay chains. ^{126}I has a half-life

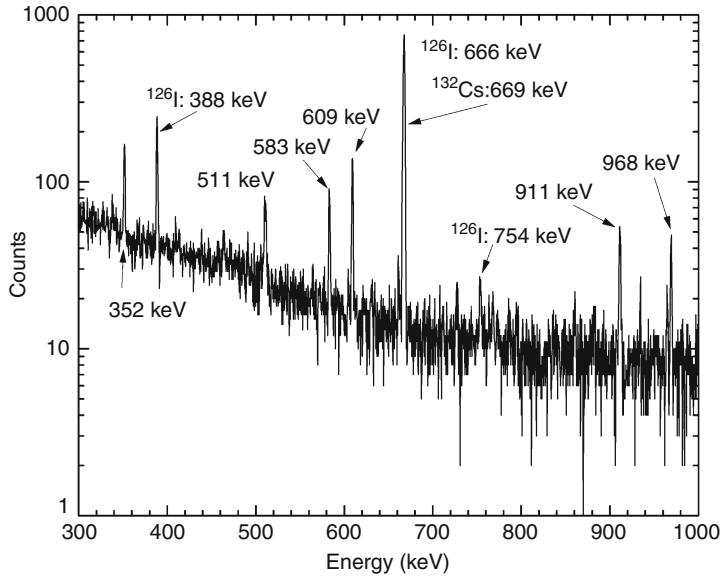


Fig. 9
Spectrum from activated CsI:Tl

of 13 days, while ^{132}Cs has a half-life of 6.47 days, implying that an activated detector should not be used for a few months.

Cesium and iodine are not the only elements that are susceptible to neutron activation. ^{186}W comprises 29% of natural tungsten and has a thermal-capture cross section of 38 barns. ^{187}W has a half-life of 23.9 hours and is a $\beta-\gamma$ emitter. ^{176}Lu, the radioactive component in LYSO and LSO, comprises 2.6% of natural Lu but has a thermal-capture cross section of 2,100 barns. ^{177}Lu has a half-life of 6.71 days and is also a $\beta-\gamma$ emitter. ^{139}La, the stable La isotope (99.9% abundance) in $LaBr_3$ and $LaCl_3$, has a 9.3 barn thermal-capture cross section. The reaction product, ^{140}La, has a 40-hour half-life and is also a $\beta-\gamma$ emitter. The thermal-capture cross section of chlorine, while high, does not lead to significant activation because of the long half-life of ^{36}Cl. On the other hand, the thermal-neutron-capture cross section of ^{79}Br is sufficiently high that in a high flux, the production of ^{80}Br (17.6 min half-life) should be noticed. Reactions with particles capable of inducing (n, 3n), (n, 4n), and spallation reactions are too numerous to be considered here. Suffice it to say that when a detector is to be used in a neutron field or exposed to energetic particles or X-rays (remember that photonuclear reactions can have the same reaction products as (n, 2n) and (n, pn) reactions), the detector designer must consider the possibility and effects of activation.

$CdWO_4$ and BGO are also used for general-purpose gamma-ray spectroscopy and have the characteristic of having extremely low background radioactivity. The latter property, coupled with their high density and average atomic number, make them attractive for use in positron emission tomography (PET). However, their relatively slow decay times limit their maximum count rates, and, therefore, the quality of imaging. The speed of LSO and LYSO, on the other hand, more than compensates for their poorer linearity and natural radioactivity in medical instruments. LSO and LYSO are not used for general gamma-ray spectroscopy because of the

radioactivity of Lu. The natural radioactivity of Lu is less of a problem in PET because a coincidence is required between a pair of 511 keV gamma rays from the patient, and the coincidence window is sufficiently short to exclude the natural β/γ activity. GSO has also been used in PET because of its high stopping power.

In addition to PET, GSO has found application in oil well logging. In this activity, a detector is lowered behind the drilling head in a module containing a strong ^{137}Cs or ^{60}Co source and a neutron generator. These sources are used for measurements of rock density and composition, and water, salt, and hydrocarbon concentration at temperatures reaching 200°C. GSO is extremely radiation resistant, and its light yield remains sufficient for spectroscopy even at high temperature. GSO is also sensitive to thermal neutrons via the ^{157,158}Gd(n,γ) reactions. In both isotopes, the reaction products are stable and only prompt-capture gammas and conversion electrons are emitted. However, the overwhelming majority of neutron-capture reactions occur within 1 mm of the surface of the crystal, preventing a peak corresponding to the Q-value of the reactions because half of the capture gammas and electrons exit the crystal.

LiI:Eu is almost exclusively used for thermal-neutron detection. Thermal neutrons impinging on the crystal are captured by ^{6}Li, which results in the deposition of 4.79 MeV split between an alpha particle and a triton. The resulting peak typically occurs at an electron equivalent energy over 3 MeV, which is higher than the vast majority of naturally occurring gamma rays. Since only a thin crystal is needed to capture thermal neutrons, the sensitivity to high-energy gammas can be limited by geometry. This scintillator is used in homeland security applications and as an alternative detector to ^{3}He in polyethylene neutron dosimeters. It is rarely, if ever, used as a replacement for ^{3}He in a mixed γ/n environment because it has a substantial gamma response.

Ce-activated Li-glass scintillator is used exclusively for thermal neutron detection. The fact that it is a glass rather than a crystal means it does not have the long-range ordering needed for the efficient transport of electrons and holes to activator sites. Consequently, it develops a continuum gamma-ray spectrum. However, a distinct peak is developed in response to the ^{6}Li(n,α) reaction (Applied Scintillator Technolgies quotes the resolution at 15–28%), and so neutron events are distinguished from gamma events by pulse amplitude. A ^{7}Li-loaded glass behaves identically to the ^{6}Li glass, except for the lack of neutron response, and is used as a "witness" detector whose counts are subtracted from the ^{6}Li glass to determine the net neutron flux from a source.

CaF_2:Eu finds application in particle detection, and low-energy X-ray detection because of its low atomic number. The crystal is rugged, inert, and not hygroscopic and can be used in more extreme conditions than some other crystals. It is bright, but the relatively long decay time precludes its use in high-count-rate applications.

BaF_2 is one of the few crystals known to exhibit core–valence luminescence, which gives it an extremely fast light component. It is an intrinsic (undoped) scintillator that is prized for its combination of timing resolution and stopping power. It has been proposed for use in PET systems and nuclear and high energy physics coincidence experiments. However, to take advantage of the fast light component, it is necessary to use a photosensor that is sensitive to 200 nm light. According to Knoll (2000), the fast component went unobserved for lack of use of an appropriate photosensor.

YAP and YAG find application in particle counting, especially in electron microscopy. Their low atomic number makes them of limited utility for gamma-ray spectroscopy. YAP has excellent proportionality, which means resolution is preserved even if gamma rays undergo multiple scatterings prior to photoelectric absorption. Unfortunately, the low Z necessitates the use of a larger crystal, and self-absorption of the light decreases net light yield and worsens resolution.

YAP also is a fast crystal (27 ns decay time), and excellent timing resolution can be obtained. YAG is unusual in that the Ce emission is shifted to 550 nm, making it less than optimal for use with photomultipliers; it is a much better match to solid-state readouts. Moszynski et al. (1997) reports that the fraction of light in the fast and slow components changes with the type of incident radiation, making it possible to devise pulse-shape discrimination schemes to distinguish between $\beta-\gamma$ and ions.

Lathanum halides have been reported in the literature since 1999. It is only available from Saint-Gobain Crystals. The crystal contains lanthanum, which is naturally radioactive, and in low-background counting applications this will limit the size of a single crystal. If arrays of crystals are needed, the designer needs to be aware that large amounts of lanthanum will have the effect of raising the local background levels. Kernan (2004) concludes from his measurements of $LaCl_3$:Ce that the rate of 1,435 keV emissions in a 7.62 cm diameter by 7.62 cm thick crystal will be 500/s. The crystal emits light between 350 and 440 nm, which is within the specifications of most photomultipliers. However, if it is desired to use solid-state readouts, it is necessary to obtain blue-enhanced devices. The crystal is not cubic and is prone to cracking if not handled carefully or subjected to rapid temperature changes.

The good energy resolution, approaching 3%, makes these crystals a good choice for high-resolution gamma-ray spectroscopy, and handheld radioactive-isotope identifiers using $LaBr_3$:Ce are on the market today. The crystals are being considered for space applications and in the oil well logging industry. The latter use is enabled by the small variation of light yield with increasing temperature, dropping by only 10% between 27 °C and 175 °C. Lanthanum halides are also being considered for medical applications in PET and SPECT imagers.

$PbWO_4$ is used almost exclusively by the high energy physics community in particle detectors in calorimeters. It is the material of choice because of its speed, high density, and small radiation length. Melcher (2005) reports that production of $PbWO_4$ was projected to peak in 2005 and was similar to worldwide demand for scintillator crystal for PET, SPECT, and X-ray CT in that year. The light yield is sufficiently small; however, it is not useful for gamma-ray spectroscopy.

ZnS was not included in the previous tables because it is not a scintillator used for gamma-ray spectroscopy. It is available only as ZnS:Ag or ZnS:Cu powder and is mixed with a binder to form thin sheets. The powder is not transparent to its own light, forcing the sheets to be at most approximately 0.5–1 mm thick with an areal density of 25 mg/cm^2. The decay time constant varies with the incident radiation, being 200 ns for heavy particles, but about 50 ns for electrons. Birks (1964) gives a discussion of the early work on this scintillator, which reported a non-exponential decay and decay times for electrons as small as 10 ns. ZnS scintillators have been used exclusively as charged-particle detectors in consumer goods (it was the phosphor in the paint in the infamous radium watch dials of the early twentieth century), and detectors for scientific purposes (Hornyak buttons for fast neutrons, mixed with LiF, ^{235}U, or B_2O_3 for thermal-neutron detectors). When mixing ZnS with LiF or B_2O_3, it is necessary to optimize the size of the particles to maximize the escape of light.

3.2 Organic Scintillators

The data in ❯ *Table 3* was obtained from the web sites of Eljen Technology, Saint-Gobain Crystals, and Scintitech, Inc. In the case of unloaded plastics, the data covers many varieties of scintillator; they are aggregated here because the base plastic is polyvinyltoluene and the phosphors themselves only comprise a small fraction of the whole.

Table 3
Properties of common organic scintillators

Material	ρ(g/cm^3)	Emission maximum (nm)	Decay time constant	Index of refraction[a]	Light yield (ph/MeV)	Air sensitive?
Plastics (unloaded)	1.03	375–600	1–300 ns	1.58	6,400–10,000	Slightly (O_2)
Plastics (B)	1.02–1.03	425	2.2 ns	1.58	7,500–9,200	Slightly (O_2)
Plastics (Pb)	1.08	425	2.1 ns	1.58	5,200	Slightly (O_2)
Liquids (HC)	0.88–1.0	425	1–300 ns	1.5	12,000	Yes
Liquids (FC)	1.6	425	~3 ns +	1.4	3,000	Yes
Liquids (DC)	0.95	425	3.5 ns +	1.5	9,200	Yes
Liquids (B)	0.92	425	3.7–300 ns	1.42	10,000	Yes
Liquids (MO)	0.85–0.87	425	2 ns +	1.47–1.49	5,000–10,000	Slightly
Liquids (Dioxane)	1.04	425	3.8 ns +	1.44	10,000	Yes
Liquids (Gd)	0.89	424	3.6 ns +	1.50	10,600	Yes
Liquids (Sn)	0.95	425	3.8 ns	1.50	5,300	Yes
Stilbene	1.22	390	3.5 ns	1.64	14,000	Yes
Anthracene	1.25	445	3.0 ns	1.62	20,000	Yes
p-Terphenyl	1.23	420	3.7 ns	1.65	27,000	Yes

[a]Index of refraction at wavelength of maximum emission
B boron loaded, *Pb* 5% lead loaded, *Sn* 10% tin loaded, *Gd* 0.25–0.5% Gd loaded, *HC* hydrocarbon (toluene, xylene, benzene), *FC* fluorocarbon liquid, *DC* deuterocarbon liquid, *MO* mineral oil, + there are additional time constants not specified by the vendor

The data in *Table 4* was obtained from the literature, vendors' web sites, and calculation by the author. The reader should take pricing information below as an approximate guide and must contact a vendor with specific requirements to obtain an accurate quote.

Organic scintillators enjoy wide use in homeland security, X-ray detectors, charged-particle detectors, heavy-ion detectors, fast- and slow-neutron detection, and electron detectors. They are typically not used for gamma spectroscopy because the primary interaction at low energy is Compton scattering; no photopeak is generated. The intrinsic resolution of organic scintillators is fairly poor, being only approximately 20% at 662 keV.

Plastics are easy to form by solvent casting for thin films and thermal casting for larger pieces, easy to machine, are not hygroscopic, and can be handled while wearing gloves. They are not affected by water, but slightly affected by oxygen. The latter is not usually a problem unless the scintillator is in a pure oxygen atmosphere for extended periods of time. The base plastic is polyvinyltoluene (PVT) or, for increased temperature stability, polystyrene (PS), and the density is always close to 1 g/cm^3.

The light yield is a strong function of particle mass. Birks (1964) shows that the light yield from electrons, protons, and alpha particles in anthracene at energies between 10 keV and 10 MeV is in the approximate ratio of 1 : 0.2 : 0.06. At energies from 20 to 160 MeV in plastic scintillator, the ratio is approximately 1 : 0.8 : 0.4. Consequently, for the same energy, gamma rays produce more light than does a proton (or neutron interacting via (n, p) scattering) or alpha particle. *Figure 10* shows the conversion as calculated by Cecil et al. (1979) for NE-213 (identical to the modern EJ-301 and BC-501A) over the energy range 0–3,000 keV. The curve is

Table 4

Additional properties of organic scintillators

Material	Neutron sensitivity	Radiation hardness (Gy)	Radiation length (cm)	Price (US$/cm^3)[c]
Plastics	F[a]	3×10^4	43.2	$0.11–0.32
Plastics (B)	F, S[b]		43.4	$11–12
Plastics (5% Pb)	F		33.2	$3
Liquids (HC)	F		50.6	$0.25
Liquids (FC)	No		21.7	
Liquids (DC)	No		49.7	
Liquids (B)	S		48.9	$5 (^{10}B) $2 (^{nat}B)
Liquids (MO)	F, S		52.5	$0.07
Liquids (Dioxane)	F		45.4	
Liquids (Gd)	S		48.1	$0.15
Liquids (Sn)	F		33.3	
Stilbene	F	4×10^4	36.0	
Anthracene	F	2×10^4	35.0	
p-Terphenyl	F	4.5×10^4	35.6	

[a]F, sensitive primarily to fast neutrons

[b]S, sensitive to slow neutrons through capture reaction

[c]Pricing data is courtesy of Dr. Charles Hurlbut of Eljen Technolgy, Inc. Solid plastic scintillator is priced for cast sheet up to 50 mm thick (thicker sheet is less expensive). Loaded plastic is priced for 50 mm diameter by 50 mm long cylinders. Liquid scintillator is priced for quantities of 3–10 liters and reflects only the cost of the scintillator

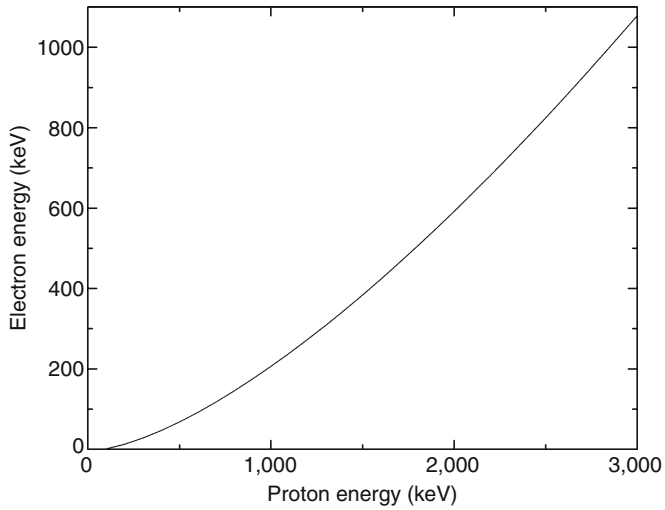


Fig. 10

Proton-to-electron conversion for EJ-301 (NE-213/BC-501A)

in general agreement with the results shown by Birks (1964) and is not linear at proton energies below approximately 2,000 keV. This conversion must be taken into account when a lower threshold and range of analog and digital electronics is to be established for a detector using organic scintillator. If care is not taken, the system may work perfectly for protons but saturate for gamma rays of interest.

Some organic scintillators (most of them liquids) also produce different shapes of light pulse, depending on the type of incident particle. Heavy charged particles tend to shift the light to longer decay times and this is exploited in pulse-shape discrimination schemes. Chapter 17 of Knoll (2000) discusses pulse-shape discrimination between neutrons and gamma rays.

Vendors specify that the temperature variation of the light yield of their scintillators does not vary from −60 °C to +20 °C, but then falls by 5% from +20 °C to +60 °C. Operation above 60 °C typically is not recommended because plastics soften and liquids are flammable. Some liquids are made with higher-flashpoint solvents (xylene or mineral oil in place of toluene), but the flashpoint of these scintillators does not exceed ~100 °C. Some plastics are made with a cross-linked polymer and do not soften until ~100 °C.

The chemistry of hydrocarbons is sufficiently rich to permit the inclusion of boron, tin, lead, and gadolinium compounds in solid and liquid scintillators. Addition of lithium is difficult (with the exception, perhaps, of lithium salicylate): lithiated hydrocarbons are pyrophoric. Boron and gadolinium provide sensitivity to thermal neutrons, while tin and lead enhance gamma sensitivity. However, the addition of heavy metals rapidly reduces the light yield, which limits the weight fraction to 5–10%. This amount of tin or lead does not significantly increase stopping power for gammas between ~100 keV and the critical energy. As can be seen in ❯ *Table 4*, the radiation length is reduced by 25% by the addition of the few percent of heavy metal.

^{10}B-loaded scintillator is generally used in small detectors because the thermal-neutron-capture cross section is 3837 b, the reaction products are ions, and the energy deposited in the scintillator is only 2.35 MeV. Consequently, a high concentration of boron is desired and only a small volume is required. When Gd captures a thermal neutron, however, the reaction products are prompt gamma rays, atomic X-rays, and conversion electrons. In a small Gd-based detector, it is difficult to distinguish neutron reaction products from photoelectrons and Compton electrons.

A large Gd-based or B-based detector can be made sensitive to fast neutrons by using sufficient loading to make the capture by Gd or B far more probable than the capture by hydrogen. In such a detector (1–1,000 liters, or larger), fast neutrons are thermalized by multiple (n, p) scatterings prior to capture by Gd or B. Each scattering produces a light pulse, and at the end of the sequence of scatterings, a "flash" from the capture occurs. When Gd is the sensitizer, 7.9 MeV (^{157}Gd) or 8.5 MeV (^{155}Gd) is released as prompt gamma rays, atomic X-rays, and conversion electrons, and if the detector's volume is sufficiently large, all the energy is captured, resulting in an identifiable pulse, generally much larger than any gamma-ray pulses. Additional discrimination may be achieved by demanding the presence of pulses from (n, p) scattering in some number of microseconds preceding the 8 MeV flash. The same principle can be applied to a boron-based system, with the exception that the flash is much smaller (equivalent to only ~100 keV electrons) because the reaction products are heavy ions. The interested reader is referred to the section in Knoll (2000) on capture-gated neutron spectrometers.

4 Conclusions

Scintillators have been used to detect radiation since the discovery of X-rays and natural radioactivity. Since that time, the field has evolved from the trial-and-error approach to the discovery of new scintillators to computational methods to understand and predict the properties of crystals and compounds. Scintillators have moved from parlor curiosities (much to their detriment, radium watch dial painters daubed themselves with radium-laced ZnS paint to amuse friends in the dark) to indispensible components in medical equipment, detectors for homeland security applications, health physics instruments, and large- and small-scale scientific experiments. They have moved from the laboratory to being a consumer product with dozens routinely commercially available, and dozens more under active investigation in research projects around the world.

5 Cross-References

- Chapter 1, "Interactions of Particles and Radiation with Matter"
- Chapter 13, "Photon Detectors"
- Chapter 14, "Neutrino Detectors"
- Chapter 16, "Semiconductor Counters"
- Chapter 17, "Gamma-Ray Detectors"
- Chapter 23, "Astrophysics and Space Instrumentation"
- Chapter 24, "Indirect Detection of Cosmic Rays"
- Chapter 25, "Technology for Border Security"
- Chapter 26, "Accelerator Mass Spectrometry and its Applications in Archaeology, Geology, and Environmental Research"
- Chapter 29, "Particle Detectors in Materials Science"
- Chapter 34, "Radiation Detectors and Art"

References

Amsler C et al (2008) Particle data group review of particle physics. Phys Lett B667: 1; Amsler C et al (2009) partial update for the 2010 edition, see Chapters 27 and 28

Birks JB (1964) The theory and practice of scintillation counting. Pergamon, London

Cecil RA, Anderson BD, Madey R (1979) Improved predictions of neutron detection efficiency for hydrocarbon scintillators from 1 MeV to about 300 MeV. Nucl Instrum Meth 161: 439

Cherepy NJ, Payne SA, Asztalos SJ, Hull G, Kuntz JD, Niedermayr T, Pimputkar S, Roberts JJ, Sanner RD, Tillotson TM, van Loef E, Wilson CM, Shah KS, Roy UN, Hawrami R, Burger A, Boatner LA, Choong W-S, Moses WW (2009) Scintillators with potential to supersede lanthanum bromide. IEEE Trans Nucl Sci 56:873

Derenzo SE, Moses WW, Cahoon JL, DeVol TA, Boatner L (1991) X-ray fluorescence measurements of 412 inorganic compounds. In: IEEE Nuclear Science Symposium Conference Record 91CH3100-5, vol 1. Sante Fe, pp. 143–147. This paper promised a more comprehensive listing in a future publication, but that manuscript was never submitted to a journal

Dorenbos P, De Haas JTM, Van Eijk CWE (1995) Non-Proportionality in the Scintillation response and energy resolution obtainable with Scintillation Crystals. IEEE Trans Nucl Sci 42:2190

Hofstadter R (1948) Alkali halide Scintillation Counters. Phys Rev 74:100

Hull G, Choong W-S, Moses WW, Bizarri G, Valentine JD, Payne SA, Cherepy NJ, Reutter BW (2009) IEEE Trans Nucl Sci 56:331

Kapusta M, Balcerzyk M, Moszynski M, Pawelke J (1999) A high energy resolution observed from a YAP:Ce scintillator. Nucl Instrum Meth Phys Res A 421:610

Kernan WJ (2004) IEEE nuclear science symposium conference record, Ergife Palace Hotel, Rome, Italy, 16–24 Oct 2004, paper N19-6

Knoll GF (2000) Radiation detection and measurement, 3rd edn. Wiley, New York

Melcher CL (2005) Perspectives on the future development of new scintillators. Nucl Instrum Meth Phys Res A 537:6

Moses WW, Weber MJ, Derenzo SE, Perry D, Berdahl P, Schwarz L, Sasum U, Boatner LA (1997) Recent results in a search for inorganic scintillators for x- and gamma ray detection. Presented at SCINT97 – the international conference on inorganic scintillators and their applications, Chinese Academy of Sciences, Shangai Branch 22–25 Sep 1997

Moszyinski M, Kapusta M, Mayhugh M, Wolski D, Flyckt SO (1997) Absolute light output of scintillators. IEEE Trans Nucl Sci 44:1052

Rodnyi PA (1997) Physical processes in inorganic scintillators. CRC, Boca Raton

Rooney BD, Valentine JD (1996) Benchmarking the compton coincidence technique for measuring electron response non-proportionality in inorganic scintillator. IEEE Trans Nucl Sci 43(3):1271–1276

Shinonoya S, Yen WM (1999) Phosphor handbook. CRC, Boca Raton

van Eijk CWE (2001) Inorganic-scintillator development. Nucl Instrum Meth Phys Res A 460:1

Wilkinson F III (2004) Scintillators. In: Wernick MN, Aarsvold JN (eds) Emission tomography: the fundamentals of PET and SPECT. Elsevier Science, San Diego, pp 229–254

Further Reading

Baccaro S, Cecilia A, Mihokova E, Nikl M, Nejezchleb K, Blazek K (2005) Influence of Si – Co doping on YAG:Ce scintillation characteristics. IEEE Trans Nucl Sci 52:1105

Balcerzyk M, Moszynski M, Kapusta M, Wolski D, Pawelke J, Melcher CL (2000) YSO, LSO, GSO and LGSO. A study of energy resolution and nonproportionality. IEEE Trans Nucl Sci 47:1319

Caffrey AJ, Heath RL, Ritter PD, Anderson DF, VanSiclen CDeW, Majewski S (1986) Radiation damage studies on BaF2 and BGO scintillator materials. IEEE Trans Nucl Sci 33:230

Chen J, Zhang L, Zhu R-Y (2005) Large size LYSO crystals for future high energy physics experiments. IEEE Trans Nucl Sci 52:3133

Curran SC, Baker WR (1948) Photoelectric alpha-particle detector. Rev Sci Instrum 19:116, The work was reported in 1944 during World War II, but withheld from publication until after the end of the war

Drozdowski W, Dorenbos P, Bos AJJ, Kraft S, Buis EJ, Maddox E, Owens A, Quarati FGA, Dathy C, Ouspenski V (2007) Gamma-ray induced radiation damage in LaBr3:5% Ce and LaCl3:10% Ce scintillators. IEEE Trans Nucl Sci 54:1387

Kraus H, Danevich FA, Henry S, Kobychev VV, Mikhailik VB, Mokina VM, Nagorny SS, Polischuk OG, Tretyak VI (2009) ZnWO4 scintillators for cryogenic dark matter experiments. Nucl Instrum Meth Phys Res A 600:594

Lecomte R, Pepin C, Rouleau D, Saoudi A, Andreaco MS, Casey M, Nutt R, Dautet H, Webb PP (1998) Investigation of GSO, LSO and YSO scintillators using reverse avalanche photodiodes. IEEE Trans Nucl Sci 45:478

Mesquita CH, Fernandes Neto JM, Duarte CL, Rela PR, Hamada MM (2002) Radiation damage in scintillator detector chemical compounds: a new approach using PPO-toluene liquid scintillator as a model. IEEE Trans Nucl Sci 49:1669

Nagornaya L, Apanasenko A, Burachas S, Ryzhikov V, Tupitsyna I, Grinyov B (2002) Influence of doping on radiation stability of scintillators based on lead tungstate and cadmium tungstate single crystals. IEEE Trans Nucl Sci 49:297

Nagornaya LL et al (2009) Large volume ZnWO4 crystal scintillators with excellent energy resolutionand low background. IEEE Trans Nucl Sci 56:994

Pausch G, Stein J (2008) Application of 6LiI(Eu) scintillators with photodiode readout for neutron counting inmixed gamma – neutron fields. IEEE Trans Nucl Sci 55:1413

Pepin CM, Berard P, Perrot A-L, Pepin C, Houde D, Lecomte R, Melcher CL, Dautet H (2004) Properties of LYSO and recent LSO scintillators for Phoswich PET detectors. IEEE Trans Nucl Sci 51:789

Pidol L, Kahn-Harari A, Viana B, Virey E, Ferrand B, Dorenbos P, de Haas JTM, van Eijk CWE (2004)

High efficiency of lutetium silicate scintillators, Ce – doped LPS, and LYSO crystals. IEEE Trans Nucl Sci 51:1084

Takayuki Yanagida (2005) Evaluation of properties of YAG(Ce) ceramic scintillators. IEEE Trans Nucl Sci 52:1836

Zhu R-Y (1998) Radiation damage in scintillating crystals. Nucl Instrum Meth Phys Res A 413:297

Zorn C et al (1994) Low dose rate evaluations of long plastic scintillator plates and bicron G2-doped wavelength shifting fibers. IEEE Trans Nucl Sci 41:746

http://scintillator.lbl.gov/. This site contains a summary list of many scintillators, some experimental, some well-established.

www.appscintech.com, web site of Applied Scintillation Technologies, a vendor of Li-glass scintillator.b

http://www.brightcrystals.com/english, web site of Bright Crystals Technology, Inc., a supplier of CsI:Tl.

www.bicron.com, web site of Saint-Gobain Crystals' scintillator division, a vendor of scintillation crystals, organic scintillator, and Li-glass scintillator.

www.eljentechnology.com, web site of Eljen Technology, a vendor of organic scintillators.

www.hamamatsu.com, web site of Hamamatsu Photonics, K.K. This site has specifications for photomultiplier tubes manufactured by the company.

www.hilger-crystals.co.uk, web site of Hilger Crystals, a vendor of scintillation crystals.

www.nist.gov/physlab/data/xcom/index.cfm, web site of the U.S. National Institute of Standards and Technology photon cross sections database.

16 Semiconductor Counters

Douglas S. McGregor
Kansas State University, Manhattan, KS, USA

1 Nomenclature 379

2 Introduction 379

3 Definitions 380
3.1 Energy Band Gap 380
3.2 Charge-Carrier Concentration 382
3.3 Dopant Impurities 383
3.4 Carrier Mobility 384
3.5 Carrier Lifetime 385
3.6 Material Resistivity 385

4 Basic Detector Configurations 385
4.1 *pn* Junction 386
4.2 *pin*-Junction Devices 389
4.3 Schottky Devices 390
4.4 Ohmic Contacts 391
4.5 Resistive Devices 391
4.6 Photoconductive Devices 392
4.7 Operation 392

5 γ-Ray and X-Ray Spectrometers 396
5.1 X-Ray Detectors Based on Si 397
5.1.1 Basic Design 398
5.2 Detectors Based on Ge 399
5.2.1 Various Designs 400
5.3 Compound Semiconductor Detectors 402
5.3.1 CdTe 404
5.3.2 CdZnTe 404
5.3.3 HgI_2 404

C. Grupen, I. Buvat (eds.), *Handbook of Particle Detection and Imaging*,
DOI 10.1007/978-3-642-13271-1_16, © Springer-Verlag Berlin Heidelberg 2012

6 Charged-Particle Detectors 405
6.1 Surface-Barrier and Implanted-Junction Detectors 405

7 Neutron Detectors 407

8 Conclusions 408

9 Cross-References 409

References 409

More References for the Interested Reader 410

Semiconductor Radiation Detector Suppliers 410

Abstract: The basic principles for the design and operation of semiconductor detectors are presented. A summary treatment of *pn*-junction and Schottky-junction formation is described. Common semiconductor configurations are discussed, including planar and coaxial detectors for γ-ray spectroscopy, and various detectors for α-particle spectroscopy.

1 Nomenclature

A	detector contact area	n	electron concentration
C_{det}	detector capacitance	p	hole concentration
C_{tot}	total coupling capacitance	q	electron charge
E	energy of radiation quantum	Q	charge collected from detector
E_A	acceptor energy	Q_0	charge excited in detector
E_C	conduction band edge	ρ_c	space-charge density
E_D	donor energy	ρ_s	semiconductor resistivity
E_F	Fermi energy	τ_e	electron lifetime
E_V	valence band edge	τ_h	hole lifetime
ϵ_s	semiconductor permittivity	ϕ_m	metal work function
ϵ_0	free space permittivity	ϕ_s	semiconductor work function
FWHM	full width at half maximum	ϕ_b	barrier potential
μ_e	electron mobility	V	applied detector voltage
μ_h	hole mobility	V_{bi}	built-in detector potential
μ_s	majority-carrier mobility	v_e	electron velocity
N_A	acceptor concentration	v_h	hole velocity
N_A^-	ionized acceptor concentration	V_{in}	detector input voltage signal
N_D	donor concentration	V_w	weighting potential
N_D^+	ionized donor concentration	W	detector active-region width
N_s	majority impurity concentration	w	average ionization energy
n_i	intrinsic concentration	χ_s	semiconductor electron affinity
ξ_e	electron extraction factor	ψ	potential
ξ_h	hole extraction factor	Z	atomic number

2 Introduction

Semiconductors are far more desirable for energy spectroscopy than gas-filled detectors or scintillation detectors because they are capable of much higher energy resolution. The observed improvement is largely due to the better statistics regarding the number of signal carriers (charges) excited by a radiation interaction. On average, it only takes 3–5 eV to produce an electron–hole pair in a semiconductor. By comparison, it takes between 25 and 40 eV to produce an electron–ion pair in a gas-filled detector and between 100 eV and 1 keV to produce a single photoelectron ejection from the photomultiplier tube (PMT) photocathode in a scintillation/PMT detector (primarily due to light reflections and poor quantum efficiency at the photocathode). Hence, a semiconductor produces more charge carriers from the primary ionization event and thus reduces the statistical fluctuation in the energy resolution.

Semiconductor radiation detectors have many different shapes, sizes, and configurations, yet there are some basic designs that, in some form or another, can be attributed to practically all semiconductor radiation detectors. These basic designs include *pn*-junction diodes, *pin*-junction diodes, Schottky diodes, resistive detectors, and photoconductors. The vast field of semiconductor radiation detectors is much too large to describe in a single book chapter. As a result, only those concepts needed to understand basic detector operations and characteristics are offered here. A selected list of literature is included at the end of the chaper that offers much more detailed accounts of various semiconductor detector configurations, characterizations, and operations. Example performances of some commercial detectors are listed.

3 Definitions

In the following section are listed and described certain basic definitions and concepts used in discussions of semiconductor detectors. For a more detailed discussion regarding these concepts, the reader is directed to the reference literature at the chapter end.

3.1 Energy Band Gap

In free space, a single atom has quantized and discrete allowed energy states. The Pauli exclusion principle states that no two electrons can occupy the same four quantum numbers (n, l, m_l, m_s), where n is the principal number referring to energy, l is the angular-momentum quantum number, m_l is the magnetic quantum number, and m_s is the spin. However, a solid material, such as a semiconductor crystal, is a matrix of atoms arranged in a lattice such that the various potentials of each of the atoms affect the surrounding neighbors and those electrons associated with them. If two atoms are forced into close proximity, each initially with identical quantum numbers, then something must change such that the exclusion principle is not violated, which is satisfied by the appearance of degenerate energy states. In other words, the original energy levels split such that two states appear where there was only one before.

Consider the issue of a solid, where typical atomic densities range from 10^{21} to 10^{23} atoms cm^{-3}. The total number of energy states must also split to accomodate the electron density, which form bands of states in place of what were once individual states for a single atom (depicted in ❯ *Fig. 1*). These bands may overlap, they may be relatively close to each other in energy with a small energy gap between them, or may form with large energy gaps between the bands. Electrons in the bands behave almost as though they are in an energy continuum, but it is actually a *quasi-continuum*, in which there is a defined density of available states in each of the bands. The density of states is still predetermined by the original total number of states of the individual atoms.

Each of these energy bands has a certain density of allowed energy states that can be occupied by an electron. The electrical conductivity of a solid is determined by many parameters, which include charge-carrier mobility, the density of free charge-carriers available in a partially filled energy band, and the availability of unfilled energy states in the partially filled band. As conceptual examples, ❯ *Fig. 2* shows simplistic band diagrams for (a) insulators, (b) conductors, and (c) semiconductors. In ❯ *Fig. 2a*, the *valence band*, which is active in chemical binding of electrons in compounds, is filled, and the next available energy band is devoid of electrons.

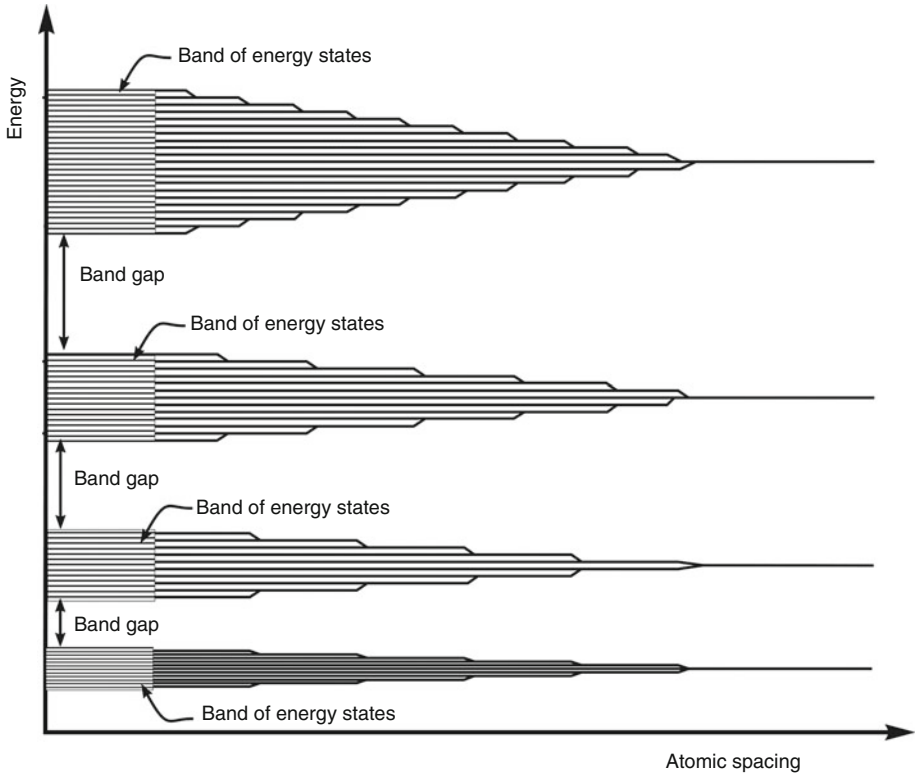


Fig. 1

As atoms are brought closer together, their allowed energy states split into degenerate states. In a solid medium, the high atomic density causes these degenerate states to form quasi-continua referred to as energy bands

In typical notation, the upper energy limit of the valence band is denoted E_V and the lower energy limit of the *conduction band* is denoted E_C. The energy difference between E_C and E_V is a forbidden energy region, referred to as the energy band gap and denoted E_g. If the band-gap energy is large such that electrons are not thermally excited from the valence band into the conduction band the material is considered an insulator. Conduction can only take place provided that there are empty states for charge carriers to occupy in lateral energy space. Since the valence band is completely full of charge carriers, there are no empty states, hence no conduction. Further, the conduction band has empty states, but no charge carriers, hence again conduction does not take place.

In *Fig. 2b*, there are two examples of conductors. In the first example, the valence band and the conduction band overlap such that electrons can easily move from the filled valence band into the partially filled conduction band without a change of energy. Hence, there are plenty of unfilled states with an overlapping reservoir of electrons (the valence band) that can move to the conduction band, thereby giving rise to free conduction. In the other example, the valence band is filled, and the conduction band is partially filled with a high density of electrons, even at low temperature. Again, the conditions exist for free conduction.

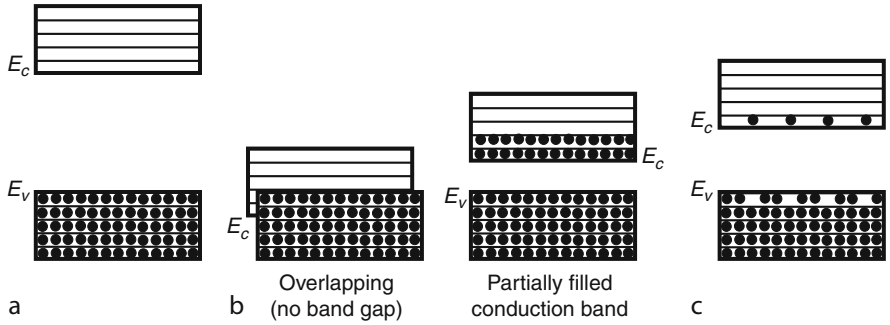


Fig. 2

Shown are depictions of simple band diagrams for (a) insulators, (b) conductors, and (c) semiconductors. In (b) there are two depictions for conductors, one in which a filled valence band overlaps the conduction band, and the other in which the valence band is full with a partially filled conduction band

In *Fig. 2c*, there is a band gap similar to the insulator example, except the band gap is relatively small. As a result, some electrons are thermally excited from the valence band into the conduction band where they can freely conduct. However, the density of the electrons in the conduction band is determined largely by the band-gap energy and the temperature. At low enough temperature, the electrons will all return to the valence band and the material will behave as an insulator. As the temperature is increased, more and more electrons will cross the band gap into the conduction band, and the material conductivity will continue to increase. Often this class of materials is separated into *semiconductors* and *semi-insulators*, roughly defined by the band-gap energy. Typically, band-gap energies ranging up to approximately 1.4 eV constitute a class of materials commonly designated as semiconductors, while band-gap energies ranging from 1.4 up to 5 eV are considered semi-insulators. Band gaps exceeding 5 eV form the insulator class of materials. However, these ranges are not rigidly classified, and often semi-insulators and semiconductors are treated as the same, which will be the case in this chapter.

3.2 Charge-Carrier Concentration

The probability distribution of electrons with energy E_e is determined by Fermi–Dirac statistics,

$$F(E) = \frac{1}{1 + \exp\left[\dfrac{E_e - E_F}{kT}\right]}, \tag{1}$$

where k is Boltzmann's constant, T is the absolute temperature, and E_F is the Fermi energy level (*Table 1*). The Fermi energy is the energy level, at 0 K temperature, where all states below it are filled and all above it are empty. Equation 1 can be used to determine the density of electrons in the conduction band, with knowledge that electrons will not be present in the band gap, hence Eq. 1 is valid for $E_e \geq E_C$ and $E_e \leq E_V$. From Eq. 1, the probability that an

Table 1
Some useful physical constants

Constant	Symbol	Magnitude
Avogadro's number	N_0	6.023×10^{23} molecules mol^{-1}
Boltzmann's constant	k	1.38×10^{-23} J $K^{-1} = 8.62 \times 10^{-5}$ eV K^{-1}
Electronic charge	q	1.6×10^{-19} C
Electron volt	eV	1.6×10^{-19} J
Free-electron mass	m	9.1×10^{-31} kg
Permittivity of free space	ϵ_0	8.854×10^{-14} F cm^{-1}
Permeability of free space	μ_0	1.257×10^{-8} H cm^{-1}
Planck's constant	h	6.625×10^{-34} J s
Velocity of light	c	3×10^{10} cm s^{-1}
Thermal energy at 300 K	kT	0.0259 eV

electron crosses the band gap to the conduction band will increase with increasing temperature, which should be intuitively obvious.

The concentration of electrons in the conduction band is denoted n and the concentration of empty states in the valence band is denoted p. These empty states are treated as positive charge carriers called "holes," which greatly simplifies calculations. For a pure material, the electrons in the conduction band arrive at the expense of leaving an equal density of holes in the valence band. Hence, $n = p$, which is referred to as the intrinsic case. Typically, the intrinsic concentration of both electrons and holes is denoted n_i. It can be shown that,

$$n = \int_{E_C}^{\infty} N(E)F(E)\,\mathrm{d}E \approx N_C \exp\left(-\frac{E_C - E_F}{kT}\right), \tag{2}$$

where $N(E)$ is the function describing the allowed density of states, and N_C is the effective density of allowed states in the conduction band. Similarly, describing the unfilled states or holes in the valence band,

$$p = \int_{-\infty}^{E_V} N(E)[1 - F(E)]\,\mathrm{d}E \approx N_V \exp\left(-\frac{E_F - E_V}{kT}\right), \tag{3}$$

where N_V is the effective density of allowed states in the valence band.

3.3 Dopant Impurities

Dopant impurities are often added to a semiconductor to control the electrical properties. Dopants that add excess electrons to the chemical binding are called donors, because they need only a slight amount of energy to liberate these excess electrons into the conduction band. Dopants that lack an electron to complete the valence bonding are called acceptors, because they need only a slight amount of energy to accept electrons from the valence band into their unfilled states. The simplified energy band structures for intrinsic, n-type and p-type materials are depicted in ❯ *Fig. 3*. The concentration of donor atoms is denoted N_D with energy level E_D, and the concentration of acceptor atoms is denoted N_A with energy level E_A. If donors are added to the semiconductor, then the concentration of holes is reduced. The opposite condition

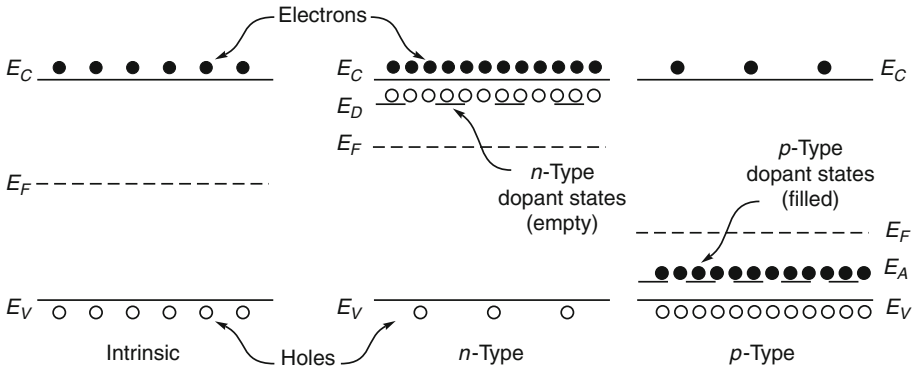


Fig. 3

The simplified energy band structures for intrinsic, *n*-type, and *p*-type materials. The electron and hole carrier concentrations are equal for the intrinsic case. Materials doped *n*-type have an excess of electrons in the conduction band, and materials doped *p*-type have an excess of holes in the valence band

is achieved if acceptors are added to the semiconductor. The general relationship between the free electron concentration and the free hole concentration is

$$np = n_i^2 . \tag{4}$$

At room temperature, almost all shallow donors and acceptors are ionized, hence $N_A \approx p$ and $N_D \approx n$. Combining Eqs. 2–4, it is easily shown that the intrinsic carrier concentration is

$$n_i = \sqrt{N_C N_V} \exp\left(-\frac{E_g}{2kT}\right). \tag{5}$$

Equation 5 clearly indicates that the intrinsic charge-carrier population n_i decreases with increasing band-gap energy E_g and decreasing temperature T.

3.4 Carrier Mobility

The motion of a charge carrier can be influenced by diffusion, magnetic fields, and electric fields. The strength of this influence is characterized by the carrier mobility. The valence and conduction bands have different periodic potentials, and for this reason electron mobility in the conduction band is different than hole mobility in the valence band. Electron mobility is denoted μ_e and hole mobility is denoted μ_h. The velocity of a charge carrier can be approximated with

$$v_{e,h} = \mu_{e,h} \mathcal{E} , \tag{6}$$

where $\mathcal{E}$ is the electric field magnitude. Equation 6 is a good approximation provided that the electric field is relatively lower than the saturation field (usually below 2×10^3 V cm^{-1}), above which the charge-carrier velocities begin to asymptotically approach a saturation limit.

3.5 Carrier Lifetime

Charge carriers in the conduction band are dynamically dropping back into the empty states of the valence band, while other electrons gain energy to cross the band gap. Overall, a somewhat constant density of electrons and holes remains in the conduction and valence bands, respectively. The time over which a charge carrier remains in either band is altered by impurity and defect states that appear in the band gap, which increase the probability of either carrier transferring from either band to an intermediate state, or eventually to completion of the recombination process. If a charge carrier falls into a defect state, it is referred to as "trapped," whereas if the charge-carrier journey is completed, where an electron falls completely back into a hole in the valence band, it is referred to as having "recombined." The average time period over which an excited electron remains in the conduction band before being trapped or recombining is the electron lifetime, denoted τ_e, and the average time period over which a hole remains in the valence band before being trapped or recombining is the hole lifetime, denoted τ_h.

3.6 Material Resistivity

The ability of a semiconductor to conduct electrons is referred to as the material resistivity, with units of Ω·cm. The resistivity of a semiconductor is found with

$$\rho = \frac{1}{q(\mu_e n + \mu_h p)}, \tag{7}$$

where q is the unit electronic charge, μ_e is the electron mobility, and μ_h is the hole mobility. In the case that $n \gg p$, Eq. 7 reduces to

$$\rho \approx \frac{1}{q\mu_e n}, \tag{8}$$

and in the case that $p \gg n$, Eq. 7 reduces to

$$\rho \approx \frac{1}{q\mu_h p}. \tag{9}$$

The resistance of a right parallelepiped block of semiconductor is described by

$$R = \rho \frac{W}{A}, \tag{10}$$

where W is the detector width or length, and A is the contact area.

4 Basic Detector Configurations

Semiconductor detectors can be fashioned into many different device configurations, including junction diodes, Schottky-barrier diodes, photoconductors, and photoresistors. To select the most appropriate device configuration, one must consider the semiconductor material and the radiation-detection application. Some semiconductor materials are composed of substances that are not easily converted into junction diodes, such as HgI_2 and PbI_2, whereas other semiconductors, such as Si and Ge, are easily fashioned into junction devices. Some materials, such as

GaAs, can be fabricated easily into either photoconductors or Schottky-barrier devices, whereas limited doping selection and other chemical constraints prevent some materials to be configured easily as either reverse-biased diodes or photoconductors. These fundamental designs are briefly described in the following sections.

4.1 *pn* Junction

If two blocks of semiconductor material, one doped with N_D donors and the other doped with N_A acceptors, are brought into contact, then the *p*-type side of the junction boundary has an excess of free-hole charge carriers and the *n*-type side of the junction has an excess of free-electron charge carriers. The concentration gradient across the junction boundary will cause holes to *diffuse* across the boundary into the *n*-type side, and electrons to *diffuse* over to the *p*-type side. The free carriers leave behind the immobile host ions, which produce regions of *space charge* of opposite polarity, as depicted in Fig. 4. The result is the production of an internal electric field with an applied force in the opposite direction of the diffusion force. The presence of space charge distorts the band potentials and causes the bands to bend across the junction boundary. The bands continue to distort and bend until the diffusion force is equal to the electric field force, thereby, producing a state of equilibrium.

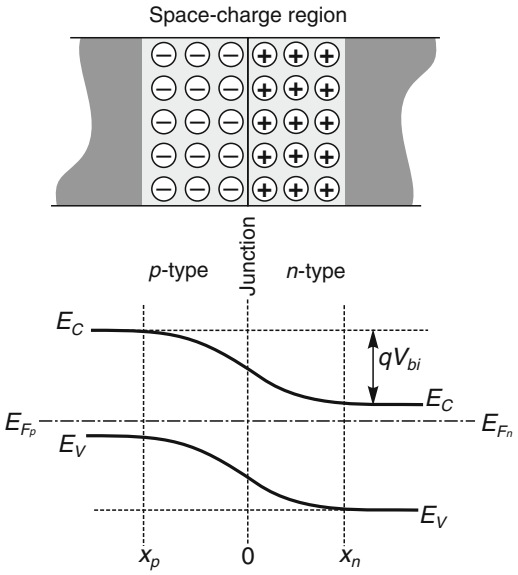


Fig. 4

In the depiction of a *pn* junction at equilibrium, the free charge carriers are swept from the space-charge or *depletion* region, leaving behind a polarized zone that produces an internal electric field. Diffusion of charges in one direction is balanced by electric field drift in the other. At equilibrium, the Fermi level is constant across the junction

Poisson's equation is used to determine the space-charge-region width,

$$\frac{\partial^2 \psi}{\partial x^2} = -\frac{\rho_c(x)}{\epsilon_s} = \frac{q}{\epsilon_s}\left(N_A^- - N_D^+ - p + n\right), \tag{11}$$

where ρ_c is the volumetric charge density, ϵ_s is the semiconductor permittivity, N_D is the dopant density on the n-type side, N_A the dopant density on the p-type side, and n and p are the electron and hole free-carrier densities. Under room-temperature conditions, the ionized acceptor concentration $N_A^- \approx N_A$ and the ionized donor concentration $N_D^+ \approx N_D$. With the assumption of an abrupt junction and uniform dopant distribution, the depletion width solution is

$$W = \left\{\frac{2\epsilon_s(V_{bi} - V)}{q}\left(\frac{N_A + N_D}{N_A N_D}\right)\right\}^{1/2}, \tag{12}$$

where V is an externally applied negative voltage and V_{bi} is the built-in potential arising from the energy-band bending, as shown in ❯ *Fig. 4*. The value of V_{bi} for common pn-junction diodes is approximately 0.7 V. In the case that one side of the junction is doped much higher than the other side, by at least an order of magnitude, ❯ Eq. 12 can be approximated by

$$W \approx \left\{\frac{2\epsilon_s(V_{bi} - V)}{qN_s}\right\}^{1/2}, \tag{13}$$

where N_s is the doping concentration of the *lighter* doped side. Note that qV_{bi} is equal to the conduction-band energy difference from the p-type side to the n-type side (see ❯ *Fig. 4*), where

$$V_{bi} \approx \frac{kT}{q}\ln\left(\frac{N_D N_A}{{n_i}^2}\right). \tag{14}$$

At room temperature, the material resistivity can be expressed as

$$\rho_s = \frac{1}{q\mu_s N_s}, \tag{15}$$

where N_s and μ_s are the background dopant concentration and mobility for the lighter doped side of the junction, and ❯ Eq. 13 can be rewritten as

$$W \approx \{2\epsilon_s \mu_s \rho_s (V_{bi} - V)\}^{1/2}. \tag{16}$$

The electric field magnitude across the device is

$$|\mathcal{E}(x)| \approx \frac{qN_s}{\epsilon_s}(W - x), \quad 0 \le x \le W. \tag{17}$$

The active radiation-sensitive volume of the pn-junction detector is defined by the space-charge region (or depletion region), and the undepleted regions act as series resistances. Detectors based on pn-junction diodes are typically operated under *reverse bias*, which is regarded as a negative voltage, but is actually a positive voltage applied to the n-type material with respect to the p-type material. Shown in ❯ *Fig. 5* is the band diagram of a reverse-biased pn junction. The large energy-band barriers prevent electrons on the n-type side from diffusing over to the p-type side and prevent holes on the p-type side from diffusing into the n-type side. As a result, the junction suppresses leakage current that would be present if the semiconductor were operated as a resistor. This necessary condition reduces the electron current such that minute charge packets excited in the depletion region by radiation interactions can be measured.

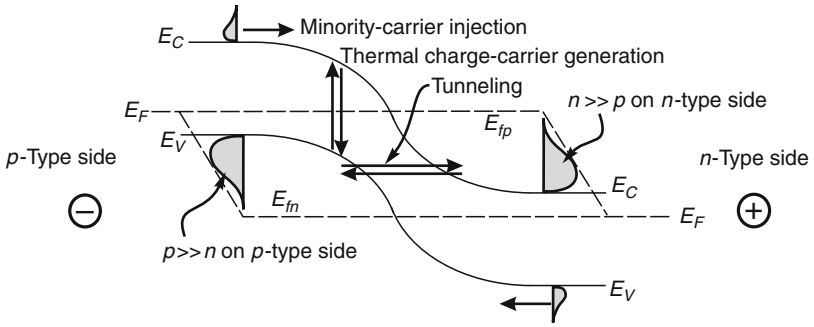


Fig. 5

In reverse bias, the density of available carriers, dominated by the *minority*-carrier concentration, determines the leakage current. At higher voltages, bulk generation and tunneling may increase the observed leakage currents

There are some sources of leakage current, as depicted in Fig. 5. Although the *majority* carriers on the *p*-type side are holes, according to Eq. 4 a small concentration of electrons is still present. These *minority* charge carriers (electrons) can diffuse into the depletion region, where they are swept across by the electric field and contribute to the leakage current. A similar case is true for holes diffusing from the *n*-type side into the depletion region. Leakage current can also occur from thermal generation of electrons directly across the band gap into the conduction band, producing electron and hole free carriers. Thermal generation of such charge carriers can be suppressed by cooling the detector while it is operating. Under high-voltage-bias conditions, charge carriers can also tunnel directly across the band gap, again contributing to the leakage current. Variations about the leakage current can be a significant source of electronic noise (shot noise), which decreases the overall energy resolution of the detector. Basically, *pn* junctions are employed to minimize leakage currents in semiconductor detectors.

The capacitance of a parallel-contact *pn*-junction detector is given by

$$C_{\text{det}} = \frac{\epsilon_s A}{W}, \tag{18}$$

where A is the device active contact area. Substituting Eq. 16 into Eq. 18,

$$C_{\text{det}} \approx A\left[\frac{\epsilon_s}{2\mu_s\rho_s(V_{bi} - V)}\right]^{1/2}. \tag{19}$$

Detector capacitance affects the input-voltage pulse height, with a large capacitance diminishing the input voltage from the detector that is measured by the amplification circuit, as

$$V_{\text{in}} \approx \frac{Q}{C_{\text{tot}}}, \tag{20}$$

where C_{tot} is the total capacitance (detector and coupling) and Q is the total charge collected from the radiation detector after a radiation event. Hence, it is important to reduce detector capacitance and coupling capacitance between the detector and the shaping electronics. From Eq. 19, it is seen that increasing the reverse-bias voltage decreases the detector capacitance, but at the expense of increasing the leakage current.

4.2 *pin*-Junction Devices

Reverse-biased diode detectors need a sizeable depletion region to maximize efficient radiation absorption. As found in Eq. 13, the depletion-region width is proportional to the square root of the applied reverse voltage, indicating that excessive voltage would be required to produce a sizeable depletion region for a common *pn*-junction diode. The usual remedy is to construct a *pin* diode, which has an intrinsic, or high-purity region, between the *p*-type and *n*-type contacts. The *p*- and *n*-type contacts can produce either blocking barriers or electrically *ohmic* contacts to the material. The application of *p*-type contacts to *p*-type material, or *n*-type contacts to *n*-type material, generally produces non-rectifying electrical contacts that follow Ohm's law.

The device may have a truly intrinsic region between the *p*- and *n*-type contacts, in which the electron and hole populations are identical, or the high-purity region may be a lightly doped material. Lightly doped *p*-type material is commonly denoted π-type, and lightly doped *n*-type material is commonly denoted ν-type. Analysis of the diode construction should take into account the "punch-through" voltage, in which the depletion region extends completely across the high-purity region, whether the material is *i*-, ν-, or π-type. Semiconductor *pin*-diode radiation detectors are commonly operated at biases above the punch-through voltage.

Intrinsic material either has dopant concentrations below the intrinsic concentration or has compensation dopants that cause the residual free-carrier concentration to be below the intrinsic concentration. In either case, the residual space charge is practically zero ($\rho_c \approx 0$), and from Eq. 11 the resulting electric field is constant across the *pin* diode under reverse bias. Many detectors are fashioned from high-purity semiconductor materials, yet these actually are not intrinsically behaving, having some residual dopant concentration that is still above the intrinsic concentration. As a result, ρ_c is nonzero, although small, and the depletion-region width is determined by V_{bi} and the applied voltage. If the lightly doped π or ν regions are relatively thin, the detector might be "fully depleted" without an applied voltage, as depicted in *Fig. 6*.

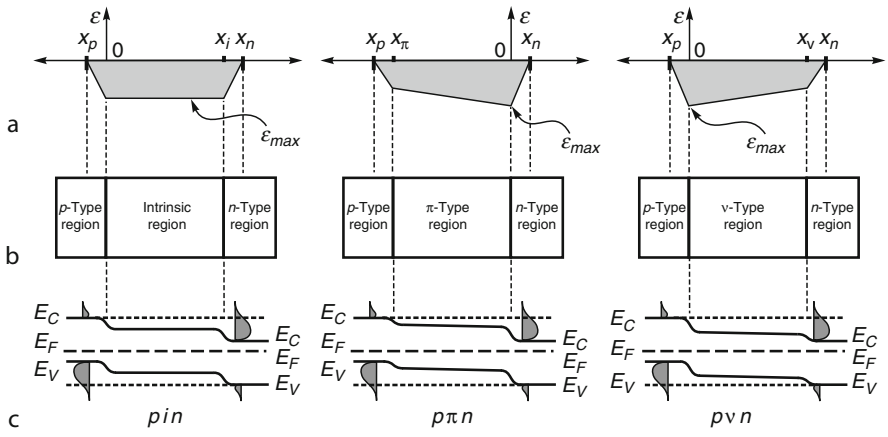


Fig. 6

Electric field (a), block (b), and band diagrams (c) of *pin*, *pπn*, and *pνn* devices. Note that the electric field across the intrinsic region of the *pin* diode is constant, but the lightly doped *pπn* and *pνn* have definite slopes to their electric fields

Usually, such is not the case and a reverse bias must be applied to extend the depletion region and electric field across the device as determined from Eqs. 13 and 17, respectively.

4.3 Schottky Devices

The application of a metal to a semiconductor surface acts to bend the semiconductor bands to form an energy barrier $q\phi_{bn}$, much like the pn-junction diode, as shown in *Fig. 7*. The Fermi energies, defined by the metal and semiconductor work functions (ϕ_m and ϕ_s) must align when the two materials are brought into contact. Because the semiconductor work function changes with doping concentration, reference is usually made to the semiconductor electron affinity ($q\chi_s$), the difference between vacuum level of full ionization and the conduction band edge. As a result of the junction formation, a built-in potential $q\phi_{bi}$ forms and potential barrier $q\phi_{bn}$ forms. Under a reverse bias, this energy barrier serves to reduce leakage current. Typically, the barrier height is lower than that for pn- or pin-junction diodes, hence detectors based on the Schottky-barrier diode typically have higher leakage currents than pn or pin diodes. Because the surfaces of semiconductors have defects, interface states and possible contaminants, the actual barrier height not only depends upon the choice of metal but also on how the surface is prepared. These surface states can effectly "pin" the detector barrier height, thereby, pre-determining the actual value of $q\phi_{bn}$ before the metal is applied. The depletion width for a Schottky-barrier detector is similar to a one-sided pn-junction diode,

$$W \approx \left\{ \frac{2\epsilon_s (V_{bi} - V)}{qN_s} \right\}^{1/2}, \tag{21}$$

where N_s is the dopant concentration in the semiconductor. Just as Eq. 13 can be rewritten as Eq. 16, Eq. 21 can be rewritten as

$$W \approx \{2\epsilon_s \mu_s \rho_s (V_{bi} - V)\}^{1/2}. \tag{22}$$

The value of V_{bi} for common Schottky diodes is 0.3 V. Note that Schottky contacts can be formed on n-type or p-type semiconductors, as depicted in *Fig. 8*. Besides being simple to

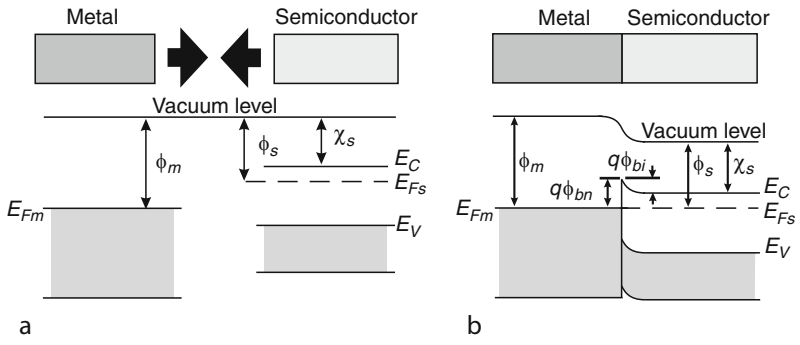


Fig. 7
The band configuration depicted for a metal and semiconductor (a) before and (b) after contact, according to the Schottky–Mott model

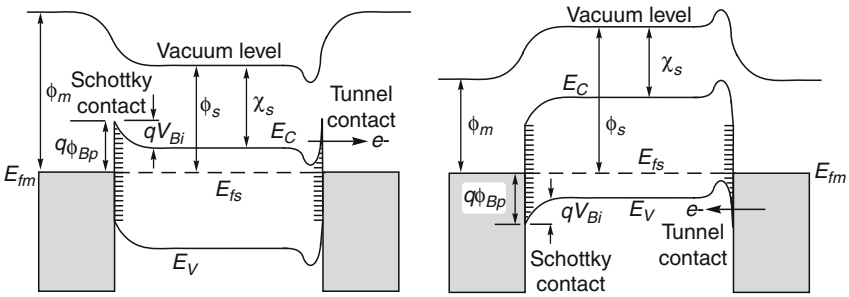


Fig. 8

Interface states alter the interface potentials and effectively "pin" the Schottky barrier. Shown are *n*-type and *p*-type Schottky barriers, along with *n*-type and *p*-type tunneling ohmic contacts

construct, Schottky-barrier detectors have a thin entrance region at the contact, unlike most *pn* and *pin* diodes, and energy attenuation in this "dead region" is kept to a minimum. As a result, Schottky-barrier detectors (sometimes called "surface-barrier detectors") are attractive as charged-particle spectrometers.

4.4 Ohmic Contacts

Efficient charge-carrier extraction from a semiconductor detector requires ohmic-behavior contacts, hence, it is important in some cases that the metal/semiconductor interface *not* form a rectifying potential barrier. For instance, metal contacts to the *n* and *p* regions of *pn* and *pin* detectors are processed to be non-rectifying, or *ohmic*, and generally follow Ohm's law. Unfortunately, due to interface state pinning, a barrier is formed for almost all metals applied to a semiconductor surface. To remedy this problem, a high concentration of dopants is applied with the metal and diffused, typically through thermal treatment, into the semiconductor. The process causes the Schottky barrier to become extremely thin so that electrons can tunnel directly through the barrier; hence the contact has ohmic behavior. Schottky diodes are typically constructed with one (or more) Schottky contact(s) as a rectifying barrier to reduce leakage current and one (or more) opposing ohmic contact(s) to allow for efficient carrier extraction, thereby, reducing electronic noise (see depiction in *Fig. 8*).

4.5 Resistive Devices

Semiconductor detectors fabricated from wide-band-gap materials (generally > 1.6 eV) have material resistivities high enough to reduce leakage currents to low levels, and as a result do not require rectifying contacts to suppress leakage currents. The devices typically have ohmic contacts for electrodes to prevent rectification and the subsequent formation of space-charge regions (which can limit the active-region volume). Radiation interactions in the detectors excite electron–hole pairs that are swept out of the detectors by an applied electric field.

The high resistivity of the device insures that the leakage current is lower than the radiation-excited current. For example, the current from a common 5 mm (width) × 10 mm × 10 mm CdZnTe detector of band gap 1.62 eV has a resistivity of 10^{11} Ω cm, giving a detector resistance of 5×10^{10} Ω. A bias of 600 V would produce a 12 nA leakage current. With average excitation energy w of 5 eV per electron–hole pair, a 662 keV γ ray will excite approximately 2×10^{-14} C. With a charge sweep-out time across the detector of 320 ns, the integrated background charge is 3.8×10^{-15} C, only 19% of the charge excited by the γ ray.

4.6 Photoconductive Devices

There is a unique class of semiconductor detectors known as photoconductors. In principle, such devices are composed of semiconducting material upon which ohmic contacts have been applied to prevent the formation of a blocking barrier or a space-charge region. From Eq. 7, the resistivity of a semiconductor material is inversely proportional to the free-carrier concentration. A single radiation absorption event will cause the local conductivity to change spontaneously, yet the small charge cloud is surrounded by higher-resistivity material on all sides. Further, the charge cloud is dissipated rapidly by an applied voltage.

Suppose instead the semiconductor block is saturated with a radiation pulse such that electron–hole pairs are evenly distributed throughout the crystal bulk. The conductivity of the entire semiconductor block changes because of the macroscopic change in the free-carrier concentration. This means that, for any constant applied voltage, the current through the device must increase. The current continues to flow, with well-fabricated ohmic contacts, since Ohm's law dictates that every electron exiting the device at the anode is replaced by another electron injected at the cathode. This *photocurrent* continues to flow, decaying away as a function of the charge-carrier lifetimes. Hence, the photocurrent is described as

$$I(t) = \left(\frac{VA}{W}\right) q \left[\mu_e (n + n_{ph} e^{-t/\tau_e}) + \mu_h (p + p_{ph} e^{-t/\tau_h})\right], \tag{23}$$

where n_{ph} and p_{ph} are the excess free-carrier concentrations excited by the radiation pulse, and n and p are the steady-state free-carrier concentrations just before the radiation pulse is produced.

This result is important. First, the current decays away as a function of the free-charge-carrier lifetimes, so that the current continues to flow even after the primary charge carriers excited in the semiconductor reach the electrodes. Second, the duration of the detector current pulse can be tailored by changing the free-charge-carrier lifetimes. High-speed photoconductive radiation detectors can be manufactured by purposely adding lifetime-shortening dopants, or by shortening the lifetimes with intentional radiation damage. These detectors have been used for fast timing measurements of radiation bursts.

4.7 Operation

Semiconductor detectors operate on the principle of induced charge, in which mobile charges drifting through the detector causes charge to flow in an externally connected circuit, typically a preamplifer for pulse-mode operation. This concept is important, mainly because the detector

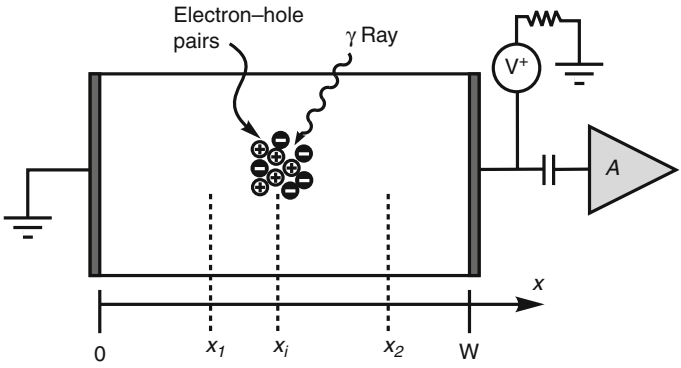


Fig. 9
Planar semiconductor detector configuration

produces voltage pulses that depend on the RC time constant of the output circuit, the capacitance of the detector, the coupling capacitance, and the charge-carrier velocities. Further, the voltage pulse induced by the detector begins to form *immediately* when the charges begin to move in the detector. From Fig. 9, it is seen that electron–hole pairs excited at point x_i in a detector with active region of width W will induce a current according to the scaled motion of electrons and holes. In general, the induced current on the ith electrode among a set of two or more electrodes is

$$i_i = \frac{\mathrm{d}Q_i}{\mathrm{d}t} = -q\nabla V_i(\mathbf{r}) \cdot \frac{\mathrm{d}\mathbf{r}}{\mathrm{d}t} = q\mathcal{E}_i(\mathbf{r}) \cdot \mathbf{v}\,, \tag{24}$$

where $\mathbf{v}$ is the charge-carrier velocity, and $\mathcal{E}_i$ is the electric field at point $\mathbf{r}$ under the condition that the "potential" at electrode i is normalized to unity (i.e., dimensionless), see e.g., He (2001), with all other electrodes grounded (set to zero). This normalized potential is often referred to as the *weighting* potential. The weighting potential for a simple planar detector is

$$V_w|_{\text{planar}} = \frac{x}{W}\,, \tag{25}$$

which is clearly a linear function of position in the detector. Hence, the current induced is directly proportional to the distance charge carriers travel across the detector width W within time t. Suppose electron–hole pairs are created at position x_i, and the holes are drifted to position x_1 while electrons are drifted to position x_2. The solution of Eq. 24 yields the total change in the induced charge from the motion of electrons and holes,

$$\Delta Q|_{\text{planar}} = Q_0 \left[\frac{x_i - x_1}{W}\bigg|_h + \frac{x_2 - x_i}{W}\bigg|_e \right] = Q_0 \left[\frac{x_2 - x_1}{W} \right]. \tag{26}$$

From Fig. 9 and Eq. 26, it becomes clear that if the holes and electrons reach their respective electrodes, where $x_2 - x_1 = W$, the change in the induced charge ΔQ is the same as Q_0, a case referred to as *complete charge collection*. Any condition in which the electrons and/or holes do not completely reach their respective electrodes results in incomplete charge collection, and $\Delta Q < Q_0$.

The induced current is not a linear function of position for detector configurations other than planar devices. Other basic detector configurations include coaxial designs and spherical

(hemispherical) designs (see Fig. 10). It can be shown that the weighting potential for a cylindrical detector is

$$V_w|_{\text{cylinder}} = \ln\left(\frac{r}{r_2}\right)\left[\ln\left(\frac{r_1}{r_2}\right)\right]^{-1}, \tag{27}$$

where r_2 is the detector outer radius and r_1 is the detector inner radius. The weighting potentials for various r_1/r_2 cases are shown in Fig. 11. With this result, the solution to Eq. 24 for a coaxial detector configuration is

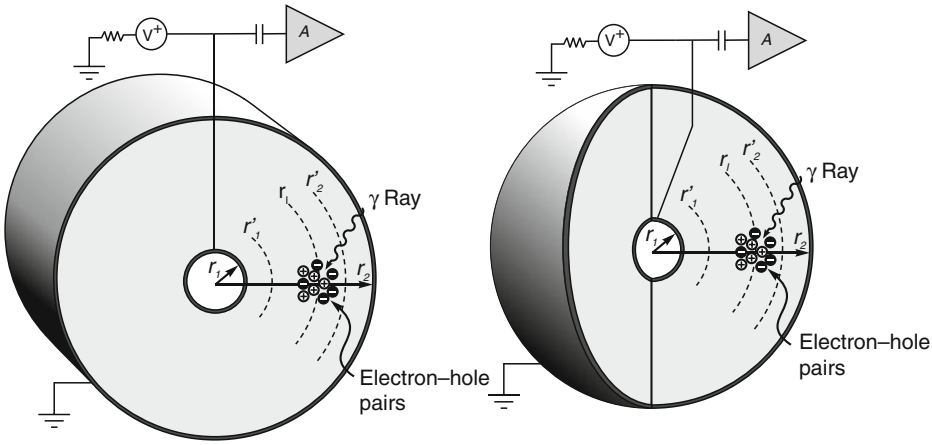


Fig. 10
Cylindrical and spherical semiconductor detector configurations

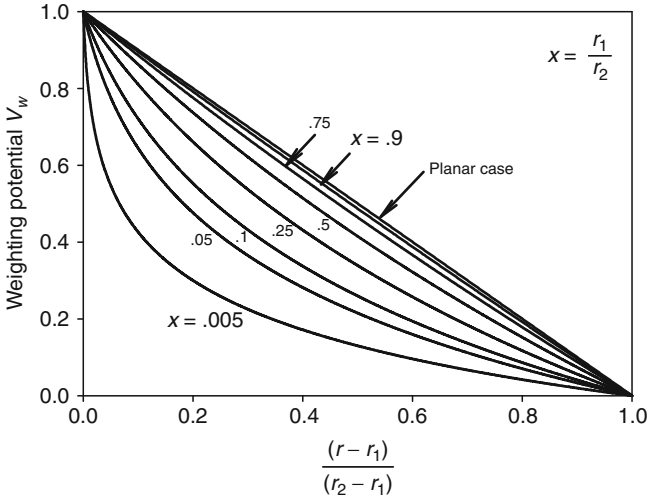


Fig. 11
Weighting potentials for various values of r_1/r_2 versus the normalized distance from r_1 to r_2 for a cylindrical detector

$$\Delta Q|_{\text{cylinder}} = Q_0 \left[\ln\left(\frac{r_2}{r_1}\right)\right]^{-1} \ln\left(\frac{r_2'}{r_1'}\right), \tag{28}$$

where r_1' and r_2' are the drift radii locations for the electrons and holes. Note that the weighting potential is not linear for the cylindrical case, and a larger change in the weighting potential is apparent in the vicinity near the inner electrode at r_1, and, therefore, the induced current is much higher for charges moving near the inner electrode than for charges moving near the outer electrode at r_2.

Similarly, it can also be shown that the weighting potential for a spherical (or hemispherical) detector is

$$V_w|_{\text{sphere}} = \left(\frac{r_1}{r_2 - r_1}\right)\left(\frac{r_2}{r} - 1\right), \tag{29}$$

where r_2 is the detector outer radius and r_1 is the detector inner radius. The weighting potentials for various r_1/r_2 cases are shown in Fig. 12. The solution to Eq. 24 for a hemispherical detector is

$$\Delta Q|_{\text{sphere}} = Q_0 \frac{r_1 r_2}{r_1' r_2'}\left[\frac{r_1' - r_2'}{r_1 - r_2}\right], \tag{30}$$

where r_1' and r_2' are the drift radii locations for the electrons and holes. Note again that the weighting potential is not linear for the spherical case, and a larger change in the weighting potential is apparent near the inner electrode at r_1. The induced current is much higher for charges moving near the inner electrode than for charges moving near the outer electrode at r_2. The weighting potential in detectors with complex geometric shapes can be found by solving Eq. 24 through numerical methods.

Because the capacitance of a detector is highly dependent upon the bias voltage, mainly because the size of the depleted active region changes with bias voltage, it is important that the preamplifier circuit used to sense the induced current is properly matched to the detector.

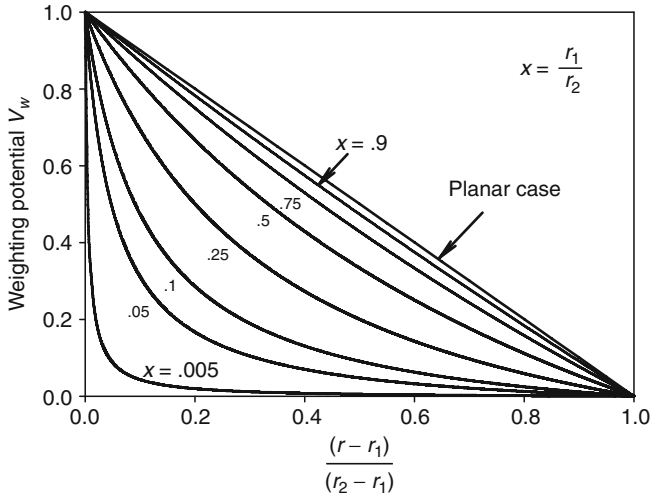


Fig. 12

Weighting potentials for various values of r_1/r_2 versus the normalized distance from r_1 to r_2 for a spherical detector

Further, the preamplifier should be a *charge-sensitive* preamplifier, in which the pulse-height output is largely insensitive to capacitance changes in the semiconductor detector, and is instead mainly dependent upon the charge collected from the detector. Commercial detectors generally have specifications with recommended preamplifier characteristics that can serve to optimize the detector performance.

5 γ-Ray and X-Ray Spectrometers

Properties sought for an ideal γ-ray spectrometer include a wide-band-gap energy, high-Z material composition, high atomic density, long charge-carrier lifetimes, high resistivity, high electron and hole mobilities, and a small ionization energy. A wide-band-gap energy (>1.5 eV) and high resistivity allow room-temperature operation that otherwise would require cryogenic cooling to reduce electronic noise. High atomic density and high-Z components increase the γ-ray interaction probability (see *Fig. 13*). Long charge-carrier lifetimes and high carrier mobilities increase the charge collection efficiency and produce better spectroscopic results. Finally, a small ionization energy causes increased numbers of excited charge carriers, thereby, improving statistics and enhancing spectroscopic energy resolution. Unfortunately, no existing semiconductor actually has all of these characteristics; hence the investigator should select a semiconductor detector best suited for the desired application. The basic properties of several semiconductors used for γ-ray and x-ray detectors are listed in *Table 2*.

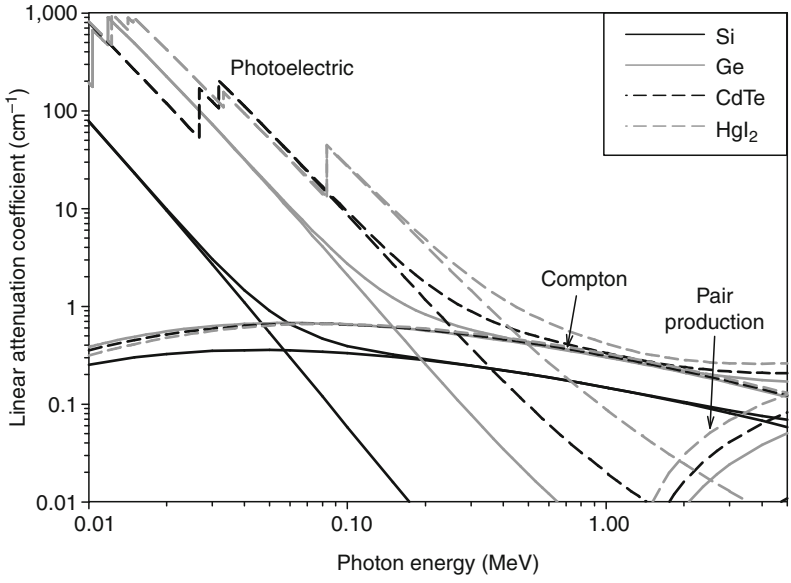


Fig. 13
Photon linear attenuation coefficients as a function of photon energy for Si, Ge, CdTe, and HgI_2 semiconductor materials

Table 2

Common semiconductors used as radiation detectors and their properties at 300 K

Semiconductor	Atomic number (Z)	Density (g cm^{-3})	Band gap (eV)	Ionization energy (eV/e–h pair)	Dielectric constant (ϵ_s/ϵ_0)
Si	14	2.33	1.12	3.61	11.9
Ge	32	5.33	0.68	2.98	16
GaAs	31/33	5.32	1.42	4.2	13.1
CdTe	48/52	6.06	1.52	4.43	10.36
$Cd_{0.9}Zn_{0.1}Te$ (CZT)	48/30/52	6.0	1.60	5.0	10.63
HgI_2	80/53	6.4	2.13	4.3	8.8
Semiconductor	**Intrinsic resistivity (Ω cm)**	**Electron mobility (cm^2/V·s)**	**Hole mobility (cm^2/V·s)**	**Electron lifetime (s)**	**Hole lifetime (s)**
Si	$>5 \times 10^4$	1,500	450	$>10^{-3}$	$>10^{-3}$
Ge	47	3,900	1,900	$>10^{-3}$	$>10^{-3}$
GaAs	$\approx 10^8$	>8,000	400	10^{-9}–10^{-8}	10^{-9}–10^{-8}
CdTe	10^9	1,050	100	3×10^{-6}	2×10^{-6}
CZT	10^{11}	1,350	120	10^{-6}	5×10^{-8}
HgI_2	10^{13}	100	4	$>10^{-6}$	$>10^{-6}$

5.1 X-Ray Detectors Based on Si

Si is a group IV elemental semiconductor with a room-temperature band-gap energy of 1.12 eV. Its low Z number and low density of electrons causes its γ-ray absorption coefficient to be small. Further, the energy at which the photoelectric effect equals the Compton scattering effect is relatively low at only 60 keV. Hence, Si is a poor choice for high-energy γ-ray spectroscopy. However, its K absorption edge appears at 1.838 keV, meaning that the absorption-edge discontinuity does not adversely affect x-ray absorption at higher energies, nor do the appearance of x-ray escape peaks cause significant issues in spectra. By comparison, the K absorption edge for Ge is 11.103 keV. The fact that higher-energy γ rays have less chance of interacting in Si serves to reduce background effects. Energy resolution is quoted in terms of energy spread at the full width at half the maximum (FWHM) of a spectral full-energy peak. Silicon detectors deliver excellent energy resolution, with a FWHM of

$$\mathrm{FWHM} = \left[\left(\mathrm{FWHM}_{\mathrm{noise}} \right)^2 + \left(2.35\sqrt{wFE} \right)^2 \right]^{1/2}, \tag{31}$$

where w is the average energy to produce an electron–hole pair, E is the photon energy, and F is the Fano factor (typically 0.1). The Fano factor is a correction factor to account for typically higher energy resolution than predicted from pure Gaussian statistics. For these reasons, Si does have importance as an x-ray spectrometer for applications such as x-ray fluorescence, x-ray microanalysis, particle-induced x-ray emission (PIXE), x-ray absorption spectroscopy (XAS), x-ray diffraction, and Mössbauer spectroscopy at energies generally below 50 keV.

5.1.1 Basic Design

Highly purified Si can be fashioned into a type of *pin* diode; yet, from Eq. 12, these devices are limited to depleted regions less than 2 mm width, an unsatisfactory thickness for efficient x-ray absorption. The problem is remedied by compensating the remaining impurities in Si with the Li-drifting technique, in which Li ions are electrically introduced deep into the semiconductor under controlled conditions. The resulting devices, denoted as Si(Li) detectors, have active regions ranging from 3–5 mm. The basic design is shown in *Fig. 14*. Although Si(Li) detectors can be operated at room temperature, they perform best when cooled to low temperatures. Various Si(Li) detectors are available coupled to either liquid-nitrogen (LN2) dewars or Peltier coolers.

The detectors are encapsulated in a protective container with a thin entrance window, typically constructed from Be. The entrance window of the detector affects the low-energy sensitivity limit. Shown in *Fig. 15* is the detection efficiency for various Si(Li) detectors with different thicknesses and different entrance windows. Note that thicker detectors increase the efficiency for higher-energy x-rays, while the appropriate choice of entrance window can increase the efficiency for low-energy x-rays.

Si(Li) detectors can be commercially acquired in a variety of segmented patterns, including strips, triangular, and square patterns. The detectors consist of *pin* diode structures individually fabricated into a single Si substrate, thereby, reducing "dead zones" between neighboring detectors. These detectors offer high x-ray energy resolution and spatial interaction information. Further, clever designs can actually improve count-rate efficiency for ion-probe instrumentation,

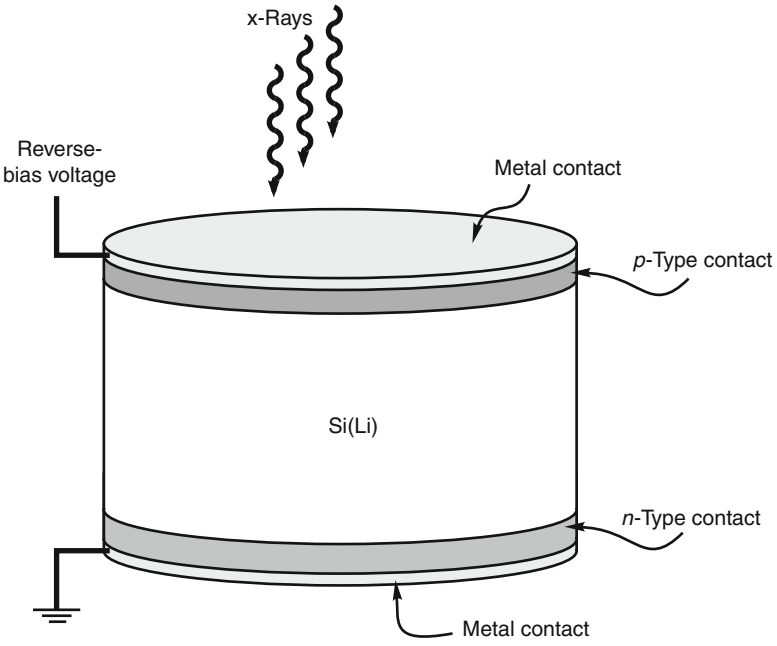


Fig. 14
General configuration of a Si(Li) detector

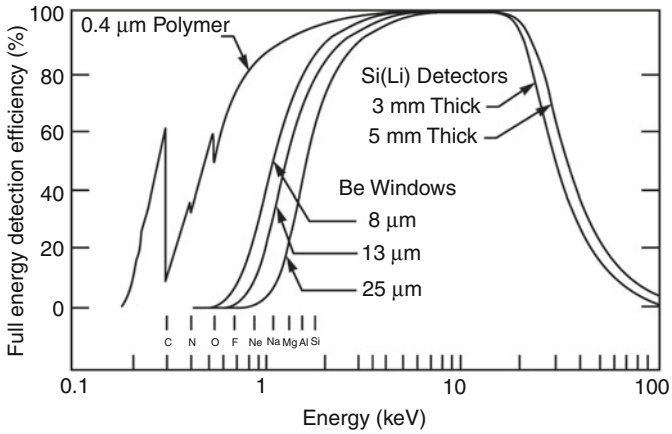


Fig. 15

The absorption efficiency of a Si(Li) detector as a function of x-ray energy, depletion thickness, and entrance window. Courtesy Canberra Industries, Incorporated

such as PIXE, by surrounding the target region with multiple detectors. Used in conjunction with other γ-ray detectors, the segmented Si(Li) detectors can be used for Compton-scatter γ-ray cameras.

5.2 Detectors Based on Ge

Presently, the most popular high-resolution γ-ray spectrometers are constructed from high-purity Ge (HPGe). The material is purified through zone-refinement, resulting in a nearly intrinsic material. Although numerous detector configurations exist, including special-order devices, a standard unit is a coaxial $p\pi n$ or pvn design with the rectifying junction on the outer surface (see *Fig. 16*). The coaxial design permits large detectors to be fabricated while minimizing the detector capacitance. With the rectifying surface on the outer surface, rather than on the inner surface, the active volume is increased and the low-energy γ-ray detection efficiency is improved.

One difficulty with HPGe detector operation is the need to chill the device while operating. Because of the small band-gap energy (0.68 eV), the intrinsic carrier concentration of electrons and holes is much too high at room temperature and significant leakage current is produced when operated at high voltage, which can damage the detector. For this reason, HPGe detectors are typically attached to a dewar and cooled with LN2, or they are attached to a low-noise refrigerator system. To ensure that damage does not occur from excessive leakage current should the LN2 be exhausted, most modern systems have a safety shutoff that disconnects the high voltage if the HPGe detector increases to a preset temperature. Portable survey detectors and laboratory units are available with either LN2 dewars or electrically cooled refrigerators. Hybrid LN2/electrical cryostats have become available, where the main cooler is electrical, backed up with LN2 cooling in case of a power outage.

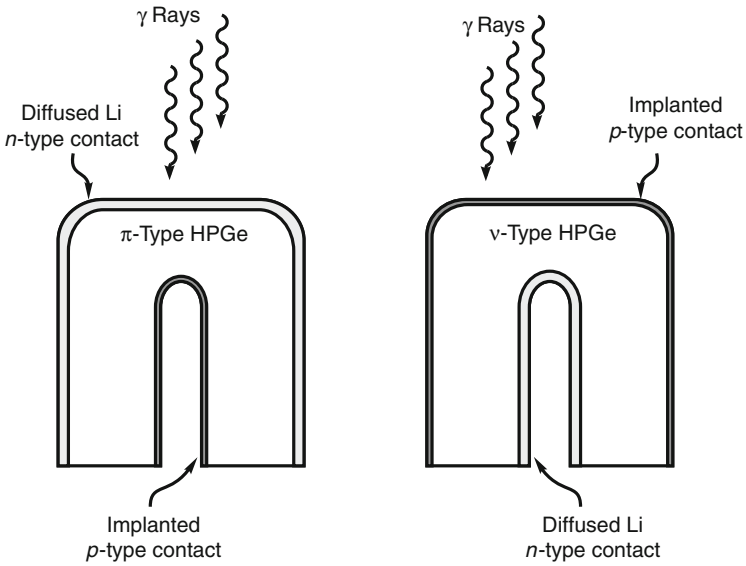


Fig. 16
Typical configurations for π-type and ν-type coaxial HPGe detectors

The usual standard for quality comparisons of HPGe detectors is to quote the energy resolution for 1.33 MeV γ rays from ^{60}Co. The expected energy resolution can be approximated by Eq. 31, where the Fano factor is approximately 0.08. Efficiency, for historical reasons, is quoted most often as a comparison to a 3 in. × 3 in. (7.62 cm × 7.62 cm) right cylindrical NaI(Tl) detector with the source placed 25 cm from the face of either detector. For instance, a relative 30% HPGe detector has 30% of the efficiency expected from a 3 in. × 3 in. NaI(Tl) detector at 1.33 MeV. Although useful as an approximation of detector performance, due to differences in detector geometries and mounting apparatuses, such sweeping generalizations can be erroneous for accurate measurements. It is best to characterize the detector efficiency and resolution, a method described by ANSI/IEEE 325-1996. A calibrated National Institute of Standards ^{60}Co check source is placed 25 cm from the front of the detector face. The number of counts appearing in the full-energy peak for the 1.332 MeV γ ray is divided by the number of emissions over that same time interval, which yields the absolute efficiency. The relative efficiency is found by dividing the absolute efficiency by 1.2×10^{-3}, which is the standard efficiency for a 3 in. × 3 in. NaI(Tl) detector under the same irradiation conditions. A comparison of relative efficiencies and energy resolutions from various HPGe detectors is shown in *Table 3*.

5.2.1 Various Designs

HPGe detectors are manufactured in various shapes, although most conform to either a planar or coaxial design. Small detectors are commonly manufactured as planar detectors. Relatively large HPGe detectors are manufactured as coaxial devices mainly to keep detector capacitance low. Small HPGe detectors usually have better energy resolution than larger devices, and the larger detectors have better γ-ray detection efficiency. The response functions for ν-type and π-type HPGe detectors are quite different at low energies.

Table 3

Typically quoted energy-resolution performance for some commercial semiconductor detectors**

Detector	Area (mm^2)	Radiation type	Energy (keV)	FWHM (keV)	Comments	Source*
Si(Li)	12.5	γ rays	5.9	.155–.175	LN2 cooled	C,O
Si(Li)	20	γ rays	5.9	.180	Peltier cooled	B
		γ rays	59.6	.450	Peltier cooled	
Si(Li)	28–30	γ rays	5.9	.165–.180	LN2 cooled	C,O
Si(Li)	80	γ rays	5.9	.175–.190	LN2 cooled	C,O
Si(Li)	200	γ rays	5.9	.220	LN2 cooled	C,O
Si pin	13	γ rays	5.9	.18–.22	Peltier cooled	A
	25	γ rays	5.9	.127–.230	Peltier cooled	
CdTe	9	γ rays	122	≤1.2	Peltier cooled	A
Schottky	25	γ rays	122	≤1.5	Peltier cooled	
CdZnTe	≈100	γ rays	122	≤6.1	Room temp	E
hemisphere	≈100	γ rays	662	≤20	Room temp	B,E
CdZnTe	100	γ rays	662	13.2–26.4	Room temp	E
coplanar	225	γ rays	662	16.5–26.4	Room temp	
Implanted	100	α particles	5,486	13	W = 100 μm	C,O
Si diode			5,486	12	W = 500 μm	
Implanted	450	α particles	5,486	17–21	W = 100 μm	C,O
Si diode			5,486	15–19	W = 500 μm	
Implanted	900	α particles	5,486	27–33	W = 100 μm	C,O
Si diode			5,486	22–28	W = 500 μm	
p-Type	50	α particles	5,486	15–17	W = 100 μm	O
SSB		α particles	5,486	15–17	W = 500 μm	
p-Type	150	α particles	5,486	16–19	W = 100 μm	O
SSB		α particles	5,486	16–18	W = 500 μm	
p-Type	900	α particles	5,486	30–40	W = 100 μm	O
SSB		α particles	5,486	30–53	W = 500 μm	
HPGe detector	**Relative eff. (%)**	**Radiation type**	**Energy (keV)**	**FWHM (keV)**	**Comments**	**Source***
p-Type	20	γ rays	122	.715–.975	LN2 cooled	B,C,O,P
coaxial	50			.9–1.2	LN2 cooled	B,C,P
	100			1.2–1.4	LN2 cooled	B,C,O,P
p-Type	20	γ rays	1,332	1.8–2.0	LN2 cooled	B,C,O,P
coaxial	50			1.9–2.1	LN2 cooled	B,C,P
	100			2.0–2.3	LN2 cooled	B,C,O,P
n-Type	20	γ rays	122	.69–1.0	LN2 cooled	C,P
coaxial	50			.86–1.2	LN2 cooled	C,P
	70			1.1–1.3	LN2 cooled	C,P
n-Type	20	γ rays	1,332	1.8–2.0	LN2 cooled	C,O,P
coaxial	50			2.1–2.3	LN2 cooled	C,O,P
	70			2.3–2.5	LN2 cooled	C,P

*A = AmpTek, B = Baltic Scientific, C = Canberra, E = EI Detection, O = Ortec, P = PGT.

**These detectors are only a few representative examples and do not account for the numerous variations available, nor a complete list of detector sources. Contact vendors to acquire a full listing of detector sizes and performance statistics.

High-purity π-type detectors are fabricated with Li, an n-type dopant, diffused at a depth of approximately 700 μm thickness around the outer surface. A much thinner implanted junction of p-type dopant (typically boron), approximately 300 nm deep, is formed as the ohmic contact. Consequently, the relatively thick "dead" layer formed by the outer contact significantly reduces the detector sensitivity to low-energy photons (typically below 40 keV). From Eqs. 27 and 28, the reverse-bias configuration and geometry cause the average output pulse to be dominated by hole motion. These $p\pi n$ HPGe detectors typically have slightly better energy resolution than $p\nu n$ HPGe detectors at high γ-ray energies.

High-purity ν-type detectors are fabricated with p-type dopants implanted and activated at a depth of approximately 300 nm for the outer rectifying contact. A much thicker diffused Li junction up to 700 μm thickness is fabricated as the ohmic contact. As result, low-energy γ rays and x-rays encounter less "dead" layer in the outer contact, thereby increasing the efficiency for these low-energy photons. To take further advantage of the thin surface contact, these ν-type detectors are typically packaged in a can that has a thin Be window, thereby, minimizing γ-ray and x-ray attenuation through the detector container. An additional advantage with ν-type HPGe detectors is their increased radiation hardness to neutron radiation. Neutron damage tends to form hole trapping sites, hence the electron-dominated pulses from ν-type HPGe detectors are somewhat less affected.

Examples of efficiency responses for a few HPGe variations are shown in *Fig. 17*. Notice in *Fig. 17* the dip in efficiency at the Ge K absorption edge (11.1 keV). Also note the efficiency reduction below 100 keV for the p-type HPGe detector, which only becomes an issue for the n-type devices represented in *Fig. 17* at energies below 10 keV. The drop in efficiency is due to a combination of photon absorption in the detector-contact dead region and the container holding the detector. Detectors specifically designed for low-energy γ-ray spectroscopy typically have thin Be windows that do not appreciably attentuate γ rays entering the device. Overall, the decision regarding which HPGe detector is best for an application requires some knowledge of the preferred energy resolution, necessary detection efficiency, and the photon energy range of interest.

5.3 Compound Semiconductor Detectors

Compound semiconductor detectors have become more important in recent years, with commercial units now available. Typically, these detectors are somewhat smaller than Si- and Ge-based detectors, mainly due to material imperfections. Regardless, a few materials, namely, CdTe, CdZnTe, and HgI_2, have desirable properties for room-temperature-operated devices, an advantage not shared by Si(Li) or HPGe detectors. The reason for this advantageous property is their larger band-gap energies that work to reduce their instrinsic carrier concentrations and substantially increase their resistivities at 300 K. Further, CdTe, CdZnTe, and HgI_2 all have relatively high-Z atomic constituents, hence have larger γ-ray absorption coefficients over those of Si and Ge. Still, because of their typical smaller size, energy resolution for these compound semiconductor detectors are usually reported relative to 662 keV γ rays of ^{137}Cs instead of 1.33 MeV.

The total charge collected is usually affected by crystalline imperfections that serve as *trapping* sites, which are energy states that remove free charge carriers from the conduction and valence bands. Charge is induced while these charge carriers are in motion; hence, their removal

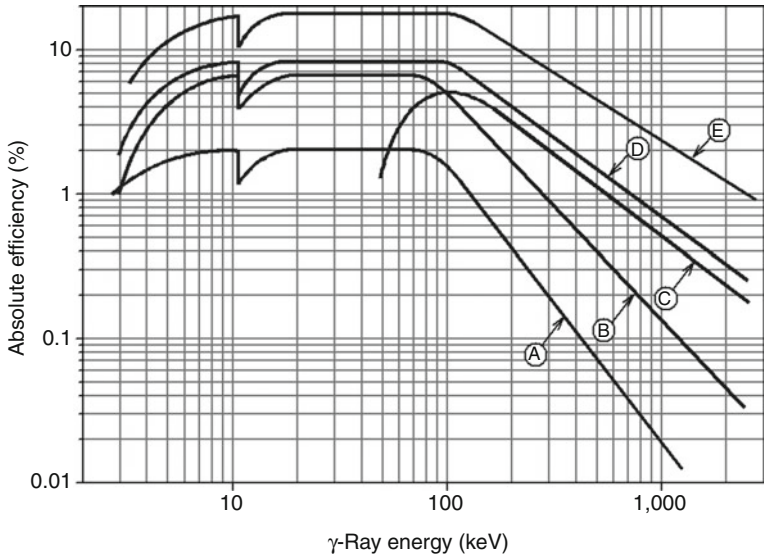


Fig. 17

The absolute detection efficiency for several HPGe detector configurations, showing a (A) 200 mm^2 × 10 mm thick low-energy v-type nominally planar HPGe detector, (B) 10 cm^2 × 15 mm thick low-energy v-type nominally planar HPGe detector, (C) coaxial π-type HPGe detector with 10% relative efficiency, (D) coaxial thin-window v-type 15% relative efficiency HPGe detector, and a (E) broad-energy-range π-type 5,000 mm^2 × 30 mm thick nominally planar HPGe detector. After Canberra Industries, Incorporated (references to Canberra Industries, Inc. and its products are made with the specific permission of Canberra Industries, Inc. for educational purposes only)

diminishes the output voltage. Although the actual trapping process is complicated, it is typical to describe the relative charge collection efficiency as a simplified function of trapping. For planar-shaped detectors, this induced charge is given by

$$\frac{Q}{Q_0} = \xi_e\left(1 - e^{-x/(\xi_e W)}\right) + \xi_h\left(1 - e^{(x-W)/(\xi_h W)}\right), \tag{32}$$

where W is the detector active-region width, Q_0 is the initial excited charge magnitude, x is the event location in the detector, and

$$\xi_{e,h} = \left(\frac{\tau_{e,h} v_{e,h}}{W}\right) = \left(\frac{\mu_{e,h} \tau_{e,h} V}{W^2}\right), \tag{33}$$

where τ is the charge-carrier lifetime, v is the charge-carrier speed, and V is the applied operating voltage. Note, that the relative charge collection is dependent upon the interaction location x, and for low values of ξ, the energy resolution is poor. Typically, good energy resolution is achieved if $\xi > 50$ for both electrons and holes, where Q/Q_0 has little deviation over the detector width W. Otherwise, the energy resolution suffers for higher-energy γ rays (more or less > 300 keV). The value of ξ can be increased by decreasing the detector size (W), increasing carrier lifetimes (τ) through material improvement, or increasing the applied voltage V. Due to practical

voltage limitations and the fundamental difficulty with improving materials, most compound semiconductor detectors are manufactured with small active widths to improve detector energy resolution, and, hence, the devices are relatively small. The $\mu\tau$ values for electrons and holes are often quoted measures of quality for compound semiconductors used as γ-ray spectrometers.

5.3.1 CdTe

CdTe, with a band-gap energy of 1.56 eV, has been explored as a commercial γ-ray detector since the 1960s, yet due to material imperfections, mainly impurities, these devices continue to be manufactured as small detectors. Although the band gap is high enough for room-temperature operation, background impurity contamination causes the leakage currents to be too high. Dopant compensation, typically with Cl, is used to create high-resistivity material. Still the detectors must be manufactured as *pn*-junction or Schottky-junction diodes to reduce leakage current to manageable levels. As a result, the detector volumes are usually no more than a few mm thick. Commercial units are available as small γ-ray spectrometers. Typically, the best energy resolution is achieved with the assistance of small electronic Peltier coolers.

5.3.2 CdZnTe

The introduction of Zn in the growth process of CdTe, nominally between 2% to 10%, has led to the production of CdZnTe detectors. The devices have the same advantage as CdTe detectors, and, due to improved materials properties, they can be manufactured much larger than conventional CdTe detectors. The band gap ranges from 1.6 to 1.65 eV for Zn concentrations ranging from 2% to 10%; hence the detectors can operate at room temperature without leakage-current issues. Because of the high resistivity, the detectors are typically manufactured with ohmic contacts for the cathodes and anodes and operate as resistive detectors. The detectors have adequate electron transport properties, but poor hole transport properties. As a result, conventional planar detectors, similar to the depiction in *Fig. 9*, seldom produce useful energy resolution for moderate- to high-energy γ rays (≥300 keV). Instead, some commercial manufacturers rely upon clever geometric detector shapes and electrode contact shapes to modify the weighting potential and electronic signal such that electrons dominate signal formation rather than holes. Energy resolution below 7 keV FWHM for 662 keV γ rays can be achieved at room temperature for these *single-carrier* detector designs. CdZnTe detectors are presently used in handheld γ-ray spectrometers and for smaller medical imaging apparatuses.

5.3.3 HgI_2

Attractive for its large-Z components, HgI_2 has long been studied as a γ-ray spectrometer with varying degrees of success. They have fewer commercial applications than CdTe or CdZnTe, mainly because of process and fabrication issues. The material is known to *polarize* over time, which is manifested as a the gradual change in spectra over time. Regardless, the material has been used for portable x-ray spectrometers for in situ analysis, and modular units are available for specialized room-temperature spectroscopy applications.

6 Charged-Particle Detectors

Semiconductor charged-particle spectrometers offer high-energy resolution for energetic ions. Typically, the devices are operated in vacuum, along with the source, to eliminate energy losses from a charged particle as its passes from the source to the detector. The detectors are typically designed with thin contacts and/or thin *pn* junctions in order to reduce particle energy loss in the nonsensitive (or dead) region of the contact. Because low-Z elements have less problems with ion backscattering, Si is typically the material choice for particle detectors. The depletion width as a function of material resistivity (n-type and p-type) is shown in ❯ *Fig. 18*. These detectors can be used for a variety of charged-particle identification and characterization purposes, including high-resolution spectroscopy of α particles, β particles, protons, and heavy ions, continuous air monitoring, and particle telescopes.

6.1 Surface-Barrier and Implanted-Junction Detectors

Si surface-barrier (SSB) detectors rely upon the production of a thin Schottky barrier for a rectifying junction. A typical SSB-detector cross section is shown in ❯ *Fig. 19*. High-purity n-type or p-type Si is etched, mounted, and epoxied into a ceramic ring. Afterward, a thin layer of Au or Al, ranging from 80 to 200 nm, is applied to the semiconductor surfaces. The thin contact region minimizes the amount of energy lost by particles that enter the device, a necessary precaution to preserve high-energy resolution. However, these delicate surface-barrier detectors can be easily damaged by improper handling and are often difficult to clean. Detectors can be obtained in a variety of sizes, ranging from a few mm to 50 mm diameter. These detectors are usually light sensitive and must be operated in darkness, although commercial companies do offer versions that can operate in ambient light, at the expense of energy resolution. Depletion depths range from approximately 100 μm up to, for special cases, 5 mm.

Implanted-junction detectors rely upon an abrupt junction *pn* diode for rectification (see ❯ *Fig. 19*). These devices are commonly fabricated from high-purity n-type Si. An oxide is grown on the devices for passivation, followed by etching windows back to the Si surface. Shallow p-type and n-type dopants are implanted on opposite sides of the Si surface and thermally activated. This process produces dead-layer junctions on the order of only 50 nm. Because there is no thin metalization layer over the detector, they are more robust and easier to clean than common SSB detectors. Implanted-junction detectors can be used for the same basic detection functions that SSB detectors are used.

SSB and implanted-junction detectors can be acquired in numerous shapes, sizes, and configurations, making them a versatile choice for particle detection and spectroscopy. Further, the detectors can be acquired as multielement arrays for position sensing and timing purposes. The adaptation of very large scale integration (VLSI) processing technology to Si detectors allows for detector arrays to be fabricated in a vast number of detector designs, including custom devices contracted to commercial vendors. The detectors are available as double-sided strip detectors with spatial resolutions as low as 25 μm and pad detectors with spatial resolution as small as 0.4 mm. Large arrays of position-sensitive Si detectors can be used in collider facilities, x-ray scattering, and Compton cameras. Finally, drift-diode configurations, a variant design that drifts electronic charge carriers laterally along the detector to a small collection contact, offer low capacitances with large sensitive areas.

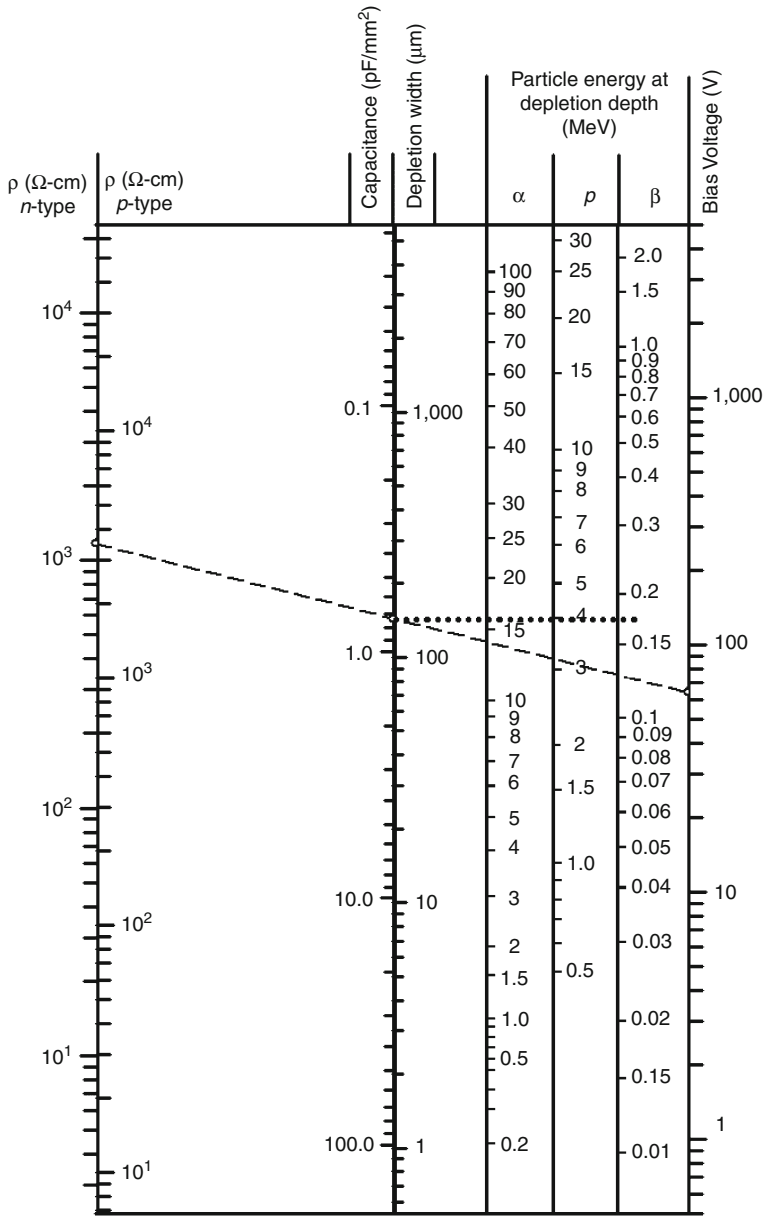


Fig. 18

The depletion region as a function of material resistivity and reverse voltage for Si particle detectors. Also shown are the ranges of α, β (or electron), and proton particles as a function of energy. A *straight line* will yield material resistivity (*n*- or *p*-type), the depletion depth and applied voltage, according to Eqs. 16 and 19. The corresponding depletion depth needed to fully absorb protons, α and β particles, as a function of energy, is found with an *intersecting perpendicular line* from the depletion-width axis (after Blankenship and Borkowski (1960))

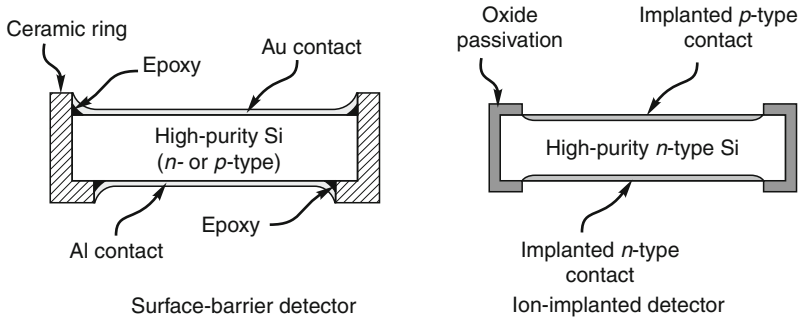


Fig. 19
General configurations for Si surface-barrier detectors and implanted-junction detectors

7 Neutron Detectors

Semiconductor radiation detectors used as neutron detectors are typically configured as *pn*-junction or Schottky-junction diodes coated with a neutron reactive material. The basic construction of such a detector is shown in *Fig. 20*, where a Schottky- or *pn*-junction diode detector has a coating of neutron-reactive material applied to the surface. Typically the devices have either ^{10}B or ^{6}LiF as the active coating. The absorption cross sections for both ^{10}B and ^{6}Li follow a $1/v$ dependence. The $^{10}\mathrm{B}(\mathrm{n},\alpha)^{7}\mathrm{Li}$ neutron reaction yields two possible de-excitation branches from the excited ^{11}B compound nucleus, namely

$$ {}^{1}_{0}\mathrm{n} + {}^{10}_{5}\mathrm{B} \longrightarrow \begin{cases} {}^{4}_{2}\mathrm{He}\,(1.4721\ \mathrm{MeV}) + {}^{7}_{3}\mathrm{Li}^{*}\,(0.8398\ \mathrm{MeV}) & (93.7\%) \\ {}^{4}_{2}\mathrm{He}\,(1.7762\ \mathrm{MeV}) + {}^{7}_{3}\mathrm{Li}\,(1.0133\ \mathrm{MeV}) & (6.3\%)\,, \end{cases} $$

where the Li ion in the 94% branch is ejected in an excited state, which de-excites through the emission of a 480 keV γ ray. Fully enriched ^{10}B has a microscopic absorption cross section for thermal neutrons of 3,840 b. With a mass density of 2.15 g cm^{-3}, the solid structure of ^{10}B has a macroscopic thermal absorption cross section of 500 cm^{-1}.

The $^{6}\mathrm{Li}(\mathrm{n},\mathrm{t})^{4}\mathrm{He}$ neutron reaction yields a single product branch,

$$ {}^{1}_{0}\mathrm{n} + {}^{6}_{3}\mathrm{Li} \longrightarrow {}^{3}_{1}\mathrm{H}\,(2.7276\ \mathrm{MeV}) + {}^{4}_{2}\mathrm{He}\,(2.0553\ \mathrm{MeV}). $$

The reaction products from the $^{6}\mathrm{Li}(\mathrm{n},\mathrm{t})^{4}\mathrm{He}$ reaction are more energetic than those of the $^{10}\mathrm{B}(\mathrm{n},\alpha)^{7}\mathrm{Li}$ reaction and, hence, are much easier to detect and discriminate from background radiations. ^{6}Li has a relatively large microscopic thermal neutron absorption cross section of 940 b, although it is less than that of ^{10}B. Unfortunately, Li is a chemically reactive metal, and, therefore, it is the stable compound ^{6}LiF, with a macroscopic cross section of 57.51 cm^{-1}, that is used as the reactive coating.

For thermal neutrons, the charged-particle reaction products are ejected in opposite directions, meaning that only one reaction product can actually enter the semiconductor detector. Further, the reaction products lose energy as they pass through the neutron absorber to the semiconductor detector, thereby limiting the effective absorber thickness. Detectors of this type are limited to less than 5% thermal-neutron detection efficiency. These devices are generally not

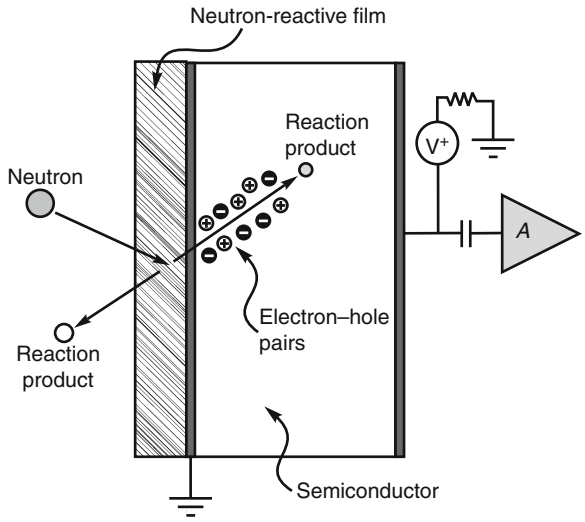


Fig. 20
Cross section of a coated semiconductor neutron detector

commercially available as independent units; rather they are sold as an active component inside some electronic dosimeter modules.

8 Conclusions

Semiconductor materials are attractive as radiation detectors for at least two main reasons. First, due to their low average ionization energy w, semiconductors produce a large number of signal charge carriers per unit energy, thereby decreasing statistical fluctuations beyond that of gas-filled and scintillation detectors, hence producing much better energy resolution. Second, semiconductor materials have energy-band structures that allow their electrical properties to be altered through the addition of impurities. These materials can be manipulated to have a majority of negative (n-type) electrical charge carriers, or electrons, or positive (p-type) charge carriers, denoted "holes." Adjacent n-type and p-type materials can be manipulated to form detectors with rectifying contacts, which work to reduce both leakage current and electrical noise.

Semiconductor detectors can be fashioned into various detectors especially designed for x-ray detection, γ detection, or charged-particle detection. Detector performance is optimized by semiconductor choice and device design. Charged-particle detectors are generally designed with low-Z material, such as Si, to reduce backscattering. Higher-Z materials, due to improved absorption efficiency, are generally used for γ-ray detectors. Low-energy x-ray and γ-ray detectors are often fabricated from Li-drifted Si (Si(Li) detectors), although the most commonly used semiconductor for photon detection is high-purity Ge (HPGe detectors).

Both Si(Li) and HPGe detectors must be cooled to low temperature for best operation. Wide-band-gap semiconductors, such as CdZnTe, can be used as room-temperature-operated γ-ray spectrometers.

Finally, semiconductor materials can be fashioned into arrays to yield spatial interaction information. These arrays can be arranged from the tiling of numerous individual detectors, or they can be fabricated as pixels upon a single semiconductor substrate. Commercial vendors offer numerous varieties of semiconductor detectors, which include particle, x-ray and γ-ray detectors, in the form of individual devices or as arrays.

9 Cross-References

Although many chapters in the present text may describe uses for semiconductor detectors, the following chapters that include information related to this chapter may be of particular interest to the reader.

Chapter 17, "Gamma-Ray Detectors," covers various detectors used for γ-ray spectroscopy, including the performance of semiconductor detectors. Also covered are the typical methods used for energy resolution characterization.

Chapter 21, "New Solid State Detectors," covers some information on recent novel detector designs.

Chapter 13, "Photon Detectors," also covers some information on detectors used for x-ray and γ-ray detection.

References

Blankenship JL, Borkowski CJ (1960) Silicon surface barrier nuclear particle detectors. IRE Trans Nucl Sci NS-7:190–195

Bertolini G, Coche A (1968) Semiconductor detectors. Wiley, New York

Granger R (1994) In: Capper (ed) Properties of narrow gap cadmium-based compounds. INSPEC, London, UK, Chapter B3.2: 433–435

He Z (2001) Review of the Shockley–Ramo theorem and its application in semiconductor gamma-ray detectors. Nucl Instrum Methods A463: 250–267

Henisch HK (1984) Semiconductor contacts; an approach to ideas and models. Oxford, Clarendon Press

Knoll GF (2010) Radiation detection and measurement, 4th edn. Wiley, New York

Lutz G (1999) Semiconductor radiation detectors. Springer, Berlin

Martini M (1986) Introduction to charged particle detectors, ORTEC technical note, ORTEC, Oak Ridge, TN

McGregor DS, Hammig MD, Gersch HK, Yang Y-H, Klann RT (2003) Design considerations for thin film coated semiconductor thermal neutron detectors. Nucl Instrum Methods A500:272–308

Pierret RF (1989) Advanced semiconductor fundamentals. Addison Wesley, Reading

Rhoderick EH, Williams RH (1988) Metal-semiconductor contacts, 2nd edn. Oxford, Clarendon Press

Schlesinger TE, James RB (1995) Semiconductors for room temperature nuclear detector applications, in semiconductors and semimetals, vol 43. Academic Press, San Diego

Schwartz B (1969) Ohmic contacts to semiconductors. Electrochemical Society, New York

Sharma BL (1984) Metal-semiconductor schottky barrier junctions and their applications. Plenum Press, New York

Sze SM (1981) Physics of semiconductor devices, 2nd edn. Wiley, New York

Tench O (2008) Germanium detectors, Canberra technical note, Canberra Industries Inc., Meriden, CT

More References for the Interested Reader

Dearnaley G, Northrop DC (1966) Semiconductor counters for nuclear radiations, 2nd. edn. Wiley, New York

Deme S (1971) Semiconductor detectors for nuclear radiation measurement. Wiley, New York

Hannay NB (1956) Semiconductors. Reinhold, New York

Martini M, Ottaviani G (1969) Ramo's theorem and the energy balance equations in evaluating the current pulse from semiconductor detectors. Nucl Instrum Methods 67:177–178

Mckelvey JP (1966) Solid state and semiconductor physics. Academic Press, New York

Nussbaum A (1962) Semiconductor device physics. Prentice Hall, Englewood Cliffs

Poenaru DN, Vilcov N (1969) Measurement of nuclear radiations with semiconductor detectors. Chemical Pub., New York

Ramo S (1939) Currents induced by electron motion. Proc IRE 27:584–585

Shockley W (1950) Electrons and holes in semiconductors. Van Nostrand, Princeton

Taylor JM (1963) Semiconductor particle detectors. Butterworth, London

Semiconductor Radiation Detector Suppliers

AmpTek, Inc.; www.amptek.com/index.html

Baltic Scientific Instruments; www.bsi.lv/

Canberra Industries, Inc.; www.canberra.com/

Constellation Technology Corporation; www.contech.com/

EI Detection and Imaging; www.evmicroelectronics.com/

Eurorad; www.eurorad.com/detectors.php

Moxtek; www.moxtek.com/

Ortec Advanced Measurement Technology, Inc.; www.ortec-online.com/

Princeton Gamma-Tech Instruments; www.pgt.com/

Radiation Monitoring Devices, Inc.; www.rmdinc.com/

Redlen Technologies; www.redlen.com/

Thermo-Eberline; www.esm-online.de/sm/contact/index.html

XRF Corporation; www.xrfcorp.com/

17 Gamma-Ray Detectors

William L. Dunn · Douglas S. McGregor
Kansas State University, Manhattan, KS, USA

1 Nomenclature 413

2 Introduction 414

3 Basic Concepts 414

4 Detector Response Models 417
4.1 Spectral Features 418
4.2 Detector Response Function 420

5 Qualitative Analysis 420

6 Quantitative Analysis 422
6.1 Area Under the Peak 422
6.2 Model Fitting 423
6.2.1 Isolated Peaks 424
6.2.2 Overlapping Peaks 425
6.2.3 Weighted Least Squares 426
6.3 Spectrum Stripping 428
6.4 Library Least Squares 429
6.4.1 Nonlinear Spectra 430
6.4.2 Compton Suppression 430
6.5 Symbolic Monte Carlo 431

7 Detectors for Gamma-Ray Spectroscopy 431
7.1 Scintillation Spectrometers 432
7.1.1 Inorganic Scintillators 432
7.1.2 Light Collection 436
7.1.3 Factors Affecting Energy Resolution 438
7.2 Semiconductor Spectrometers 439
7.2.1 Ge Detectors 441
7.2.2 Si Detectors 442
7.2.3 Compound Semiconductor Detectors 443
7.2.4 Factors Affecting Energy Resolution 444
7.3 Cryogenic Spectrometers (Microcalorimeters) 446
7.4 Crystal Diffractometers (Wavelength-Dispersive Spectroscopy) 446

C. Grupen, I. Buvat (eds.), *Handbook of Particle Detection and Imaging*,
DOI 10.1007/978-3-642-13271-1_17, © Springer-Verlag Berlin Heidelberg 2012

8 Conclusions *449*

9 Cross-References *449*

Acknowledgments *449*

References *449*

Further Reading *451*

Radiation Spectrometer Suppliers *451*

Abstract: The common methods of analyzing gamma-ray spectra obtained from detectors capable of energy discrimination are discussed. Gamma-ray spectra generally are in the form of detector response versus discrete channel number. The methods considered for gamma-ray spectroscopy are somewhat general and can be applied to other types of spectroscopy. The general objective of spectroscopy is to obtain, at a minimum, the qualitative identification of the source (e.g., source energies or nuclides present). However, most spectroscopy applications seek quantitative information also, as expressed by, e.g., the source strength or the nuclide concentration. Various different methods for qualitative and quantitative analysis are summarized, and an illustrative example is provided. A review of detectors used for gamma-ray spectroscopy is included.

1 Nomenclature

Most of the symbols used in this chapter are briefly identified here. More complete descriptions of the symbols are provided within the text.

Symbol	Description	Symbol	Description
A_j	net peak counts	$\mu_{e,h}$	charge mobility (electrons, holes)
$b(h)$	background density function	m	number of overlapping peaks
b_n	background rate in channel n	n	continuous channel number
B_n	background counts in channel n	n_d	discrete channel number
c	response density function	N	number of channels
C	response cumulative function	N_0	initial number of charge pairs
C_h	heat capacity	ν	number of degrees of freedom
d	spacing between crystal planes	ξ	nuclide concentration
Δ	channel width	$\xi_{e,h}$	carrier extraction factor
E	photon energy	$\mathbf{p}$	vector of model parameters
E^2	least-squares function	q	unit electronic charge
E'	apparent energy	Q	total induced charge
E_d	energy deposited in detector	Q_0	initial excited charge
E_j	jth discrete energy emitted by source	$\mathcal{R}$	detector resolution
e_n	stochastic error in channel n	r_n	response in channel n
f	function that relates h to E_d	$s(E)$	source density function
F	Fano factor	s_c	continuum emission rate
FWHM	full width at half maximum	s_j	emission rate at energy E_j
g	Gaussian density function	S_j	E_j emission rate over time t
G	dynode gain	σ	standard deviation
Γ	full width at half maximum	t	counting time
h	pulse height	T	absolute temperature
h_0	centroid of pulse-height peak	$\tau_{e,h}$	charge-carrier lifetime
η	detector efficiency	u	detector response function
j	subscript for discrete energy	$v_{e,h}$	charge-carrier velocity
J	number of discrete energies	w	average ionization energy
k	subscript for nuclide	W	weight factor
k	Boltzmann's constant	W_d	detector width
K	number of nuclides	χ^2_ν	reduced chi-square function
λ	photon wavelength	y	response model
M	subset of N	z	composite detector kernel
μ	peak channel centroid	z_{jn}	discrete response kernel

2 Introduction

Gamma rays and X rays are photons of electromagnetic radiation that are capable of causing ionization. Technically, X rays differ from gamma rays in their source of origin but, for practical purposes, this is irrelevant and photon spectra can be analyzed by the same methods whether the source photons are X rays or gamma rays. As a practical matter, photons of energy less than about 10 keV are difficult to detect because they are easily absorbed by the detector housing. The concepts that are discussed here can be applied, in principle, to spectra from photons of energy less than 10 keV and also to spectra generated by other particles, such as electrons. For instance, X-ray photoelectron spectroscopy (XPS) and electron scattering for chemical analysis (ESCA) lead to electron spectra that can be analyzed by the methods discussed here. Thus, it is understood that reference to gamma-ray spectroscopy is an oversimplification and many of the methods discussed here can be applied to spectra generated by X-rays, gamma rays, or other types of radiation.

Gamma-ray spectroscopy is a general area of study within which spectra are analyzed in order to determine qualitative and, if possible, quantitative information about a sample under investigation. Spectra generally refer to collections of data for which the independent variable is channel number (or a related quantity such as pulse height, energy, or wavelength) and the dependent variable is a detector response that depends on the independent variable. Spectra are generated in various processes, such as energy-dispersive X-ray fluorescence (EDXRF), neutron activation analysis (NAA), prompt-capture gamma-ray neutron activation analysis (PGNAA), XPS, and general counting of unknown radioactive sources. Often, spectroscopy is directed at the quantitative objective of identifying concentrations of specific elements or isotopes (henceforth the generic term "nuclides" is used) that are present in samples, but it also can be employed in a qualitative manner to identify whether specific gamma-ray-emitting nuclides, such as those in special nuclear materials, are present in samples.

In general, a sample that is under investigation emits photons whose energies are characteristic of the nuclides present in the sample. The photons may be excited by an external source or the sample may emit these photons naturally. A careful spectroscopic investigation generally seeks to determine either the energies emitted and their intensities or the nuclides present and their concentrations. In the remainder of this chapter, attention is given to photon spectra that are generated from samples interrogated by any of a number of means, active or passive, to determine information about the constituents of the sample.

3 Basic Concepts

Various photon detectors can provide responses that are proportional to the energy deposited in the detector. These include proportional counters, scintillation detectors, and semiconductor detectors. Regardless of the detector, a voltage pulse is created whose amplitude h, generally called the *pulse height*, is a function of the energy deposited E_d in the detector, i.e.,

$$h = f(E_d), \tag{1}$$

where f is some function. For many detectors, f is linear and

$$h = \alpha + \beta E_d. \tag{2}$$

However, scintillation detectors, in particular, can exhibit nonlinearities, especially at low photon energies (Knoll 2010), and this should be taken into account. (Linearity of a detector is often expressed in terms of the pulse height per unit energy as a function of energy, which is constant for a linear detector.) In any event, a good spectroscopist should know f, the functional relationship between pulse height and energy, for any detector that he or she uses.

In gamma-ray spectroscopy, the pulse height h is measured and the deposited energy is obtained by inversion of Eq. 1, namely

$$E_d = f^{-1}(h), \tag{3}$$

or, if the spectrometer is linear, then

$$E_d = \frac{h - \alpha}{\beta}. \tag{4}$$

However, the pulse heights produced by repeated deposition of energy E_d in a detector are not exactly the same; rather, they are distributed about a mean value h_0. In effect, the detector systems that produce the voltage pulses operate on the energy deposited with a spreading kernel $z(E_d \rightarrow E')$ that transforms E_d into a range of "apparent" energies E', which are centered on E_d. In gamma-ray spectroscopy, this spreading kernel usually is assumed to have a normal or Gaussian shape, as depicted in *Fig. 1*. Thus, the distribution of pulse heights is given by the probability density function (PDF)

$$g(h|h_0, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-[(h-h_0)^2/(2\sigma^2)]}, \tag{5}$$

where h_0 is the mean value or centroid and σ is the standard deviation.

The peak value of the Gaussian PDF occurs at the centroid and is given by

$$g_{max} = \frac{1}{\sqrt{2\pi}\sigma}.$$

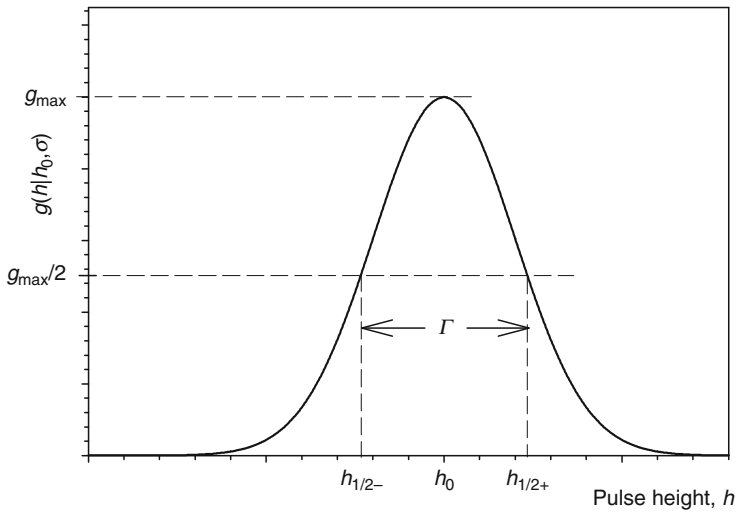


Fig. 1

A Gaussian distribution of pulse heights around a centroid of h_0

The resolution $\mathcal{R}$ of a detector is a function of h_0 and is defined to be the full width at half maximum (FWHM) Γ of the Gaussian peak divided by the value of the centroid of the peak, i.e.,

$$\mathcal{R} = \frac{\Gamma}{h_0}. \tag{6}$$

As shown in Fig. 1, the FWHM is the width of the peak when the Gaussian is half the maximum, which occurs at the two pulse heights of $h_{1/2-}$ and $h_{1/2+}$. For the PDF to have the value $g_{\max}/2$, it must be true that

$$e^{-[(h_{1/2+}-h_0)^2/(2\sigma^2)]} = 0.5. \tag{7}$$

One can take logarithms of Eq. 7 to show that, because $\Gamma = 2(h_{1/2+} - h_0)$, the FWHM can be expressed in terms of the Gaussian standard deviation as

$$\Gamma = 2\sqrt{2\ln 2}\sigma \simeq 2.355\sigma. \tag{8}$$

It is worth noting that the energy resolution of semiconductor detectors is, in general, significantly better than for scintillation detectors or proportional counters. McGregor (2008) gives a good comparative description of the resolution of various detector types. Each type of detector has its advantages and disadvantages. Section 7 of this chapter and the chapters in this book on the various detector types present useful information on the characteristics of some types of detectors.

The pulse heights are scaled to be within a finite interval $(h_{\min}, h_{\max})$. Typically, these limits are $h_{\min} = 0$ and $h_{\max} = 10$ V. In any case, the variables E_d and h are continuous variables. Whatever the pulse-height limits are, the response of a detector is typically binned into discrete "channels." Let n_d refer to individual discrete channel numbers, N be the total number of channels, and

$$\Delta = \frac{h_{\max} - h_{\min}}{N}$$

be the channel width. The channel numbers are related to the detected magnitudes of the voltage pulses (pulse heights) by the relations

$$\begin{aligned} n_d &= 1, \text{ if } h_{\min} < h \le h_{\min} + \Delta \\ &= 2, \text{ if } h_{\min} + \Delta < h \le h_{\min} + 2\Delta \\ &\vdots \\ &= N, \text{ if } h_{\min} + (N-1)\Delta < h \le h_{\min} + N\Delta. \end{aligned} \tag{9}$$

Thus, the measured continuous pulse heights are converted into discrete channels and each pulse registers a count in one and only one channel. Typical spectroscopic systems now operate with $N = 2^{13} = 8{,}192$ channels.

Although there are a finite number of discrete channels, corresponding to the pulse-height intervals specified in Eq. 9, the peak centroid can occur at any continuous value of h. Thus, it is customary to specify a linear relationship between the continuous values of h and a continuous channel number n, given by

$$n = \frac{h - h_{\min}}{\Delta}, \tag{10}$$

which varies continuously between 0 and N. Henceforth, the term channel number is used to mean either the discrete channel number n_d or the continuous channel number n; the context can be used to infer the intended variable.

4 Detector Response Models

Consider a source that emits photons with some source distribution $s(E)$ and a spectrometer that detects the photons and produces a pulse-height spectrum. The basic spectroscopic relationship in its continuous form can be written

$$c(h) = \frac{\mathrm{d}C(h)}{\mathrm{d}h} = \int_0^\infty s(E)z(h|E)\,\mathrm{d}E + b(h), \quad h \in [h_{\min}, h_{\max}], \tag{11}$$

where

- $c(h)$ is the detector response such that $c(h)\,\mathrm{d}h$ is the number of counts within $\mathrm{d}h$ about h per unit time
- $s(E)$ is the source strength, in photons per unit time, such that $s(E)\,\mathrm{d}E$ is the number of source photons emitted within $\mathrm{d}E$ about E per unit time
- $z(h|E)$ is a composite kernel such that $z(h|E)\,\mathrm{d}h$ gives the probability that a particle of energy E strikes the detector and produces a pulse whose height is within $\mathrm{d}h$ about h
- $b(h)$ is the background count rate such that $b(h)\,\mathrm{d}h$ is the expected number of counts within $\mathrm{d}h$ about h, per unit time, that are due to background radiation

Note that $C(h) = \int_0^h c(h')\,\mathrm{d}h'$ is the cumulative number of counts per unit time due to pulses whose heights are less than h and that $C(h_2) - C(h_1) = \int_{h_1}^{h_2} c(h)\,\mathrm{d}h$ is the total number of counts per unit time whose pulse heights are between h_1 and h_2. Note also that $z(h|E)\,\mathrm{d}E$ is the probability of transport of source photons whose energies are within $\mathrm{d}E$ about E to and within the detector, deposition of energy E_d in the detector, and spreading of the deposited energy into an apparent energy E' that produces a pulse whose pulse height is h.

In general, a source can emit photons at J discrete energies and also over a continuum of energies. For such sources,

$$s(E) = \sum_{j=1}^{J} s_j \delta(E - E_j) + s_c(E), \tag{12}$$

where s_j is the emission rate of photons of energy E_j from the source, $\delta(E - E_j)$ is the Dirac delta function, and $s_c(E)$ is the source strength of the photons emitted over a continuum of energies such that $s_c(E)\,\mathrm{d}E$ is the expected number of photons emitted within $\mathrm{d}E$ about E per unit time. For a detector that sorts counts into discrete channels, one can substitute ❯ Eq. 12 into ❯ Eq. 11, integrate over each channel width, and write the discrete form of the pulse-height spectrum in the form

$$c_n = \sum_{j=1}^{J} s_j z_{jn} + \int_0^\infty s_c(E) z_n(E)\,\mathrm{d}E + b_n, \quad n = 1, 2, \ldots, N, \tag{13}$$

where c_n is the number of counts recorded in channel n per unit time, z_{jn} is the probability that a particle of initial energy E_j that is emitted by the source produces a count within the nth channel, b_n is the expected number of counts recorded in channel n, per unit time, due to background radiation, and

$$z_n(E) = \int_{(n-1)\Delta}^{n\Delta} z(h|E)\,\mathrm{d}h.$$

In general, spectra are obtained over a counting time t, in order to improve the statistical precision of the measurements. The total counts obtained over counting time t per channel can

be obtained by integrating Eq. 13 over the counting time. The continuum source term can be due to photons emitted from the source over a continuum of energies and/or to photons emitted at discrete energies from the source that scatter into the detector. Because the primary objective of spectroscopy is to find the E_j and s_j, for $j = 1, 2, \ldots, J$, it is customary to combine the continuum and background terms into a generalized background. Doing so and integrating over counting time, one obtains

$$\boxed{r_n = \sum_{j=1}^{J} S_j z_{jn} + B_n}, \quad n = 1, 2, \ldots, N, \tag{14}$$

where $r_n = c_n t$ is the detector response in channel n, $S_j = s_j t$, and

$$B_n = \left[\int_0^\infty s_c(E) z_n(E)\, \mathrm{d}E + b_n \right] t. \tag{15}$$

A plot of r_n versus n is called a *pulse-height spectrum*. Integrating over any range of channels gives the total counts within those channels.

Equation 14 poses an inverse problem, in which the specific j and S_j, $j = 1, 2, \ldots, J$ are sought, given $M \leq N$ measured pulse-height responses. The generalized background is typically not known explicitly, complicating the inversion process. There are different methods that can be used to solve this inverse problem, and the number of channels, $M \leq N$, used depends on the method chosen. Because most spectrometers have thousands of channels, most spectroscopy inverse problems are overdetermined in the sense that there are considerably more responses available than unknowns.

It is noted that a variant of this inverse problem often is posed in terms of the $k = 1, 2, \ldots, K$ nuclides present in a sample, each of which can emit one or more photons at discrete energies. Rather than look for the J discrete energies and their intensities, one looks for the K nuclides and their concentrations. The form of this inverse problem is similar to the form of Eq. 14 (see Eq. 34 later).

4.1 Spectral Features

Consider a monoenergetic source that emits photons of energy E_1. A given photon that enters the detector may experience any of several outcomes, including the following:

1. It may be completely absorbed by photoelectric absorption, in which case the deposited energy is the photon energy, i.e., $E_d = E_1$.
2. It might undergo a sequence of one or more scatters within the detector and then be absorbed within the detector by photoelectric absorption, which again leads to $E_d = E_1$.
3. It might scatter one or more times in the detector and then escape with remaining energy E_r, which leads to a deposited energy of $E_d = E_1 - E_r$.
4. If $E_1 > 1.022$ MeV, it might undergo pair production in the detector, producing an electron–positron pair that usually leads to the production of two 0.511-MeV photons by positron annihilation (although positron annihilation occasionally takes place at higher than thermal energies and so photons of other energies are rarely produced). If both annihilation photons deposit all of their energy in the detector, then $E_d = E_1$. If one of the annihilation

photons escapes and the other is absorbed, the energy deposited is $E_d = E_1 - 0.511$ MeV. If both annihilation photons escape the detector, the deposited energy is $E_d = E_1 - 1.022$ MeV. If either or both annihilation photons scatter within the detector and then escape, an intermediate energy within the range $E_1 - 1.022\,\text{MeV} < E_d < E_1$ is deposited.

5. It might not interact at all in the detector and so $E_d = 0$.

Note that outcomes 1 and 2 both contribute to a "full-energy" peak. Voltage pulses result whose pulse heights are distributed from zero up to a maximum determined by the energy E_d and the energy resolution of the detector. Thus, a monoenergetic source produces a pulse-height spectrum that is distributed over the N channels. An example pulse-height spectrum obtained from a scintillation spectrometer exposed to the photons from a ^{22}Na source, which emits 1.28-MeV photons and 0.511-MeV annihilation photons, is shown in *Fig. 2*.

The features in the spectrum, from right to left, include a full-energy peak centered about the channel corresponding to $E_1 = 1.28$ MeV, a Compton edge and a continuum extending from about channel 1250 downward, a full-energy peak due to 0.511-MeV annihilation photons, a Compton edge and Compton continuum for the annihilation photons, and a backscatter peak. The maximum amount of energy that a photon can give up in a Compton scatter occurs in a "backscatter" through an angle of π rad, and so the Compton edge occurs over those channels that represent Compton scatters in the detector through angles near π. Photons that are emitted by the source and backscatter (within either the source or the material behind the source) have energies that are near the energy given by Compton scatter through π rad. The backscatter peak

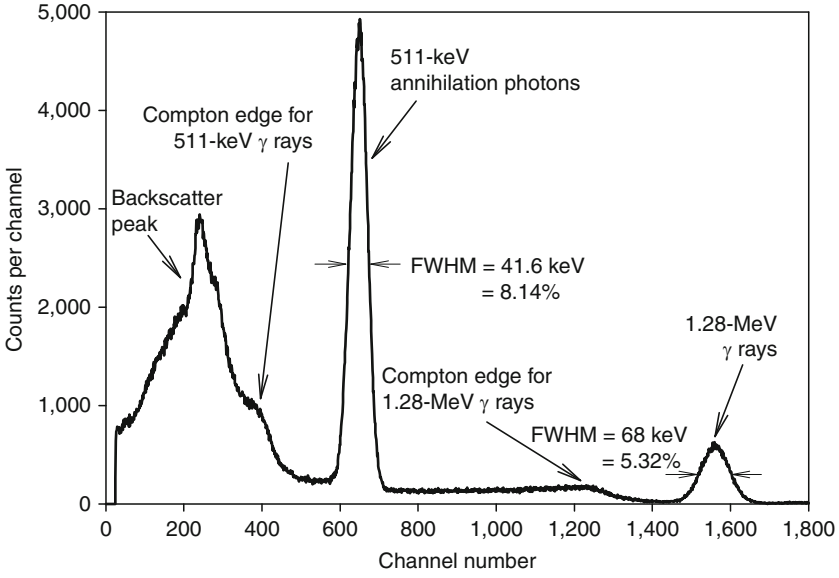


Fig. 2

The pulse-height spectrum obtained by a NaI(Tl) scintillation detector exposed to a ^{22}Na source. The source emits photons at 0.511 MeV, as a result of positron annihilations, and at 1.28 MeV, emitted as the product ^{22}Ne transitions from the excited state to the ground state (McGregor 2008)

in the spectrum of Fig. 2 is caused by source gamma rays that scatter, in or near the sample, through angles close to π and then deposit full energy in the detector. This peak typically has a FWHM that is larger than the FWHM of a full-energy peak for monoenergetic photons. The larger FWHM occurs because source photons can backscatter in or near the source over a range of angles near π rad, and thus, these scattered photons that reach the detector do not all have the same energy.

4.2 Detector Response Function

A spectrum such as that in Fig. 2, normalized to unit concentration, is called the *detector response function* for a given radionuclide. Each nuclide has a characteristic detector response function for each spectrometer for a specified source–detector geometry. Let the subscript k refer to a specific nuclide. Then the detector response function u_{nk} is the expected response (number of counts) of the spectrometer in channel n per unit concentration of nuclide k.

Detector response functions also can be associated with monoenergetic photons. If a source of monoenergetic photons of energy E_j, in some specific source–detector geometry, irradiates a detector, then the detector response function u_{nj} gives the expected response in channel n per source particle of energy E_j emitted from the source. In either case, it is not only the full-energy peaks that are of interest; the entire spectrum contains information about the source. Ways to exploit this fact are considered later.

5 Qualitative Analysis

For some purposes, it is necessary only to identify whether or not a sample emits photons at certain discrete energies E_j. This may be the case, for instance, if one wants to know if a sample contains a particular radionuclide. Alternatively, a procedure such as EDXRF, NAA, or PGNAA can be used to excite characteristic photons from a sample under investigation. If the element of interest is present, the characteristic photons emitted from the sample should create full-energy peaks in a pulse-height spectrum collected from the sample.

Photons of energy E_j that are emitted by the source produce full-energy pulses whose magnitudes are distributed about a mean value of

$$h_j = f(E_j). \tag{16}$$

A pulse of magnitude h_j produces a count in discrete channel n_d if

$$h_{\min} + (n_d - 1)\Delta < h_j \leq h_{\min} + n_d\Delta.$$

For purposes of energy determination, it is useful to consider the continuous channel number n_j corresponding to the pulse height h_j of the full-energy peak. Then the gamma-ray energy can be estimated by determining the generalized channel number n_j of the centroid of each peak.

It is then straightforward to obtain the corresponding h_j from

$$h_j = h_{\min} + n_j\Delta. \tag{17}$$

For the usual case, where $h_{\min} = 0$ and $h_{\max} = 10$ V, this reduces to

$$h_j = \frac{10}{N} n_j. \tag{18}$$

The expected source energy is then easily obtained from

$$E_j = f^{-1}(h_j). \tag{19}$$

For a linear detector whose response is given by Eq. 2, this is simply

$$E_j = \frac{h_j - \alpha}{\beta}. \tag{20}$$

If a nuclide emits several characteristic-energy photons, then one can gain confidence in the conclusion that the element is present if peaks occur at all energies for which the photon abundance and the detection efficiency would lead one to suspect that peaks should occur.

The simplest way to estimate the channel corresponding to the peak is by inspection of the plotted spectrum. This technique often is adequate. One merely estimates the channel number n_j that corresponds to the apparent centroid of the full-energy peak due to photons of energy E_j. For instance, in the spectrum shown in Fig. 3, one can obtain estimates of the channel numbers of the centroids of the two peaks shown. It should be apparent that this procedure is subjective (different researchers may estimate slightly different centroids) and thus has limited precision. Nevertheless, this procedure may suffice for some applications.

When estimation of the centroid by inspection is deemed insufficient, other methods must be employed. The wavelet transform has been used in various spectroscopic applications,

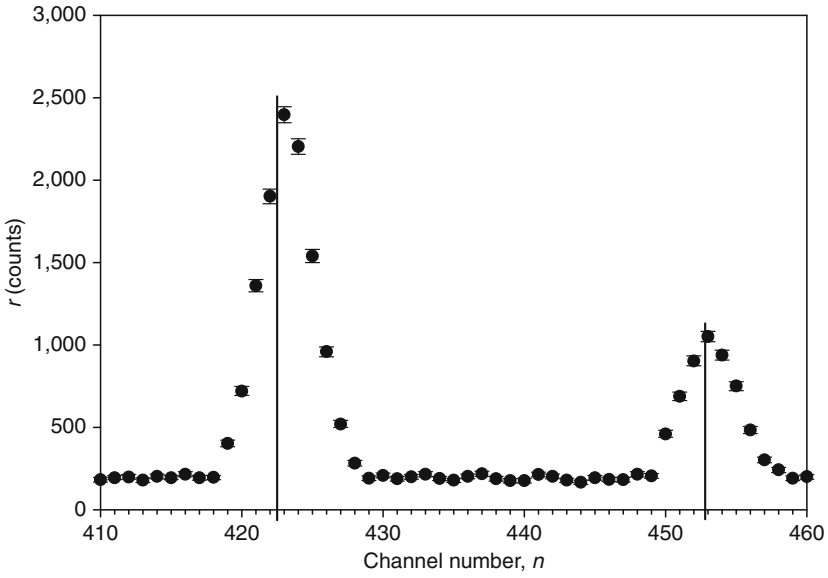


Fig. 3

An example of visual inspection to yield the channel numbers corresponding to the centroids of two peaks in a spectrum. The vertical lines are used to connect the apparent peaks to their centroid channel numbers

including nuclear magnetic resonance (Barache et al. 1997) and Raman spectroscopy (Xu et al. 1994). However, this approach is not commonly employed in gamma-ray spectroscopy and thus is not further considered here. Rather, it is common that the centroids of the peaks present are identified as part of a process that can also determine quantitative information about both the characteristic energies E_j and their relative abundances s_j (or $S_j = s_j t$). Alternatively, such methods may seek to identify the nuclide k and the nuclide concentration ξ_k. Such methods are discussed in the next section.

6 Quantitative Analysis

The model of Eq. 14 identifies a general inverse problem in which many channels contain information about the source distribution. One typically seeks to determine not only the characteristic energies E_j emitted by the source but also the individual source strengths s_j from the measured responses. Alternatively, one may use the spectral responses to identify the nuclide k and its concentration ξ_k in a sample. Quantitative analysis refers to the determination of quantities such as s_j and ξ_k. There are several approaches to quantitative analysis in spectroscopy, including the following:

1. Area under isolated peaks
2. Model fitting
3. Spectrum stripping
4. Library least squares
5. Symbolic Monte Carlo

Summaries of how these methods are typically implemented are given below. In general, one seeks to obtain both the E_j, $j = 1, 2, \ldots, J$, and the net areas under each of the J peaks. For spectra that are linearly related to the source strengths, the net area A_j under the jth full-energy peak is related to S_j, the number of gamma rays emitted by the source during the counting time t, by

$$S_j = \frac{A_j}{\eta_j},$$

where η_j is the detector efficiency, presumed known, at energy E_j,

$$A_j = \int_{n_1}^{n_2} (r_n - B_n)\, \mathrm{d}n, \tag{21}$$

and n_1 and n_2 are channel numbers that define the peak. Similarly, in linear systems, the concentration ξ_k of nuclide k is directly related to the net peak area.

6.1 Area Under the Peak

For isolated peaks, i.e., those that do not overlap with other peaks, a simple approach can be used to estimate the net peak area. Generally, the peak is superimposed on a generalized background, as shown, for an ideal case, in *Fig. 4*. To obtain the net area A, one identifies the channel

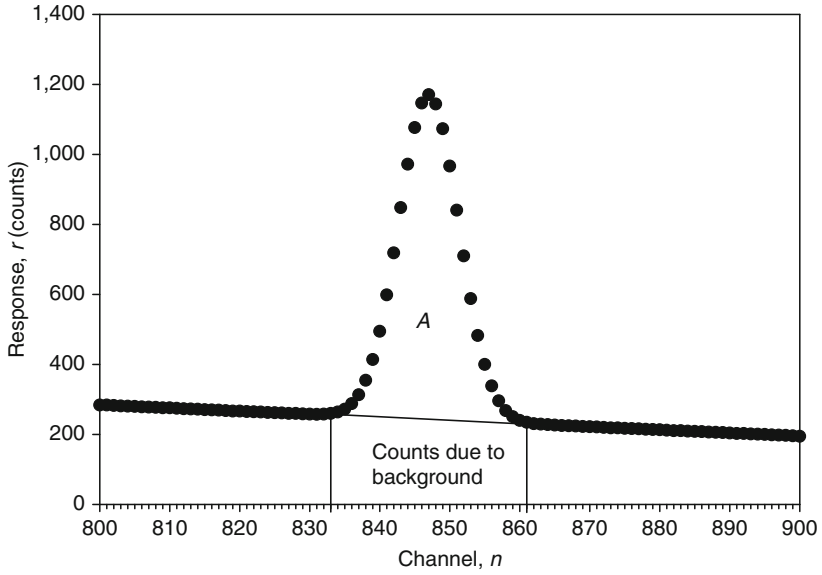


Fig. 4
The net area *A* under the peak but above background

numbers, n_1 and n_2, at which the peak disappears into the background. Then, the net peak area is estimated as

$$A = \sum_{n=n_1}^{n_2} r_n - (n_2 - n_1)\frac{r_{n_1} + r_{n_2}}{2}. \tag{22}$$

The net area is the total number of counts between channels n_1 and n_2 minus the area under a linear background between r_{n_1} and r_{n_2}. For instance, for the spectrum shown in *Fig. 4*, one might choose $n_1 = 833$ and $n_2 = 861$. The total area between these channels is the sum of the counts in each channel and the background is the area under the straight line connecting r_{833} and r_{861}; thus, the net area A is the area beneath the peak and above the background line in the figure. This simple procedure cannot be applied to overlapping peaks and gives only approximate net areas because the background may not be linear under the peak and identifying the channels n_1 and n_2 is subjective, especially when the standard deviations of the responses, $\sigma(r_{n_1})$ and $\sigma(r_{n_2})$, are large relative to the values r_{n_1} and r_{n_2}.

6.2 Model Fitting

A second technique, which can be used to obtain both the peak centroid and the area under the peak, is to fit a model to the data. Each peak is modeled as consisting of a peak function and a background function. In most gamma-ray spectroscopy applications, the peak function is assumed to be the product of a magnitude and a Gaussian PDF. Because the integral of the Gaussian PDF is unity, the magnitude is just the desired net peak area. The background can be assumed to have any functional form but most often is assumed to be a polynomial. As a rule of thumb, there is little extra work in treating the background function as quadratic

(or even cubic) as opposed to linear, and better results are generally obtained if this is done. If the best fit occurs for a linear background function, the coefficients of the higher-order terms are determined to be zero (or very small); if not, then appropriate coefficients for the higher-order terms assume appropriate values. Note that the background model is not physically based but is empirical. Nevertheless, this model-fitting approach works quite well, for both isolated and overlapping peaks.

6.2.1 Isolated Peaks

For purposes of illustration, consider a quadratic background function. Thus, for any apparently isolated peak, the model has the form

$$y(n) = Ag(n, \mu, \sigma) + a_0 + a_1 n + a_2 n^2, \tag{23}$$

where

$$g(n, \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-[(n-\mu)^2/(2\sigma^2)]} \tag{24}$$

is the Gaussian PDF centered at continuous channel μ and having standard deviation σ. The model of Eq. 23 generally is fit to the data only locally near each isolated peak in the spectrum. Thus, the values of a_0, a_1, and a_2 are generally different for different peaks. This model is fit to the data by a procedure such as least squares. If the energy resolution of the detector is well known as a function of energy, then the known value can be used for σ. It is also possible to consider σ as a model constant to be fit to the data. If the value obtained for σ is greater than the value expected (for the energy of the peak), then this might be an indication that the apparent isolated peak is actually composed of multiple overlapping peaks. In any event, consider $\mathbf{p} = \{p_k, k = 1, 2, \ldots, K\}$ to be the vector of model constants to be determined by fitting the model of Eq. 23 to the spectral data.

The least-squares approach proceeds as follows. For some range of channels (n_1, n_2) that encompasses the peak of interest, form the least-squares function

$$E^2(\mathbf{p}) = \sum_{n=n_1}^{n_2} \left[r_n - Ag(n, \mu, \sigma) - a_0 - a_1 n - a_2 n^2 \right]^2. \tag{25}$$

In order to assure that the entire peak is fit, it is suggested that the channel n_1 be at least 3σ below the centroid channel and n_2 be chosen to be at least 3σ above the centroid channel. This is because the area under a Gaussian over the range $\mu \pm 3\sigma$ is 0.997 of the total area under the Gaussian and, thus, the parameter A is a very good approximation to the total area under the Gaussian.

The parameters that best fit the model to the data minimize the least-squares function. The minima can be obtained by setting the partial derivatives to zero. Thus,

$$\frac{\partial}{\partial p_k} E^2(\mathbf{p}) = 0, \quad k = 1, 2, \ldots, K, \tag{26}$$

forms a system of K algebraic equations in the K unknown parameters. Note that the model is linear in most of the model constants and nonlinear in only two (μ and σ). Bevington (1969) and Press et al. (1996) discuss algorithms for solving systems of equations such as Eq. 26.

It is always useful to estimate not only the model parameters but also their standard deviations. The model constants for which the relative standard deviations are small should be

considered to be good estimates; a large relative standard deviation for a parameter indicates that the parameter value obtained is not known very precisely. Large relative standard deviations of one or more parameters put in doubt the validity of the assumed response model.

It generally is not possible to obtain exact error estimates. Rather, approximate methods are used. The least-squares procedure, in essence, uses a parabolic expansion of the E^2 function over the parameter space. Bevington (1969) states that the standard deviation in the retrieved value of parameter p_k can be estimated by

$$\sigma(p_k) \simeq \left[\sum_{n=1}^{N}\left(\frac{\partial p_k}{\partial r_n}\right)^2 \sigma^2(r_n)\right]^{1/2}, \tag{27}$$

where the $\sigma(r_n)$ are the standard deviations of the response values. If the spectrum consists of raw counts (e.g., not net counts obtained after subtracting background), then the standard deviations are well approximated by assuming the counts to be Poisson-distributed, for which $\sigma(r_n) = \sqrt{r_n}$. In other cases, appropriate methods should be used to estimate the standard deviations. The partial derivatives in Eq. 27 are generally easily estimated for the linear parameters. If the model is such that the partial derivative for the nonlinear parameters also can be easily estimated, then application of Eq. 27 is straightforward. When this is not the case, then the partial derivatives can be estimated numerically by difference formulas (see, e.g., Press et al. (1996) or Hornbeck (1975)).

Strictly speaking, this procedure, which requires initial estimates of the nonlinear parameters, is not guaranteed to lead to a global minimum. Thus, it is good practice to select a value of μ that is near the channel number (by inspection) corresponding to the centroid of the peak and a value of σ near the presumed energy resolution of the detector at the approximate channel corresponding to μ. Alternatively, one can estimate the FWHM Γ of the peak and set $\sigma = \Gamma/2.355$.

Of course, the values of μ and A are of most interest. Once these values are determined, μ is treated as the generalized channel number and the corresponding pulse height is estimated from either Eq. 17 or Eq. 18, as appropriate. Then E is found either from Eq. 19 or Eq. 20, as appropriate. Given A, one can estimate s, the rate at which the source emits photons of energy E, as

$$s = \frac{A}{\eta t}, \tag{28}$$

where, as before, t is the counting time and η is the detector efficiency at energy E.

6.2.2 Overlapping Peaks

Sometimes two or more peaks overlap. In cases where a peak is asymmetric or has a FWHM larger than expected, one can try to fit multiple peaks to the data locally. In such cases, employ a model of the form

$$y(n, \mathbf{p}) = \sum_{j=1}^{m} A_j g(n, \mu_j, \sigma_j) + a_0 + a_1 n + a_2 n^2, \tag{29}$$

where m is the number of peaks you suspect might be overlapped. As a practical matter, you might try $m = 2$ first and see if you obtain reasonable results. If not, try larger values of m. Of course, this model introduces new model constants for each additional peak. The fitting of multiple overlapping peaks, with asymmetric peak models, to XPS spectra is considered in detail by Dunn and Dunn (1982).

6.2.3 Weighted Least Squares

Another approach, perhaps better than standard least squares, is to use weighted least squares. In this approach, one fits a model such as that given by Eq. 29 to the data by forming the reduced chi-square function

$$\chi_\nu^2 = \frac{1}{\nu}\sum_{n=n_1}^{n_2}\frac{\left[r_n - \sum_{j=1}^{m} A_j g(n,\mu_j,\sigma_j) - a_0 - a_1 n - a_2 n^2\right]^2}{\sigma^2(r_n)}, \tag{30}$$

where $\nu = n_2 - n_1$ is the number of degrees of freedom and $\sigma^2(r_n)$ is the variance of the pulse-height response in channel n. Generally, the r_n are either gross counts in channel n or, for background-subtracted spectra, net counts in channel n. (Note that background subtraction only removes part of the generalized background, the $b_n t$ of Eq. 15.) If the r_n are gross counts, then one presumes that Poisson statistics apply and $\sigma^2(r_n) = r_n$. If the spectrum is background-subtracted, then $\sigma^2(r_n) = r_n + b_n$. If the r_n result from some other process, then one should use the appropriate variances in Eq. 30. In any case, one proceeds as in standard least squares by setting the partial derivatives of χ_ν^2 with respect to each of the model parameters equal to zero and solving the resulting set of equations.

An advantage of this approach is that the better the model fits the data, the closer the value of χ_ν^2 should be to unity. This is because if the model is an accurate description of the data, then the square of the deviations between the data and the model in the numerator of Eq. 30 should, on average, be about as large as the natural variance in the data in the denominator of Eq. 30.

This procedure is illustrated in the following example. Suppose you are given the spectral data in *Table 1*. These data suggest that two peaks overlap. Thus, you might fit a model of the form

$$y(n,\mathbf{p}) = A_1 g(n,\mu_1,\sigma_1) + A_2 g(n,\mu_2,\sigma_2) + a_0 + a_1 n + a_2 n^2 \tag{31}$$

to the data.

To implement the χ_ν^2 fitting procedure, it is customary to list the nonlinear parameters first and the linear parameters next. Thus, in this case, one might choose $p_1 = \mu_1, p_2 = \sigma_1, p_3 =$

Table 1
A portion of a spectrum giving channels and gross counts

n	r_n	n	r_n	n	r_n	n	r_n
117	2,210	127	4,756	137	45,568	147	3,827
118	2,253	128	6,176	138	36,698	148	3,416
119	2,333	129	9,761	139	23,773	149	3,156
120	2,487	130	17,016	140	14,669	150	2,896
121	2,763	131	26,462	141	10,054	151	2,713
122	2,869	132	30,846	142	7,666	152	2,448
123	2,984	133	28,392	143	6,284	153	2,335
124	3,312	134	26,822	144	5,387	154	2,119
125	3,629	135	32,667	145	4,792	155	1,957
126	4,077	136	42,618	146	4,224	156	1,754

$\mu_2, p_4 = \sigma_2, p_5 = A_1, p_6 = A_2, p_7 = a_0, p_8 = a_1$, and $p_9 = a_2$. Then, a system of equations is formed by setting partial derivatives of the χ^2_ν function to zero, i.e.,

$$-\frac{\partial}{\partial p_k}\chi^2_\nu(\mathbf{p}) = \frac{2}{\nu}\sum_{n=n_1}^{n_2}\frac{[r_n - y(n,\mathbf{p})]}{\sigma^2(r_n)}\frac{\partial y(n,\mathbf{p})}{\partial p_k} = 0, \quad k = 1, 2, \ldots, K, \tag{32}$$

where, to be explicit,

$$\begin{aligned}
\frac{\partial y(n,\mathbf{p})}{\partial \mu_j} &= A_j\frac{n-\mu_j}{\sigma_j^2}g(n,\mu_j,\sigma_j), \quad j = 1,2,\\
\frac{\partial y(n,\mathbf{p})}{\partial \sigma_j} &= \frac{A_j}{\sigma_j}\left[\frac{(n-\mu_j)^2}{\sigma_j^2} - 1\right]g(n,\mu_j,\sigma_j), \quad j = 1,2,\\
\frac{\partial y(n,\mathbf{p})}{\partial A_j} &= g(n,\mu_j,\sigma_j), \quad j = 1,2,\\
\frac{\partial y(n,\mathbf{p})}{\partial a_0} &= 1.0,\\
\frac{\partial}{\partial a_1}y(n,\mathbf{p}) &= n,\\
\frac{\partial}{\partial a_2}y(n,\mathbf{p}) &= n^2.
\end{aligned} \tag{33}$$

With methods discussed by Bevington (1969), it is possible to perform an iterative search on the four nonlinear parameters while using matrix inversion to determine the linear parameters. The procedure proceeds as follows:

- Specify initial estimates for the nonlinear parameters.
- For these specified values, perform a matrix inversion to obtain the best-fit linear parameters.
- Use a method to update the nonlinear parameter values. Methods that depend on the local partial derivative of the model for each parameter are better than methods that proceed blindly, but any method, properly implemented, suffices for the task.
- For the new values of the nonlinear parameters, perform another matrix inversion to obtain the linear parameters.
- Proceed in an iterative manner until the minimum of the reduced χ^2_ν function is found to within some convergence criterion.

It is noted that while searching through parameter space for the values of the parameters that best fit the model to the data, it is possible for unrealistic parameter values to occur. Thus, it is preferable to introduce constraints on the parameter values. For instance, negative values of means and standard deviations are not physically meaningful, although they may be mathematically allowable.

Solutions of the system of equations given by Eq. 33 for the data of *Table 1*, produced the results plotted in *Fig. 5*. The best-fit parameters were as follows:

$$\mu_1 = 131.745 \pm 0.013,$$

$$\sigma_1 = 1.493 \pm 0.010,$$

$$A_1 = 92{,}012.6 \pm 871.8,$$

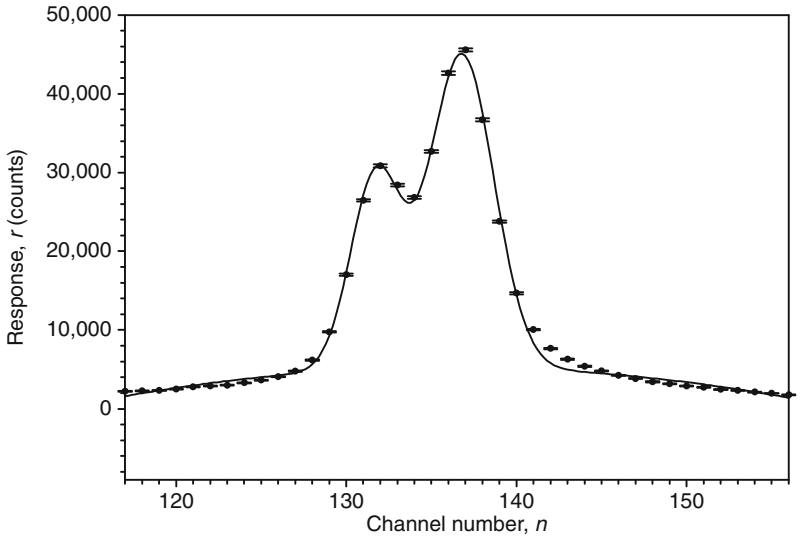


Fig. 5
A fit of two overlapping peaks with a quadratic background

$$\mu_2 = 136.786 \pm 0.010,$$

$$\sigma_2 = 1.891 \pm 0.009,$$

$$A_2 = 188{,}560 \pm 1264,$$

$$a_0 = -177{,}406 \pm 138,$$

$$a_1 = 2681.9 \pm 1.9,$$

$$a_2 = -9.8445 \pm 0.0074.$$

It is seen that the model fits the data quite well. One would expect that a good model would be within the error intervals for about 68% of the data points and this seems to be the case here. The uncertainty in each of the model parameters is small, relative to the parameter value, which also indicates a good fit. It is noted that the values of $n_1 = 117$ and $n_2 = 156$ are well beyond the 3σ range for the σ_1 and σ_2 obtained. Thus, the A_1 and A_2 values should be good estimates of the net areas under the two peaks.

6.3 Spectrum Stripping

If response functions can be collected or generated for all the sources (or nuclides) that are expected to be present in an unknown sample, and if the dependence of the response model is linearly related to the source strengths or nuclide concentrations, then the method of spectrum stripping can be applied. This procedure involves the following:

- Collect a spectrum from an unknown sample.
- Identify the highest-energy peak in the spectrum.

- Subtract the response function for that peak, weighted by a constant, such that the peak is effectively removed.
- Proceed down the spectrum while subtracting other weighted response functions until all peaks are removed.

If the residuals are randomly distributed about zero or about some smooth background, then the specific response functions stripped from the spectrum for the unknown identify the nuclides present and the weighting constants estimate the source strength or concentration of each nuclide.

6.4 Library Least Squares

Because a detector produces a spectrum, even for a monoenergetic input, one can try to utilize the entire spectral response, or at least a significant part of it, rather than just the response values near each peak or set of overlapping peaks. The library least-squares approach, originally introduced by Marshall and Zumberge (1989), asks the following question. Why focus only on the peaks? Basically, the library least-squares (LLS) approach, as implemented, for instance, by Gardner and Sood (2004) and Gardner and Xu (2009), involves storing into a library the detector response functions of all the candidate nuclides that might be present in the sample being analyzed. Then some fitting technique, such as least squares or weighted least squares, is used to fit the library spectra to the data.

This approach was not possible many years ago because one could not experimentally measure good spectra from all candidate nuclides. However, Monte Carlo modeling has become sufficiently robust that detector response functions can be calculated for essentially any nuclide. When Monte Carlo is used to generate detector response functions, the method often is referred to as Monte Carlo library least squares or MCLLS. In this method, one calculates and stores the detector response functions, for any nuclides suspected to be present in a sample, in a library and proceeds as follows.

Let u_{nk} be the response expected in channel n due to a unit concentration of nuclide k. For an unknown sample that may contain K nuclides, write the response in channel n in the form

$$r_n = \sum_{k=1}^{K} \xi_k u_{nk} + e_n, \tag{34}$$

where e_n is the stochastic error in channel n. Then,

$$e_n = r_n - \sum_{k=1}^{K} \xi_k u_{nk}$$

and one can fit $M \leq N$ channels of the spectrum in a weighted least-squares manner, namely

$$\chi_\nu^2 = \frac{1}{\nu} \sum_{n=1}^{M} W_n e_n^2, \tag{35}$$

where W_n is a weight factor, often taken as $W_n = 1/\sigma^2(r_n)$, and $\nu = M - K$ is the number of degrees of freedom. The ξ_k identify the concentrations of the nuclides present. Note that if ξ_k is negative or near zero for one or more k, then the corresponding nuclides might not be present

in the sample. In this case, it is a good practice to remove the detector response functions for these nuclides and repeat the analysis to see if a good fit is obtained.

The process just described assumes the model is linear in the nuclide concentrations and thus can be referred to as the *linear LLS* approach. This approach is very similar to application of the weighted least-squares process in the model-fitting approach, but attempts to use more features of the spectrum than just the full-energy peaks. There are instances, however, where the model is not linear in the nuclide concentrations. One such case, for PGNAA, is considered in the next subsection and another, for EDXRF, is considered in Sect. 6.5.

6.4.1 Nonlinear Spectra

Generally, in NAA, the samples are small, the mass of the sample can be accurately measured, and one is generally looking for trace elements. Thus, NAA analysis is well approximated as a linear process and application of linear LLS is appropriate. However, in bulk samples, the PGNAA response is nonlinear, primarily for the following reasons:

- Sample mass, which often is not known, affects the flux density within the sample.
- The composition of the sample, which is unknown in advance, affects the spectrum and the macroscopic capture cross section of the sample. In particular, moisture content strongly affects the thermal neutron flux density, which is what gives rise to the prompt gamma rays. Also, neutron absorbers affect the thermal flux density.

Thus, nonlinear models are needed for PGNAA.

The general MCLLS approach in the nonlinear case proceeds as follows.

1. Assume values for the concentrations and use Monte Carlo to generate a complete spectrum for a sample of this assumed composition.
2. Keep track of the individual spectral responses for each element within the Monte Carlo code, so as to provide library spectra u_{nk} for each nuclide.
3. Use linear LLS to estimate the nuclide concentrations $\xi_k, k = 1, 2, \ldots, K$, from the sample spectrum.
4. If the calculated $\xi_k, k = 1, 2, \ldots, K$, match the assumed composition closely enough, you are done. If not, pick a new composition, based on the calculated concentrations, and repeat the process.
5. Iterate until you converge to the actual composition, to within a desired tolerance.

Results of this general procedure are given, for instance, by Gardner and Xu (2009).

6.4.2 Compton Suppression

Because prompt gamma rays tend to be of high energy (most are 1 MeV $< E <$ 12 MeV); there is a large Compton component to the spectra, especially for thin semiconductor detectors. There are ways to reduce the Compton continuum. One is to partially surround a high-resolution germanium detector with scintillators, such as BGO, which has high efficiency because of its high density. The basic idea is that a photon that Compton scatters in the germanium detector has a

reasonable chance of interacting in the scintillation crystals. These interactions occur at essentially the same time as the Compton scatter in the germanium and can thus be suppressed by an anti-coincidence gate (the gate only accepts pulses in the germanium that are anti-coincident with the scintillators). Fairly dramatic results can be achieved (Molnar 2004).

6.5 Symbolic Monte Carlo

In X-ray fluorescence, the responses are due to the *elements* present, but each element is composed of nuclides and the convention was introduced earlier to refer only to nuclides. Hence in the discussion below, elemental concentrations will be called nuclide concentrations. Nonlinear matrix effects lead to absorption and enhancement in EDXRF. For instance, the characteristic X rays of nuclide *a* can be absorbed by elements with lower atomic number, reducing the signal from nuclide *a* and enhancing the signals for the lower atomic number nuclides. This means that the models in EDXRF also are not linear in the nuclide concentrations. Another implementation of Monte Carlo has been used in the EDXRF case. The method, originally called Inverse Monte Carlo (IMC) (Dunn 1981), was applied to EDXRF by Yacout and Dunn (1987) for primary and secondary X-rays. Mickael (1991) extended the work to account for tertiary fluorescence.

The term IMC has been used for other purposes, e.g., to mean solving inverse problems by iterative Monte Carlo simulations as the unknown parameters are varied until simulated and measured results agree sufficiently. The acronym IMC also has been used for "implicit Monte Carlo" (Gentile 2001). Thus, Dunn and Shultis (2009) recently proposed renaming the version of IMC that is noniterative in the Monte Carlo simulations *symbolic Monte Carlo* (SMC) because the method proceeds by using symbols in the Monte Carlo scores for the unknown parameters.

Symbolic Monte Carlo is a specialized technique in which the inverse problem of estimating the k and ξ_k is solved by a system of algebraic equations generated by a single Monte Carlo simulation. For purposes of illustration, a ternary system (one that contains three elemental nuclides) is considered. In essence, SMC creates models, with symbols for the unknown concentrations ξ_k, for the areas under all of the various X-ray peaks (e.g., K_α and K_β) in a single simulation. The models depend on the detector efficiency as a function of energy; the nuclide concentrations; the primary, secondary, and tertiary fluorescence produced in the sample; and the background. For a ternary system, three equations result. The equations are rather complex (see Yacout and Dunn (1987) and Mickael (1991)) but can be developed using only a single Monte Carlo simulation. The advantage of this approach is that there is no need to iteratively run full Monte Carlo simulations as the assumed concentrations are varied. The disadvantage is that development of the model is involved and the algebraic equations are quite complex. Nevertheless, the method has been shown to work well in X-ray spectroscopy and can, in principle, be applied to other spectroscopy applications, such as PGNAA, where the responses are nonlinear functions of the concentrations.

7 Detectors for Gamma-Ray Spectroscopy

Typical methods of gamma-ray and X-ray spectroscopy can be categorized as either energy dispersive or wavelength dispersive. Energy-dispersive spectroscopy (EDS) is perhaps the more popular technique, in which the energy of photons is preserved and recorded.

Energy-deposition indicators in the detector include light emission, electrical current, and thermal changes. Wavelength-dispersive spectroscopy (WDS) is a technique in which a specific photon wavelength is measured by a witness detector. The witness-detector position is translated in space so as to accumulate a spectrum of different photon wavelengths.

There are numerous detectors that can be used for radiation spectroscopy, which include versions of gas-filled, scintillation, and semiconductor detectors. However, only those commercial devices commonly used as analytical tools for gamma-ray and X-ray spectroscopy are described in the present section. EDS devices include scintillation, semiconducting, and cryogenic spectrometers. Diffractive spectrometers are classified as WDS devices, and typically use gas-filled proportional counters as witness detectors.

7.1 Scintillation Spectrometers

Scintillators are generally separated into two classes, those being inorganic and organic. The method by which either produces scintillation light is physically different, hence the distinction. Inorganic scintillators can be found as crystalline, polycrystalline, or microcrystalline materials. Organic scintillators come in many forms, including crystalline materials, plastics, liquids, and even gases. However, organic scintillators, being composed of low-Z materials, are ineffective as gamma-ray spectrometers, and, thus, are not covered here.

The scintillation principle is quite simple. Radiation interactions occurring in a scintillator cause either the atomic or molecular structure in the scintillator to become excited such that electrons are increased in energy. These excited electrons will de-excite, some of which will radiate light energy. The light emissions can then be detected with light-sensitive instrumentation.

A typical scintillation spectrometer consists of a scintillating material hermetically sealed in a reflecting cannister. Typical cannisters are cylindrical, with one end of the cylinder being an optically transparent window with all remaining surfaces being Lambertian reflectors. The optically transparent window is coupled to a light collection device, such as a photomultiplier tube (PMT), with an optical compound. The optical compound helps match the indices of refraction between the scintillation cannister and the light collection device so as to reduce reflective losses. The PMT provides a voltage output that is linear with respect to the light emitted from the scintillator. Hence, the voltage "spectrum" is a linear indication of the radiation energy spectrum deposited in the detector. It is typical for commercial vendors to provide the scintillation cannister and the PMT as one complete unit, although they can be acquired separately.

7.1.1 Inorganic Scintillators

Inorganic scintillators depend primarily on the crystalline energy band structure of the material for the scintillation mechanism. Referring to ❯ *Fig. 6*, there is shown an energy band diagram typical of an inorganic scintillator. A lower energy band, referred to as the valence energy band, has a reservoir of electrons. It is this band of electrons that participates in the binding of atoms. The next higher band is commonly referred to as the conduction band, which for inorganic scintillators is usually devoid of electrons. Between the two bands is a forbidden region where electrons are not allowed to exist, typically referred to as the energy band gap.

If a radioactive energy quantum, such as a gamma ray or charged particle, interacts in the scintillation material, it can excite numerous electrons from the valence band and the tightly

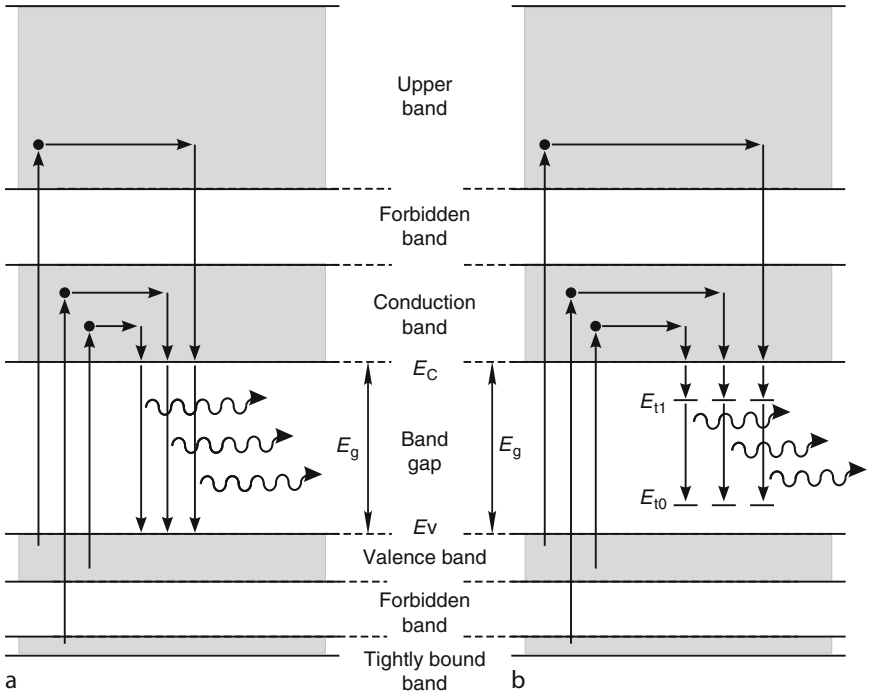


Fig. 6

Shown are two basic methods by which an inorganic scintillator produces light. The intrinsic case (a) has no added activator dopants. Absorbed radiation energy excites electrons from the valence and tightly bound bands up into the higher conduction bands. These electrons quickly de-excite to the lowest conduction-band edge, E_C. As they drop back to the valence band, they release light photons. Some intrinsic scintillators emit light through optical transitions from ionized elemental constituents, which can be of lower energy than the band gap. The extrinsic case (b) has activator dopants that produce energy levels in the band gap. Absorbed radiation energy excites electrons from the valence and tightly bound bands up into the higher conduction bands. These electrons quickly de-excite to the lowest conduction-band edge, E_C, as before. However, many drop into the upper-activator-site energy state, E_{t1}. As they then drop to the activator ground state E_{t0}, they release light photons of lower energy than the scintillator band gap (McGregor 2008)

bound bands up into the conduction bands (see *Fig. 6a*). These electrons rapidly lose energy and fall to the conduction-band edge, E_C. As they de-excite and drop back into the valence band, they can lose energy through light emissions. Unfortunately, because the radiated energy of the photons is equivelent to the band-gap energy, these same photons can be reabsorbed in the scintillator and again excite electrons into the conduction band. Hence, the scintillator can be opaque to it own light emissions. There are exceptions in which intrinsic scintillators work well. For example, bismuth germanate (BGO) releases light through optical transitions of Bi^{3+} ions, which release light that is lower in energy than the band gap, hence is relatively transparent to its own light emissions.

However, if an *impurity* or *dopant* is added to the crystal, it can produce allowed states in the band gap, as depicted in ❯ *Fig. 6b*. Such a scintillator is referred to as being *activated*. In the best of cases, the impurity atoms are uniformly distributed throughout the scintillator. When electrons are excited by a radiation event, they migrate through the crystal, many of which drop into the excited state of the impurity atom. Upon de-excitation, an electron will yield a photon equal in energy to the difference between the impurity atom excited and ground states. Hence, it will most likely not be reabsorbed by the scintillator material. Careful selection of the proper impurity *dopant* can allow for the light-emission wavelength to be tailored specifically to match the sensitivity of the light collection device.

NaI(Tl) Scintillation Detectors

The most used inorganic scintillator current with the writing of this handbook is NaI(Tl), the nomenclature meaning that the scintillator is the salt NaI that has been activated with the dopant Tl. NaI(Tl) yields approximately 38,000 photons per MeV of energy absorbed in the crystal. Light is emitted from NaI(Tl) in a continous spectrum, yet the most probable emission is at 415 nm, which matches well to typical commercial photomultiplier tubes. The decay time of NaI(Tl) is 230 ns, which refers to the time required to release 63.2% of the scintillation photons. It is the availability of large sizes and the acceptably linear response to gamma rays that makes NaI(Tl) important. Many different sizes are available, ranging in size from cylinders that are only 0.5 inch in diameter to almost a meter in diameter. Yet, the most preferred geometry remains the 3 inch × 3 inch (7.62 cm × 7.62 cm) right circular cylinder. It is the most characterized NaI(Tl) detector size with extensive efficiency data in the literature. Further, it is the standard by which all other inorganic scintillators are measured.

Because of its high efficiency for electromagnetic radiation (see ❯ *Fig. 7*), NaI(Tl) is widely used to measure X-rays and gamma rays. X-ray detectors with a thin entrance window

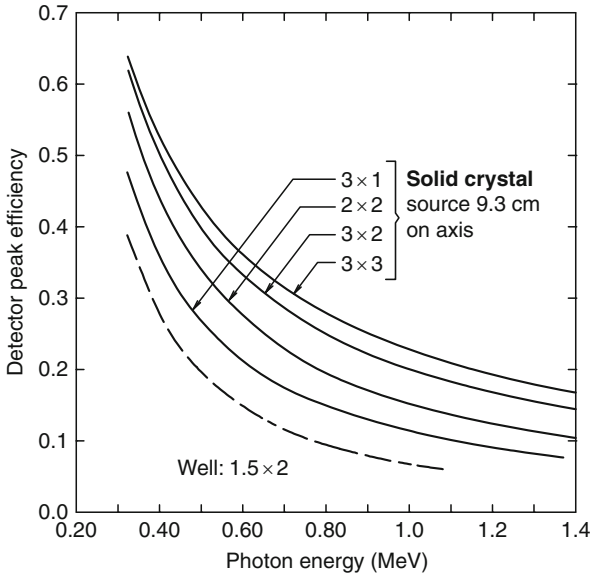


Fig. 7
Intrinsic peak efficiency for NaI(Tl) detectors (McGregor 2008)

containing a very thin NaI(Tl) detector are often used to measure the intensity and/or spectrum of low-energy electromagnetic radiation. NaI(Tl) detectors do not require cooling during operation and can be used in a great variety of applications. The bare NaI(Tl) crystal is hygroscopic and fragile. However, when properly packaged, field applications are possible because they can operate over a long time period in warm and humid environments, resist a reasonable level of mechanical shock, and are resistant to radiation damage. Basically, for any application requiring a detector with a high gamma-ray efficiency and a modest energy resolution, the NaI(Tl) detector is clearly a good choice. A gamma-ray spectrum from a 2 × 2 right circular cylindrical NaI(Tl) detector is shown in *Fig. 2* (see Sect. 4.1).

Other Inorganic Scintillation Detectors

Since the discovery of NaI(Tl) in 1948, the search has continued for a better scintillator for higher-energy-resolution gamma-ray spectroscopy. There has been some limited success, which includes those scintillators listed in *Table 2*. For instance, CsI(Na) is similar in performance to NaI(Tl), but has a longer decay time. CsI(Tl) has much higher light output than NaI(Tl), but the emission spectrum maximizes at 550 nm, which does not couple well to PMTs. However, CsI(Tl) has been coupled to Si-photodiode sensors quite successfully. Bismuth germanate (BGO) has lower light output, but is much denser and a better absorber of gamma rays. As a result, BGO is used for medical imaging systems, which helps to reduce the overall radiation dose that a patient receives during the imaging procedure. LiI(Eu) is a scintillator that is primarily used for neutron detection, relying upon the ^{6}Li content in the crystal. In recent years, $LaBr_3$(Ce), a new scintillator with exceptional properties for gamma-ray spectroscopy, has become available, demonstrating lower than 3% FWHM for 662 keV gamma rays. $LaBr_3$(Ce) has much higher light yield and a much shorter decay constant than NaI(Tl). Further, it is composed of higher-Z elements, hence is a better gamma-ray absorber than NaI(Tl). However, it is extremely hygroscopic and fragile, hence is difficult to produce and handle. Although it has recently

Table 2

Common inorganic scintillator materials and properties

Scintillator	Wavelength of maximum emission (nm)	Decay time (ns)	Light yield in photons/MeV	Relative PMT response compared to NaI(Tl)
NaI(Tl)	415	230	38,000	1.00
CsI(Na)	420	680, 3340	39,000	1.10
CsI(Tl)	550	460, 4180	65,000	0.49
LiI(Eu)	470	1400	11,000	0.23
BGO	480	300	8,200	0.13
CaF_2(Eu)	435	900	24,000	0.50
GSO(Ce)	440	56, 400	9,000	0.20
YAP(Ce)	370	27	18,000	0.45
YAG(Ce)	550	88, 302	17,000	0.50
LSO(Ce)	420	47	25,000	0.75
$LaCl_3$(Ce)	350	28	49,000	0.70–0.90
$LaBr_3$(Ce)	380	16	63,000	1.30

become commercially available, it is presently 40 times more expensive than NaI(Tl) due to production and fabrication problems. Overall, there are numerous inorganic scintillators available for special radiation-detection purposes.

7.1.2 Light Collection

The light produced from a scintillation detector is collected by a photo-sensitive device, such as a photomultiplier tube, a photodiode, or a microchannel plate. The light is then converted to a measureable voltage pulse.

Photomultiplier Tubes

The photomultiplier tube (PMT) is commonly used to measure the light output from a scintillation detector. Referring to *Fig. 8*, the basic PMT has a photocathode which is located so as to absorb light emissions from a light source such as a scintillating material. When light photons strike the coating on the photocathode, they excite electrons which can diffuse to the surface facing the vacuum of the tube. A fraction of these excited electrons will escape the surface and leap into the vacuum tube. A voltage applied to the tube will guide the liberated electrons to an adjacent electrode named a dynode. As an electron approaches the dynode, it gains velocity and energy from the electric field. Hence, when it strikes the dynode, it will again cause more electrons to become liberated into the tube. These newly liberated electrons are then guided to the next dynode where more electrons are liberated and so on. As a result, the total number of

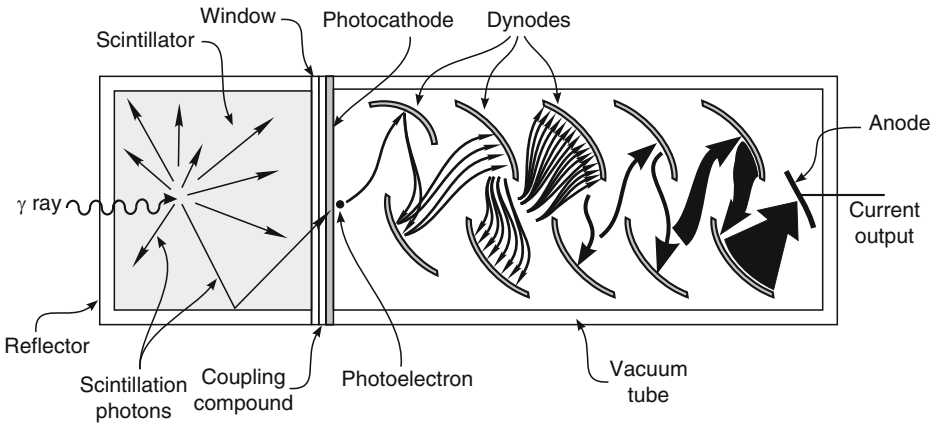


Fig. 8

The basic mechanism of a photomultiplier tube (PMT). An absorbed γ ray causes the emission of numerous light photons which can strike the photocathode. A scintillation photon that strikes the photocathode excites a photoelectron. The photoelectron is accelerated and guided to the first dynode with an electric field, where it strikes the dynode and ejects more electrons. These electrons are accelerated to the next dynode and excite more electrons. The process continues through the dynode chain until the cascade of electrons is collected at the anode, which is used to produce a voltage pulse (McGregor 2008)

electrons released is a function of the number of dynodes in the PMT and the photoefficiency of the photocathode and the dynodes.

The total charge released in the PMT is

$$Q = qN_0G^n, \tag{36}$$

where q is the charge of an electron, N_0 is the initial number of electrons released at the photocathode, G is the number of electrons released per dynode per electron (the gain), and n is the total number of dynodes in the PMT. For instance, suppose that a PMT has ten dynodes each operated with a gain of four. An event that initially releases 1,000 electrons (N_0) will cause over 10^9 electrons to emerge from the PMT.

The photomultiplier tube is an important tool in radiation detection, as it is the device that allowed scintillation materials to be used as practical detectors. It can take a minute amount of light produced in a scintillator from a single radiation absorption and turn it into a large electrical signal. It is this electrical signal, typically converted to a voltage pulse, that is measured from the scintillation detector system.

PMTs are stable and electronically quiet (low noise). Modern varieties have exceptional photocathode and dynode efficiencies, often referred to as quantum efficiency, with gains that can exceed 30. However, there is a drawback. PMT materials used as photocathodes are generally fabricated from alkaline metals, which are most sensitive to light in the 350–450 nm range. Scintillators emitting light outside of these boundaries can still be used under some circumstances, yet their effectiveness can be severely compromised.

Microchannel Plates

Microchannel plates are an alternative method of amplifying signals from a scintillator. Microchannel plates are glass tubes with the insides coated with secondary-electron-emissive materials. A voltage is applied across the tube length which causes electrons to cascade down the tube. Every time an electron strikes the tube wall, more electrons are emitted, much like with dynodes in a PMT. Hence, a single electron can cause a cascade that can eventually produce 10^6 electrons emitted from the other end of the tube. Typically, hundreds of these microchannels are bonded together to form a plate of channels running in parallel. The microchannel plate can be fastened to a common scintillator, whether organic or inorganic, which operates in a similar fashion as a PMT. Light photons entering the microchannel plate cause the ejection of primary photoelectrons, which cascade down the microchannels to liberate millions more electrons. The main advantage of a microchannel plate is its compact size, in which a microchannel plate only one inch thick can produce a signal of similar strength as a common PMT. The main problem with microchannel plates is that the signal produced per monoenergetic radiation event is statistically much noisier than that produced by a PMT; hence, the energy resolution for spectroscopy is typically worse than provided by a PMT.

Photodiodes

Photodiodes are actually semiconductor devices formed into a *pn*- or *pin*-junction diode. When photons strike the semiconductor, usually Si- or GaAs-based materials, electrons are excited. A voltage bias across the diode causes the electrons to drift across the device and induce charge much like a gas-filled ion chamber. The quantum efficiency of the semiconductor diode varies with device configuration and packaging. For instance, various different commercial Si photodiodes have peak efficiencies at wavelengths ranging between 700 and 1000 nm. Regardless, they are typically more sensitive to longer wavelengths than common

commercial PMTs. As a result, CsI(Tl) emissions match better to Si photodiodes than PMTs. Photodiodes operate with low voltage, are small, rugged, and relatively inexpensive, hence offer a compact method of sensing light emissions from scintillators. However, they typically do not couple well to light emissions near the 400 nm range (blue-green) and have low gain, if any at all. Consequently, the signals from photodiodes need more amplification than signals from PMTs, and scintillator/photodiode systems generally produce worse energy resolution than do scintillator/PMT systems.

7.1.3 Factors Affecting Energy Resolution

The energy resolution achievable from a scintillation spectrometer is determined by a number of factors, including brightness, reflector efficiency, light collection efficiency, PMT quantum efficiency, activator uniformity, and the scintillator-light-yield linearity. Simplistically, the energy resolution can be described as

$$\mathcal{R}^2 = R_s^2 + R_p^2, \tag{37}$$

where R_s is the energy-resolution contribution relating to the scintillation crystal and encapsulation and R_p is the energy-resolution contribution relating to the photon detection device. The term R_s can be described by

$$R_s^2 = R_n^2 + R_i^2 + R_t^2 + R_d^2, \tag{38}$$

where R_n is resolution broadening from statistical fluctuations in the number of electrons excited by monoenergetic absorption events, R_t is resolution broadening from variations in light transfer to the light detection device, R_i is resolution broadening from inhomogeneity, and R_d is the nonproportionality of response. The variance R_n^2 can be expected to follow a somewhat Gaussian distribution, corrected with the appropriate Fano factor to correct for the energy band structure. The distribution of dopants in the scintillator may not be homogeneously distributed in the material, causing spatially dependent variations R_i^2 in the number of electrons diffusing to the photon-producing activator sites. Although reflectors around scintillator detectors can be quite efficient, the angle at which photons strike the light collector (PMT, for instance) may cause variations R_t^2 in the number of photons that actually enter the light detector. Other factors affecting the light transfer variation include nonuniform clarity of the scintillator, imperfections in the reflector, and nonuniformity of the coupling compound between the scintillator and the light detection device. Most scintillators have a nonlinear light yield as a function of absorbed photon energy, which is especially true in the energy range between a few keV up to 1 MeV. As a result, there will be a variation R_d^2 in the number of photons produced for monoenergetic gamma rays, depending on if the photon was absorbed through a single photoelectric event or numerous Compton scatters.

The light collection terminates at the photo-detection device, typically a PMT, in which the light is converted into an electron signal. The conversion efficiency and variance R_p^2 of the light collection efficiency is affected by the wavelength of light striking the device and uniformity of the collecting layers. For instance, scintillators emit a spectrum of photon wavelengths, which may or may not match well to the quantum conversion efficiency of the PMT photocathode. Futher, the photocathode layer will have some amount of variance in thickness, which affects the variance in the number of electrons ejected from the photocathode per initial gamma-ray interaction event.

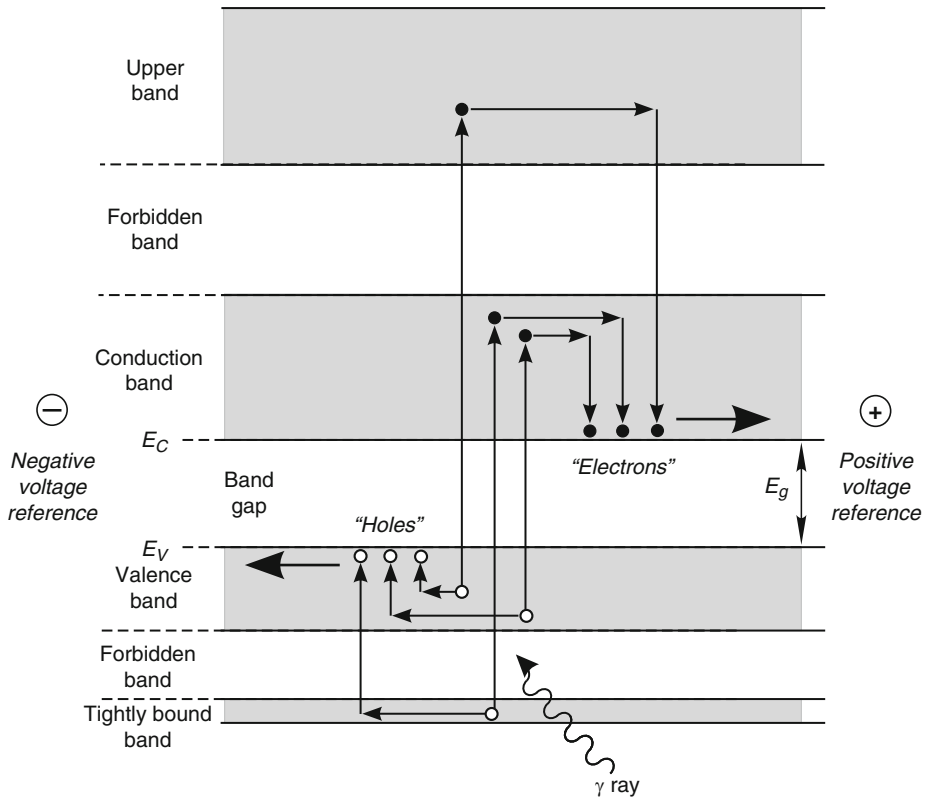


Fig. 9

Absorbed radiation energy excites electrons from the valence and tightly bound bands up into the higher conduction bands, in a similar manner as excitation occurs in a crystalline inorganic scintillator. The empty states left behind, referred to as *holes*, behave as positive charges. The electrons quickly de-excite to the lowest conduction-band edge, E_C, and the holes rapidly de-excite to the top of the valence band, E_V. A voltage applied to the detector causes the electron and hole charge carriers to drift to the device contacts, much in the same manner as electron–ion pairs drift to the electrical terminals in a gas-filled ion chamber (McGregor 2008)

7.2 Semiconductor Spectrometers

In some way, the operation of a semiconductor detector combines the concepts of the charge excitation method in a crystalline inorganic scintillator and the charge collection method of a gas-filled ion chamber. Referring to *Fig. 9*, gamma rays or charged particles that are absorbed in the semiconductor will excite electrons from the valence and tightly bound energy bands up into the numerous conduction bands. The empty states left behind by the negative electrons will behave as positively charged particles generally referred to as *holes*. The excited electrons will rapidly de-excite to the conduction-band edge, E_C. Likewise, as electrons high in the valence

band fall to lower empty states in the valence and tightly bound bands, it gives rise to the effect of holes moving up to the valence-band edge, E_V.

A single major difference between a semiconductor and almost all scintillators is that the mobility of charge carriers in semiconductors is high enough to allow for conduction, whereas scintillation materials are mostly insulating materials that do not conduct. As a result, a voltage can be applied across a semiconductor material that will cause the negative electrons and positive holes, commonly referred to as *electron–hole pairs*, to drift in opposite directions, much like the electron–ion pairs in a gas-filled ion chamber. In fact, at one time, semiconductor detectors were referred to as "solid-state ion chambers." As these charges drift across the semiconductor, they induce a current to flow in an external circuit which can be measured as a current or stored across a capacitor to form a voltage. Semiconductors are far more desirable for energy spectroscopy than gas-filled detectors or scintillation detectors because they are capable of much higher energy resolution. The observed improvement is largely due to the better statistics regarding the number of charges produced by a radiation interaction. Typically, it only takes 3–5 eV to produce an electron–hole pair in a semiconductor. By comparison, it takes between 25–40 eV to produce an electron–ion pair in a gas-filled detector and between 100 eV and 1 keV to produce a single photoelectron ejection from the PMT photocathode in a scintillation/PMT detector (primarily due to light reflections and poor quantum efficiency at the photocathode). Hence, statistically, semiconductors produce more charge carriers from the primary ionization event, which determines the statistical fluctuation in the energy resolution (*Table 3*).

Most semiconductor detectors are configured as either planar or coaxial geometries, as shown in *Fig. 10*. Small semiconductor detectors are configured as planar devices and can be used for charged-particle detection and gamma-ray detection. Large semiconductor gamma-ray spectrometers are usually configured in a coaxial form to reduce the capacitance of the detector (which can affect the overall energy resolution). There are three basic methods generally used to reduce leakage currents through semiconductor detectors. Most commonly, the semiconductors are formed into reverse-biased *pn*- or *pin*-junction diodes, which is the case for Ge, Si, GaAs, CdTe, and InP detectors. Alternatively, highly resistive semiconductors, such as CdZnTe and HgI_2, need only have ohmic contacts because the bulk resistance of the material is high enough to reduce leakage currents to manageable levels. Finally, large detectors, such as high-purity Ge detectors and lithium-drifted Si detectors, are chilled with liquid nitrogen (LN2) or a mechanical refrigerator to reduce thermally generated leakage currents.

Table 3

Common semiconductors and properties

Semiconductor	Atomic numbers Z	Density (g cm^{-3})	Band gap (eV)	Ionization energy (eV per e–h pair)	Fano factor
Si	14	2.33	1.12	3.61	≈ 0.10
Ge	32	5.33	0.72	2.98	≈ 0.08
GaAs	31/33	5.32	1.42	4.2	≈ 0.18
CdTe	48/52	6.06	1.52	4.43	≈ 0.15
$Cd_{0.8}Zn_{0.2}Te$	48/30/52	6.0	1.60	5.0	≈ 0.09
HgI_2	80/53	6.4	2.13	4.3	≈ 0.19

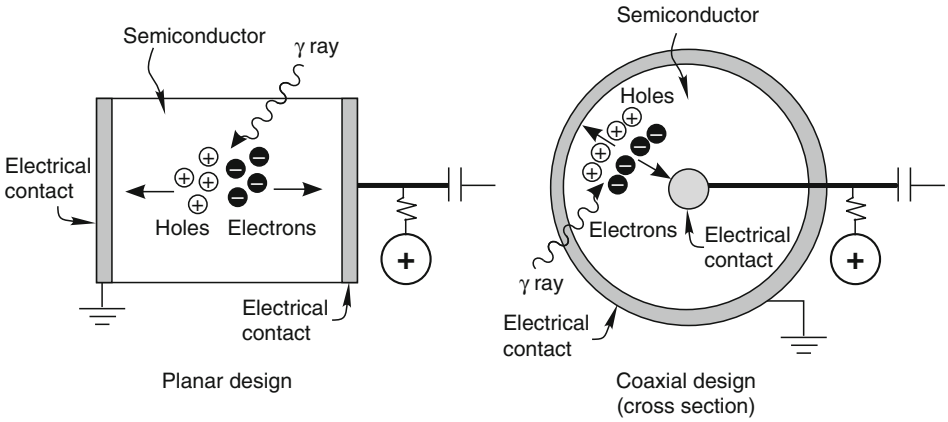


Fig. 10
The most common designs for semiconductor detectors are the planar and coaxial configurations (McGregor 2008)

7.2.1 Ge Detectors

Although Li drifting allowed for the construction of Ge-based semiconductor gamma-ray spectrometers, denoted Ge(Li) detectors, the detectors came with problems. Li is highly mobile in Ge and must be locked into place by immediately freezing the Ge crystal with LN2 after the drifting process is finished. Further, if Ge(Li) detectors were ever allowed to warm up, the Li would diffuse and redistribute, hence ruining the detectors. As a result, Ge(Li) detectors had to constantly be kept at LN2 temperatures, a major inconvenience. Zone refinement of Ge materials allows for impurities to be removed from the material such that Li drifting is no longer necessary. Hence, Ge(Li) detectors have been replaced by high-purity Ge detectors, denoted HPGe detectors. However, HPGe detectors must still be chilled with LN2 *when operated* in order to reduce excessive thermally generated leakage currents.

HPGe detectors have exceptional energy resolution compared to scintillation and gas-filled detectors. The dramatic difference in the energy resolution between NaI(Tl) and HPGe spectrometers is shown in *Fig. 11*, where there is a spectroscopic comparison of measurements made of a mixed ^{152}Eu, ^{154}Eu, and ^{155}Eu radiation source. HPGe detectors are standard high-resolution-spectroscopy devices used in the laboratory. Their high-energy resolution allows them to easily identify radioactive isotopes for a variety of applications, which includes impurity analysis, composition analysis, and medical isotope characterization. Portable devices with small LN2 dewars are also available for remote spectroscopy measurements, although the dewar capacity allows for only 1 day of operation. Hence, a source of LN2 must be nearby. More recently, portable HPGe detectors with small mechanical refrigerators have become available, thereby, eliminating the need for LN2.

The gamma-ray absorption efficiency for Ge (Z = 32) is much less than that for the iodine (Z = 53) in NaI(Tl). Due to the higher atomic number and generally larger size, NaI(Tl) detectors often have higher detection efficiency for high-energy gamma rays than do HPGe detectors (but much poorer energy resolution). When first introduced, Ge detector efficiency was commonly compared to that of a 3 inch diameter × 3 inch long (3×3) right circular cylinder of

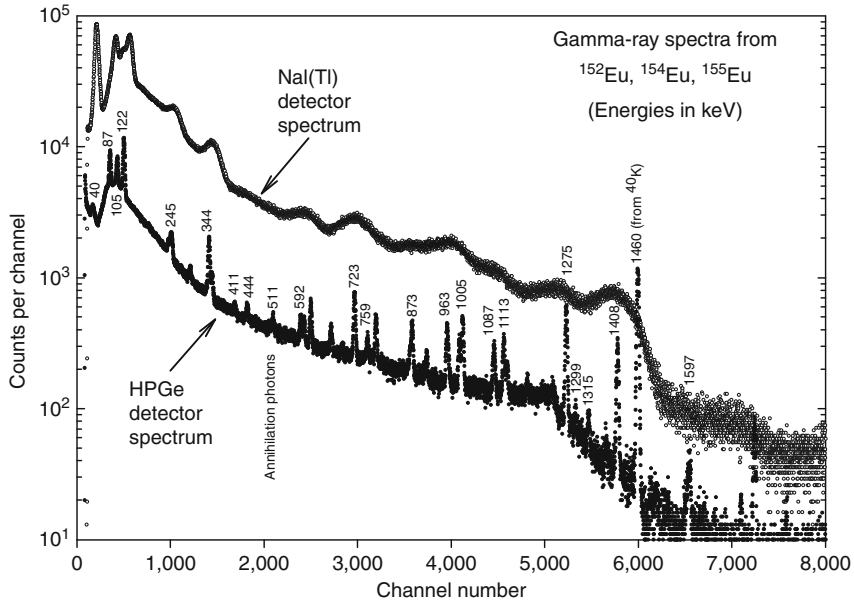


Fig. 11

Comparison of the energy resolution between a NaI(Tl) and an HPGe detector. The gamma-ray source is a mixture of ^{152}Eu, ^{154}Eu, and ^{155}Eu (McGregor 2008)

NaI(Tl) for 1,332 keV gamma rays from ^{60}Co. Even today, efficiency of a Ge detector is quoted in terms of a 3 × 3 NaI(Tl) detector (Fairstein 1996). For instance, an HPGe detector denoted as 60% relative efficiency will have 60% of the efficiency that a 3 × 3 NaI(Tl) detector would have for 1,332 keV gamma rays from ^{60}Co. HPGe detectors are much more expensive than NaI(Tl) detectors, hence are best used when gamma-ray energy resolution is most important for measurements. If efficiency is of greatest concern, it is often wiser to use a NaI(Tl) detector. Although very expensive, modern manufacturers do produce larger HPGe detectors with 200% relative efficiency.

7.2.2 Si Detectors

The problem with Li redistribution does not apply to Si; hence, Si(Li) detectors are still available. Because Si(Li) detectors have a much lower atomic number than HPGe, their relative efficiency per unit thickness is significantly lower for electromagnetic radiation. However, for X-ray or gamma-ray energies less than about 30 keV, commercially available Si(Li) detectors are thick enough to provide performance which is comparable to HPGe detectors. For example, a 3–5 mm thick detector with a thin entrance window has an efficiency of 100% near 10 keV. Si(Li) detectors are preferred over HPGe detectors for low-energy X-ray measurements, primarily due to the lower-energy X-ray escape-peak features that appear in a Si(Li) detector spectrum as opposed to an HPGe detector spectrum. Further, background gamma rays tend to interact more strongly with HPGe detectors than with Si(Li) detectors, which also complicates the X-ray

spectrum. Based upon the fact that a majority of the detector applications require a thin window, Si(Li) detectors are often manufactured with very thin beryllium windows. Typically, Si(Li) detectors are chilled with LN2 to reduce thermal leakage currents, improving performance.

High-purity Si detectors, which do not incorporate Li drifting, are also available, but are significantly smaller than HPGe and Si(Li) detectors. Such devices are typically only a few hundred microns thick and are designed for charged-particle spectroscopy. They range in diameter from 1 cm to several cm. The detectors are formed as diodes to reduce leakage currents, and use either a thin metal contact or a thin implanted-dopant-layer contact to produce the rectifying diode configuration. The devices are always operated in reverse bias to reduce leakage currents. Heavy charged particles, such as alpha particles, rapidly lose energy as they pass through a substance, including the detector contacts. In order to preserve the original energy of charged particles under investigation, the detector contacts and implanted junctions are relatively thin, typically being only a few hundred nanometers thick to reduce energy loss in the contact layer. Further, the measurements are typically made in a vacuum chamber to reduce energy loss otherwise encountered by the alpha particles in air. Because the detectors are not very thick, they do not have much thermal charge-carrier generation and consequently do not need to be cooled during operation.

7.2.3 Compound Semiconductor Detectors

Although HPGe and Si(Li) detectors have proven to be useful and important semiconductor detectors, the fact that they must be chilled with LN2 is a considerable inconvenience. Hence, much research has been devoted to the search for semiconductors that can be used at room temperature. The main requirement is that the band-gap energy (E_g) be greater than 1.4 eV, which seriously limits the field of candidates. Further, the material must be composed of high atomic numbers for adequate gamma-ray absorption. As a result, there are only a few candidates, all of which are compound semiconductors, meaning that they are composed of two or more elements. Hence, the issues regarding crystal growth defects and impurities become far more problematic. Still, there are several materials that show promise, three of which are briefly mentioned here.

HgI_2, CdTe, and CdZnTe Detectors

Mercuric iodide (HgI_2) has been studied since the early 1970s as a candidate gamma-ray spectrometer, and has been used for commercial X-ray spectrometry analysis tools. The high atomic numbers of Hg ($Z = 80$) and iodine ($Z = 53$) make it attractive as an efficient gamma-ray absorber, and its large band gap of 2.13 eV allows it to be used as a room-temperature gamma-ray spectrometer. However, the bright red crystals are difficult to grow and manufacture into detectors. The voltage required to operate the devices is excessive, usually 1,000 V or more for a device only a few mm thick. HgI_2 detectors can degrade over time, an effect referred to as polarization, which is another reason why they do not enjoy widespread use.

Cadmium telluride (CdTe) has been studied since the late 1960s as a candidate gamma-ray spectrometer. They have relatively good gamma-ray absorption efficiency, with Cd ($Z = 48$) and Te ($Z = 52$). The band gap of 1.52 eV allows CdTe to be operated at room temperature. Compared to HgI_2, the crystals are easier to grow and are not as fragile. Further, although still difficult to manufacture, detectors are easier to produce than HgI_2. There are commercial vendors of CdTe detectors, although the devices are relatively small, typically being only a few mm thick with area

of only a few mm^2. CdTe detectors have been used for room-temperature-operated low-energy gamma-ray spectroscopy systems, and also for electronic personal dosimeters. Over time, CdTe detectors also suffer from polarization.

Cadmium zinc telluride (CdZnTe or CZT) has been studied as a gamma-ray spectrometer since 1990. By far, the most studied version of CZT has 10% Zn, 40% Cd, and 50% Te molar concentrations, which yields a band-gap energy of approximately 1.6 eV. CZT detectors offer an excellent option for low-energy X-ray spectroscopy where cooling is not possible. Although the detectors are quite small compared to HPGe and Si(Li) detectors, they are manufactured in sizes ranging from 0.1 cm^3 to 2.5 cm^3, depending on the detector configuration. Still, due to their small size, they perform best at gamma-ray energies below 1.0 MeV. Various clever electrode designs have been incorporated into new CZT detectors to improve their energy resolution, and CZT has become the most used compound semiconductor for gamma-ray spectroscopy. Some detector cooling (near −30 °C), usually performed with miniature electronic Peltier coolers, improves the resolution performance, although excellent performance can be achieved at room temperature. The average ionization energy is 5.0 eV per electron–hole pair, which is greater than Ge (2.98 eV) or Si (3.6 eV). Hence, the resolution of CZT detectors is not as good as HPGe or Si(Li) detectors, although much better than gas-filled and scintillation detectors. When LN2 chilling is not an option, CZT detectors are a good choice for radiation measurement applications requiring good energy resolution (*Fig. 12*). Typically, CZT detectors do not show polarization effects.

7.2.4 Factors Affecting Energy Resolution

The energy resolution achievable from a semiconductor spectrometer is largely determined by the average ionization energy, leakage currents, electronic noise, mean free drift times, and the charge-carrier mobilities. Energy resolution is quoted in terms of energy spread at the full width at half the maximum (FWHM) of a spectral full-energy peak,

$$\mathrm{FWHM} = \left[\left(\mathrm{FWHM}_{\mathrm{noise}}\right)^2 + \left(2.35\sqrt{wFE}\right)^2 \right]^{1/2}, \tag{39}$$

where w is the average energy to produce an electron–hole pair, E is the photon energy, and F is the Fano factor (typically 0.1). The Fano factor is a correction factor to account for typically higher energy resolution than predicted from pure Gaussian statistics (see *Table 3*). For semiconductor detectors with short charge-carrier mean free drift times $\tau_{e,h}$ and low charge-carrier mobilities $\mu_{e,h}$, the energy resolution will suffer from loss of charge carriers during the collection process, hence a variance in the current measured from monoenergetic gamma-ray events as a function of the interaction position and detector size.

The total charge collected is usually affected by crystalline imperfections that serve as *trapping* sites, which are energy states that remove free charge carriers from the conduction and valence bands. Charge is induced while these charge carriers are in motion; hence, their removal diminishes the output voltage. Although the actual trapping process is complicated, it is typical to describe the relative charge collection efficiency as a simplified function of trapping. For planar-shaped detectors, the induced charge is given by

$$\frac{Q}{Q_0} = \xi_e\left(1 - \mathrm{e}^{-x/(\xi_e W_d)}\right) + \xi_h\left(1 - \mathrm{e}^{(x-W_d)/(\xi_h W_d)}\right), \tag{40}$$

where W_d is the detector active-region width, Q_0 is the initial excited charge magnitude, x is the event location in the detector, and

$$\xi_{e,h} = \left(\frac{\tau_{e,h}\nu_{e,h}}{W_d}\right) = \left(\frac{\mu_{e,h}\tau_{e,h}V}{W_d^2}\right), \tag{41}$$

where ν is the charge-carrier speed, V is the applied operating voltage, and the e and h subscripts denote properties for electrons and holes, respectively. Note that the relative charge collection is dependent upon the interaction location x, and for low values of $\xi_{e,h}$, the energy resolution is poor. Typically, good energy resolution is achieved if $\xi_{e,h} > 50$ for both electrons and holes, where Q/Q_0 has little deviation over the detector width W_d. Otherwise, the energy resolution suffers for higher-energy γ rays ($\gtrsim 300$ keV). The value of $\xi_{e,h}$ can be increased by decreasing the detector width W_d, increasing carrier mean free drift times $\tau_{e,h}$ through material improvement, or increasing the applied voltage V. Due to practical voltage limitations and the fundamental difficulty with improving materials, most compound semiconductor detectors are manufactured with small active widths to improve detector energy resolution, and, hence, the devices are relatively small. The $\mu\tau$ values for electrons and holes are often quoted measures of quality for compound semiconductors used as γ-ray spectrometers.

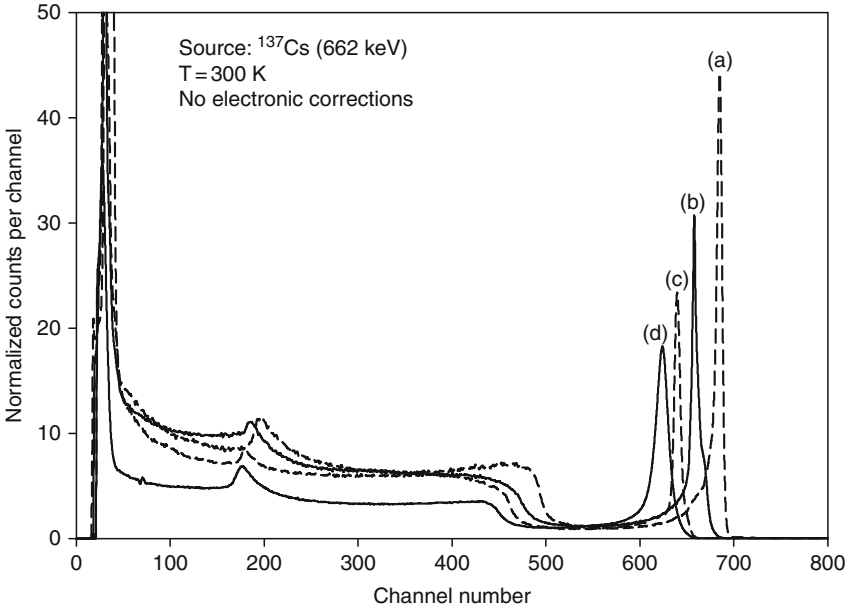


Fig. 12

Performance of various-size Frisch collar CdZnTe detectors exposed to 662 keV gamma rays from ^{137}Cs. Resolutions shown are (a) 0.9% FWHM for a 4.7 × 4.7 × 9.5 mm^3 device, (b) 1.1% FWHM for a 6.5 × 6.5 × 13.0 mm^3 device, (c) 1.2% FWHM for a 7.8 × 7.8 × 15.6 mm^3 device, and (d) 2.4% FWHM for a 11.0 × 11.0 × 22.0 mm^3 device. The detectors were operated at room temperature without the assistance of electronic correction methods (after Kargar et al. 2009)

7.3 Cryogenic Spectrometers (Microcalorimeters)

Microcalorimeter detectors are energy-dispersive spectrometers that measure the thermal change ΔT in an absorber rather than the change in charge concentration ΔQ. The detector consists of an absorber in contact with a type of low-temperature (mK range) thermometer. When absorbed, an X-ray produces heat in the absorber material which can then be measured as $\Delta T \approx E/C_h$, where C_h is the heat capacity of the absorber and E is the initial X-ray energy. Hence, a measurement of the thermal rise in temperature can yield the photon energy.

Early microcalorimeters used semiconductor thermistors as the thermometer. An X-ray absorption causes the resistance in the thermistor to increase, hence producing a change in voltage for current-biased devices. These voltages can be measured as an indication of the ΔT observed in the detector. Although effective, yielding energy resolutions below 8 eV for 5.9 keV gamma rays from ^{55}Fe, the resolution is limited by the heat capacity of the absorbers.

The heat capacity is a function of the absorber volume and T^3. In general, the change in FWHM can be approximated by

$$\Delta_{\mathrm{FWHM}} \approx 2.35\eta\sqrt{\mathtt{k}T^2C_h}, \tag{42}$$

where $\mathtt{k}$ is Boltzmann's constant, T is the absolute temperature, and η is an experimental constant dependent upon thermal conductance and heat capacity. From Eq. 42, it becomes clear that the energy resolution improves as the sample volume decreases, yet this resolution improvement comes at the expense of detection efficiency.

Another form of the microcalorimeter utilizes superconducting transition edge sensor (TES) thermometers. The device is chilled well below the transition edge, and heated ohmically by applying a constant voltage bias to the absorber. The bias is adjusted such that the temperature of the device is maintained slightly below the transition edge. The absorption of an X-ray causes the superconducting absorber to become normal conducting, thereby increasing the resistance and decreasing the current. The current is measured through induction with a superconducting quantum interference device (SQUID) current amplifier.

Typically, the choice of absorber depends greatly upon the photon energy of interest. Energy resolution of below 2 eV has been achieved for 5.9 keV gamma rays from ^{55}Fe using Bi absorbers on Mo–Au TES thermometers. Higher-energy gamma rays, yet generally below 100 keV, have good results from superconducting Sn, producing an energy resolution below 30 eV for 102 keV gamma rays.

The response time is limited by the heat capacity, in which the reset time is dependent upon the time it takes to return the detector temperature to equilibrium, where the cooling time is represented by $\tau = C_h/G$, where G is the thermal conductance between the thermometer and the cryostat. Arrays of microcalorimeters can be used to maintain fast response time while increasing detection efficiency. *Figure 13* shows comparison spectra for a typical HPGe semiconductor detector and a TES microcalorimeter array.

7.4 Crystal Diffractometers (Wavelength-Dispersive Spectroscopy)

Ultrahigh resolution can be achieved for low-energy gamma and X rays with wavelength-dispersive spectroscopy (WDS), which can yield X-ray peak resolution better than

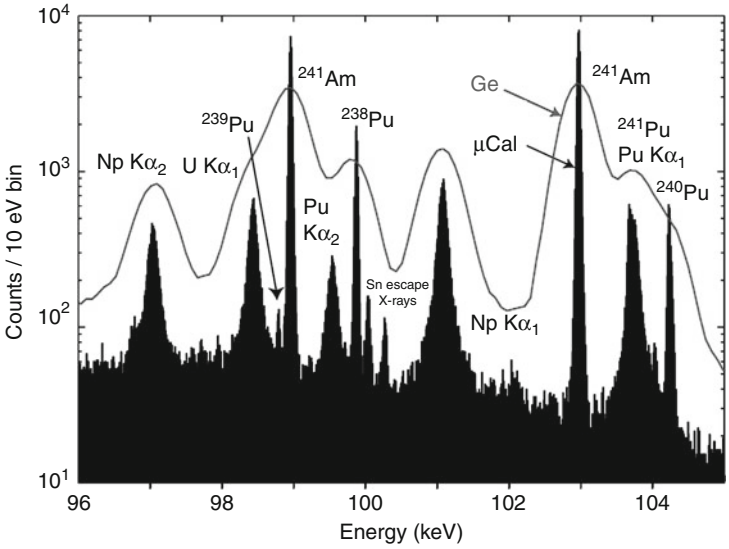


Fig. 13

Pu spectrum from a microcalorimeter array using data from 11 of 13 active pixels. The combined array resolution is approximately 45 eV. At this resolution, the broad X-ray peaks can be readily distinguished from gamma-ray peaks. The solid curve is a spectrum taken with a conventional HPGe detector (Bacrania 2009)

semiconductor or cryogenic detectors. The method utilizes Bragg scattering, in which the Bragg condition must be satistified,

$$n\lambda = 2d \sin\theta, \tag{43}$$

where n is an integer, d is the spacing between crystalline planes, λ is the wavelength of the photon under inspection, and θ is the angle at which radiation intersects the crystal from the parallel condition. It is difficult to make a portable system; hence, these instruments are generally attached to an electron microprobe or scanning electron microscope.

Figure 14 shows a common arrangement for the tool, in which a sample under inspection is irradiated with an electron beam, thereby producing characteristic X rays from the sample. These X rays intersect a slightly bent diffraction crystal. Those X rays satisfying the Bragg condition will diffract into a detector and be recorded, whereas other X rays are absorbed, scatter randomly, or pass through the crystal. Because only the number of counts at a given diffraction angle need be recorded, the detector need not be a high-resolution spectrometer; hence, a gas-filled proportional counter is commonly used as the X-ray detector. During operation, the crystal and detector are rotated through a Rowland circle, which allows for the Bragg condition to be maintained as the arrangement rotates through a continuum of wavelengths.

As a result, the X-ray detector records the number of counts as a function of wavelength. The stringent requirement for the Bragg condition results in ultrahigh resolution, which can be plotted as a function of photon energy (see *Fig. 15*). The important advantage of WDS is the superior identification ability it provides to the user. Unfortunately, the system can be used only on photons of energy low enough to Bragg diffract. Several commercial systems have a rotating rack of different diffraction crystals that extend the sensitive range. WDS systems are laboratory based, hence are not generally considered portable.

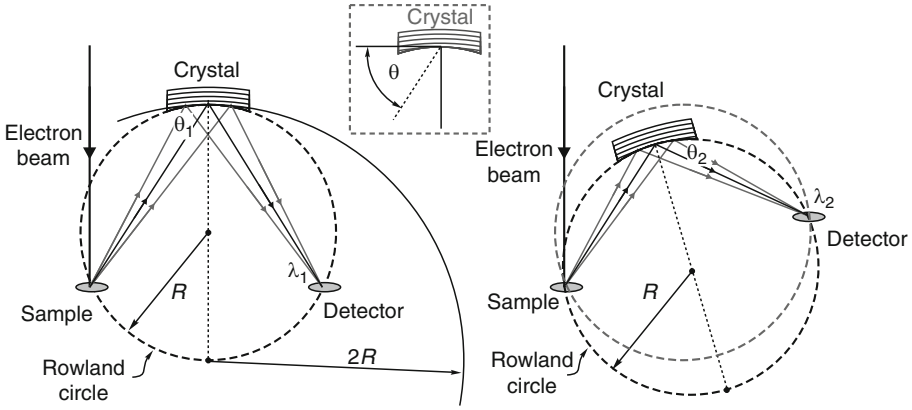


Fig. 14

A typical WDS diffraction arrangement is aligned on a Rowland circle. The sample location remains stationary. The diffraction crystal is bent with a radius twice that of the Rowland circle radius *R*, and it is typically ground with radius *R*. The Bragg condition is maintained for various values of λ by moving both the crystal and detector, with the sample remaining stationary, such that all points remain on the Rowland circle

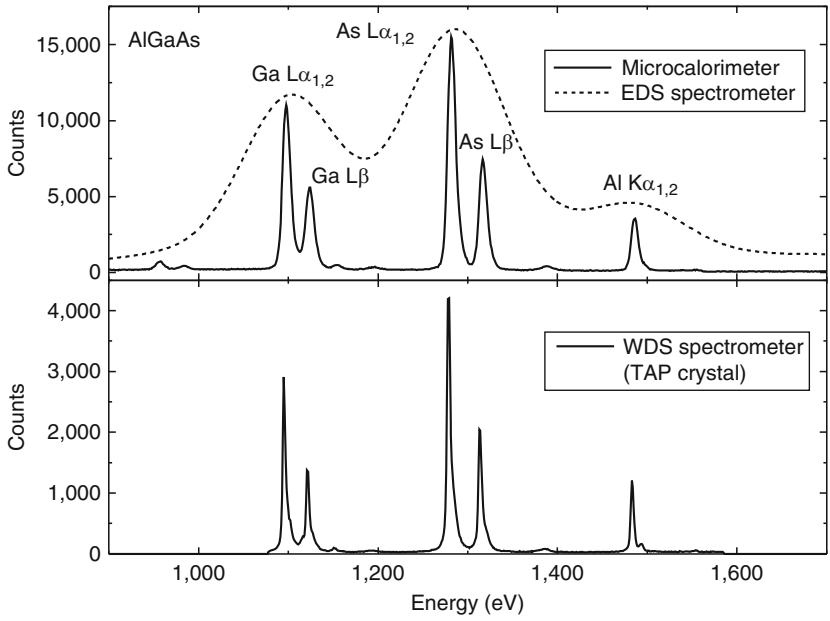


Fig. 15

Shown are (*top*) a comparison of EDS spectra from a Si(Li) detector and a microcalorimeter detector and (*bottom*) an additional comparison to a WDS detector (courtesy of Wollman et al. (1997))

8 Conclusions

Gamma-ray spectroscopy seeks to determine, first, the gamma-ray energies emitted by a sample or the nuclides (which emit gamma rays of certain energies) present within a sample. This goal can be called *qualitative analysis*. However, generally, one also seeks either the source strength of the particular gamma rays or the concentration of the nuclide emitting the gamma rays. This often is referred to as *quantitative analysis*.

One achieves the objectives of gamma-ray spectroscopy by analyzing spectra. In this chapter, pulse-height spectra in which number of counts are specified by discrete channel number are considered. This approach is used because channel number is directly proportional to pulse height and photon energy can then be obtained from pulse height whether or not the spectrometer is linear. Spectroscopic analysis techniques begin by attempting to determine the continuous channel number that corresponds to the centroids of the full-energy peaks in the spectrum that are of interest. The MCLLS approach focuses on the entire spectrum and seeks to determine the parameters of models that best fit all or major portions of the spectra. The symbolic Monte Carlo approach holds promise for spectroscopic applications in which the model is nonlinear in terms of the nuclide concentrations of samples.

Depending upon the need, there are several devices that can be used for gamma-ray spectroscopy. Efficiency with adequate energy resolution can be provided with large-volume scintillators, whereas high-energy resolution can be achieved with semiconductor detectors. Both scintillator and semiconductor detectors can be acquired as portable units with good detection efficiency for gamma rays. For ultrahigh energy resolution, microcalorimeters or WDS diffractometers offer excellent performance. Yet, microcalorimeters and WDS spectrometers are generally restricted to laboratory-based instrumentation for low-energy gamma rays and X rays. Note that the spectrometers discussed in the present chapter represent only a select sample of variations commercially available. More information can be found in the book chapters dedicated to semiconductor and scintillation detectors.

9 Cross-References

- Chapter 15, "Scintillation Counters"
- Chapter 16, "Semiconductor Counters"

Acknowledgments

The assistance of Prof. J.K. Shultis at Kansas State University is much appreciated.

References

Bacrania MK et al (2009) Large-area microcalorimeter detectors for ultra-high-resolution X-ray and gamma-ray spectroscopy, IEEE Trans Nuc Sci 56(4):2299–2302

Bale G, Holland A, Seller P, Lowe B (1999) Cooled CdZnTe detectors for X-ray astronomy. Nucl Instrum Meth Phys Res A 436: 150–154

Barache D, Antoine J-P, Dereppe J-M (1997) The continuous wavelet transform, an analysis tool for NMR spectroscopy. J Magnetic Res 128:1–11

Bevington PR (1969) Data reduction and error analysis for the physical sciences. McGraw-Hill, New York

Birks JB (1964) The theory and practice of scintillation counting. Pergamon Press, Oxford

Dunn WL (1981) Inverse Monte Carlo analysis. J Comput Phys 41(11):154–166

Dunn WL, Dunn TS (1982) An assymetric model for XPS analysis. Surf Interface Anal 4(3): 77–88

Dunn WL, Shultis JK (2009) Monte Carlo analysis for design and analysis of radiation detectors. Radiat Phys Chem 78:852–858

Fairstein E et al (1996) IEEE standard test procedures for germanium gamma-ray detectors, IEEE Std 325-1996, NIDC

Gardner RP, Sood A (2004) A Monte Carlo simulation approach for generating NaI detector response functions (DRFs) that accounts for nonlinearity and variable Flat Continua. Nucl Instrum Meth Phys Res B 213:87–99

Gardner RP, Xu L (2009) Status of the Monte Carlo library least-squares (MCLLS) approach for nonlinear radiation analyzer problems. Radiat Phys Chem 78:843–851

Gentile NA (2001) Implicit Monte Carlo diffusion an acceleration method for Monte Carlo time-dependent radiative treansfer simulations. J Comput Phys 172:543–571

Goldstein JI, Newbury DE, Echlin P, Joy DC, Fiori C, Lifshin E (1981) Scanning electron microscopy and X-ray microanalysis. Plenum Press, New York

Hornbeck RW (1975) Numerical methods with numerous examples and solved illustrative problems. Quantum Publishers, New York

Irwin KD (1995) An application of electrothermal feedback for high resolution cryogenic particle detection. Appl Phys Lett 66:1998–2000

Irwin KD (1996) X-ray detection using a superconducting transition-edge sensor microcalorimeter with a electrothermal feedback. Appl Phys Lett 69:1945–1947

Iyomoto N et al (2008) Close packed arrays of transition-edge X-ray microcalorimeters with high spectral resolution at 5.9 keV. Appl Phys Lett 92:01358

Kargar A, Brooks AC, Harrison MJ, Chen H, Awadalla S, Bindley G, McGregor DS (2009) Effect of crystal length on CdZnTe frisch collar device performance. IEEE Nucl Sci Symp Conf Rec, Orlando, 24 Oct–1 Nov 2017–2022

Knoll GF (2010) Radiation detection and measurement, 4th edn. Wiley, New York

Marshall III, JH, Zumberge JF (1989) On-line measurements of bulk coal using prompt gamma neutron activation analysis. Nucl Geophys 3:445–459

McGregor DS (2008) In: Shultis JK, Faw RE (eds) Detection and measurement of radiation, in fundamentals of nuclear science and engineering, 2nd edn. CRC Press, New York

McGregor DS, Hermon H (1997) Room-temperature compound semiconductor radiation detectors. Nucl Instrum Meth Phys Res A 395: 101–124

Mickael MW (1991) A complete inverse Monte Carlo model for energy-dispersive X-ray fluorescence analysis. Nucl Instrum Meth Phys Res A 301: 523–542

Molnar GL (2004) Handbook of prompt gamma activation analysis with neutron beams. Kluwer Academic Publishers, Boston

Press WH, Teukolsky SA, Vetterling WT, Flannery BP (1996) Numerical recipes in FORTRAN 77 the art of scientific computing, 2nd edn. Cambridge University Press, New York

Redus RH, Pantazis JA, Huber AC, Jordanov VT, Butler JF, Apotovsky B (1998) Fano factor determination for CZT. Proc MRS 487: 101–107

Rodnyi PA (1997) Physical processes in inorganic scintillators. CRC Press, Boca Raton

Schlesinger TE, James RB (1995) Semiconductors for room temperature nuclear detector applications. In: Semiconductors and semimetals, vol 43. Academic Press, San Diego

Takahashi T et al (2001) High resolution CdTe detector and applications to imaging devices. IEEE Trans Nucl Sci 48:287–291

Wollman DA, Irwin KD, Hilton GC, Dulcie LL, Newbury DE, Martinis JM (1997) High-resolution, energy-dispersive microcalorimeter spectrometer for X-ray microanalysis. J Microscopy 188(3):196–223

Xu Y, Weaver JB, Healy DM Jr, Lu J (1994) Wavelet transform domain filters: a spatially selective noise filtration technique. IEEE Trans Image Proc 3(6):747–758

Yacout AM, Dunn WL (1987) Application of the inverse Monte Carlo method to energy-dispersive X-ray fluorescence. Adv X-Ray Anal 30:113–120

Further Reading

Price WJ (1964) Nuclear radiation detection, 2nd edn. McGraw-Hill Book Co., New York

Tsoulfanidis N, Landsberger S (2011) Measurement and detection of radiation, 3rd edn. Taylor and Francis, Washington, DC

Radiation Spectrometer Suppliers

AmpTek, Inc.; www.amptek.com/index.html
Baltic Scientific Instruments; www.bsi.lv/
Bruker; www.bruker-axs.de/
Canberra Industries, Inc.; www.canberra.com/
Constellation Technology Corporation; www.contech.com/
EI Detection and Imaging; www.evmicroelectronics.com/
Eurorad; www.eurorad.com/detectors.php
Moxtek; www.moxtek.com/
Ortec Advanced Measurement Technology, Inc.; www.ortec-online.com/
Oxford Instruments, http://www.oxinst.com/Pages/home.aspx
Princeton Gamma-Tech Instruments; www.pgt.com/
Radiation Monitoring Devices, Inc.; www.rmdinc.com/
Redlen Technologies; www.redlen.com/
Rigatu; www.rigaku.com/
Thermo-Eberline; www.esm-online.de/sm/contact/index.html
XRF Corporation; www.xrfcorp.com/

18 Cherenkov Counters

Blair Ratcliff[1] · *Jochen Schwiening*[2]
[1]Stanford Linear Accelerator Center, Stanford University, Menlo Park, CA, USA
[2]Hadronenphysik 1, GSI Helmholtzzentrum für Schwerionenforschung GmbH, Darmstadt, Germany

1 Introduction 454

2 Basic Cherenkov Theory 454

3 Cherenkov Counters 457
3.1 Cherenkov Counter Components: Radiators 458
3.2 Cherenkov Counter Components: Detectors 459

4 Counter Types 460
4.1 Threshold Counters 460
4.2 Imaging Counters 461

5 Examples of Cherenkov Counters 463
5.1 Accelerator-Based Particle Identification Detectors 463
5.1.1 Threshold Cherenkov Counters 464
5.1.2 Imaging Cherenkov Counters: RICH 464
5.2 Astroparticle Physics 466
5.2.1 Underground Neutrino Detectors 466
5.2.2 Neutrino Detectors in Natural Water or Ice 467
5.2.3 Imaging Air Cherenkov Telescopes 468

6 Conclusions 468

Acknowledgment 469

References 469

Further Reading 471

C. Grupen, I. Buvat (eds.), *Handbook of Particle Detection and Imaging*,
DOI 10.1007/978-3-642-13271-1_18, © Springer-Verlag Berlin Heidelberg 2012

Abstract: When a charged particle passes through an optically transparent medium with a velocity greater than the phase velocity of light in that medium, it emits prompt photons, called Cherenkov radiation, at a characteristic polar angle that depends on the particle velocity. Cherenkov counters are particle detectors that make use of this radiation. Uses include prompt particle counting, the detection of fast particles, the measurement of particle masses, and the tracking or localization of events in very large, natural radiators such as the atmosphere, or natural ice fields, like those at the South Pole in Antarctica. Cherenkov counters are used in a number of different fields, including high energy and nuclear physics detectors at particle accelerators, in nuclear reactors, cosmic ray detectors, particle astrophysics detectors and neutrino astronomy, and in biomedicine for labeling certain biological molecules.

This chapter begins with a brief history of the Cherenkov effect. It then describes some salient features of the radiation that leads to its unique value in particle detection. Several different classes of Cherenkov detectors will be described, along with the technology needed to build them. The chapter will conclude with a review of a number of different Cherenkov counters, including some historically important counters, more recent devices now in operations, and devices that remain under research and development that make use of innovative technologies.

1 Introduction

More than 100 years ago, Marie and Pierre Curie enjoyed seeing beautiful, if slightly eerie, bluish-white luminescence from their concentrated radioactive solutions (Curie 2001), but their observations occurred long before the complex light-emitting effects in these solutions were understood, or, indeed, before the health dangers of the ionizing radiation producing the effects were realized. Early investigations were deterred not only by competing effects, such as fluorescence, but by the very limited number of photons emitted, and by the lack of light detectors with sufficient sensitivity. Thorough, inventive experimental investigations to fully explore the phenomena, now called Cherenkov radiation, were carried out with quite simple apparatus beginning in 1934 by P. Cherenkov (see, for example, the earliest in a series of papers by Cherenkov 1934). These detailed observations both agreed with and were fully explained theoretically by I. Frank and I. Tamm using classical electromagnetic theory in a landmark paper in 1937 (Frank and Tamm 1937), resulting in the award of the Nobel Prize to these three physicists in 1958.

2 Basic Cherenkov Theory

Cherenkov light is an electromagnetic analog of the more familiar sonic boom produced by an aircraft moving faster than the speed of sound in air, and is possible only because the phase velocity of light in transparent materials with refractive index n is slower than the speed of light (c) in a vacuum. The shock wave (sometimes called the Mach cone) can be observed as a very

prompt pulse of light emitted uniformly in azimuth (ϕ_c) around the particle direction with the characteristic Cherenkov polar opening angle (θ_c),

$$\cos\theta_c = \frac{1}{n(\lambda)\beta},$$

where $\beta = v_p/c$, v_p is the particle velocity, and $n(\lambda)$ is the index of refraction of the material. Since the index of refraction, n, is a function of the photon wavelength, in normal optical materials there is an intrinsic Cherenkov-angle resolution smearing that depends on the bandwidth of the detected photons. No Cherenkov light emission occurs below a particle threshold velocity $\beta_t = 1/n$. Since $\gamma_t = 1/\left(1-\beta_t^2\right)^{1/2}$, $\beta_t\gamma_t = p_t/m = 1/(2\eta+\eta^2)^{1/2}$, where p_t is the threshold particle momentum for a particle with rest mass m, and $\eta = n-1$.

Cherenkov emission is a weak effect, and, as such, causes no significant loss to the particle energy. The number of Cherenkov photons N_{photons} produced by a charged particle of charge z, within the total Cherenkov bandwidth, is given by the Frank–Tamm equation (Frank and Tamm 1937)

$$N_{\text{photons}} = L\frac{\alpha^2 z^2}{r_e\, m_e c^2}\int \sin^2\theta_c(E)\,\mathrm{d}E,$$

where L is the length of the particle path through the radiator in cm, E is the photon energy in eV. The integral is taken over the region where $n(E)$ is greater than 1, and $\alpha^2/(r_e\, m_e c^2) = 370\ \text{cm}^{-1}\,\text{eV}^{-1}$.

As first shown in a classical paper by Tamm in 1939 (Tamm 1939), the conical Cherenkov radiation shell is not quite perpendicular to the Cherenkov propagation angle in normal optical media, which are always dispersive. With angles defined in ❯ *Fig. 1*, the half-angle of the cone opening (η_c) is given by,

$$\cot\eta_c = \left[\frac{\mathrm{d}}{\mathrm{d}\omega}(\omega\tan\theta_c)\right]_{\omega_0} = \left[\tan\theta_c + \beta^2\omega\, n(\omega)\frac{\mathrm{d}n}{\mathrm{d}\omega}\cot\theta_c\right]_{\omega_0},$$

where ω_0 is the central value of the small frequency range under consideration. As Motz and Schiff pointed out in 1953 (Motz and Schiff 1953), the presence of the second term means

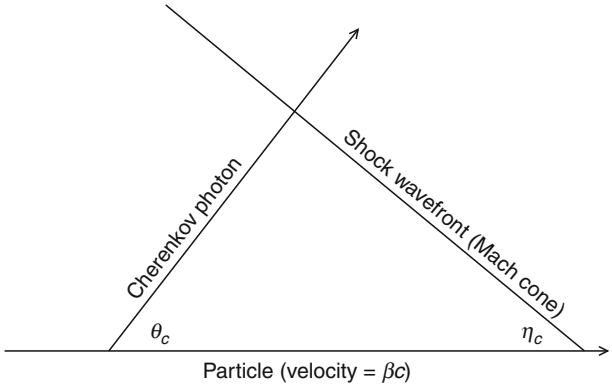


Fig. 1
Schematic showing the Cherenkov cone and angles defined in the text

that the Mach-cone half-angle (η_c) is the complement of the Cherenkov angle (θ_c) only for a nondispersive medium where $dn/d\omega = 0$. Though subtle and unimportant in many Cherenkov applications, this can affect the performance of many modern devices that either are large, or that have very fast photon timing.

In principle, since three measurements (a space point (x, y) and a propagation time (t_p)) may be made at a fixed z location to measure the two Cherenkov angles (θ_c, ϕ_c) with respect to a known track, there is a nominal over-constraint, even for a single photoelectron. In practice, exploiting the time dimension requires very fast photon detectors and/or rather long propagation times, such as those that occur with the DIRC or large water detectors discussed later, so it is less frequently utilized.

For pedagogical purposes, it is useful to recall specifically how the measured quantities are related to the Cherenkov angles. Consider a frame (q), as shown in ❯ *Fig. 2a*, where the particle moves along the (z) axis. The direction cosines of the Cherenkov photon emission in this frame (q_x, q_y, and q_z) are related to the Cherenkov angles by,

$$q_x = \cos\phi_c \sin\theta_c,$$
$$q_y = \sin\phi_c \sin\theta_c,$$
$$q_z = \cos\theta_c.$$

Defining the emission point (z_e) and the detection point (z_d), the propagation time t_p over a length L_p is given by

$$t_p = \frac{L_p n_g}{c} = \frac{(z_d - z_e) n_g}{c q_z},$$

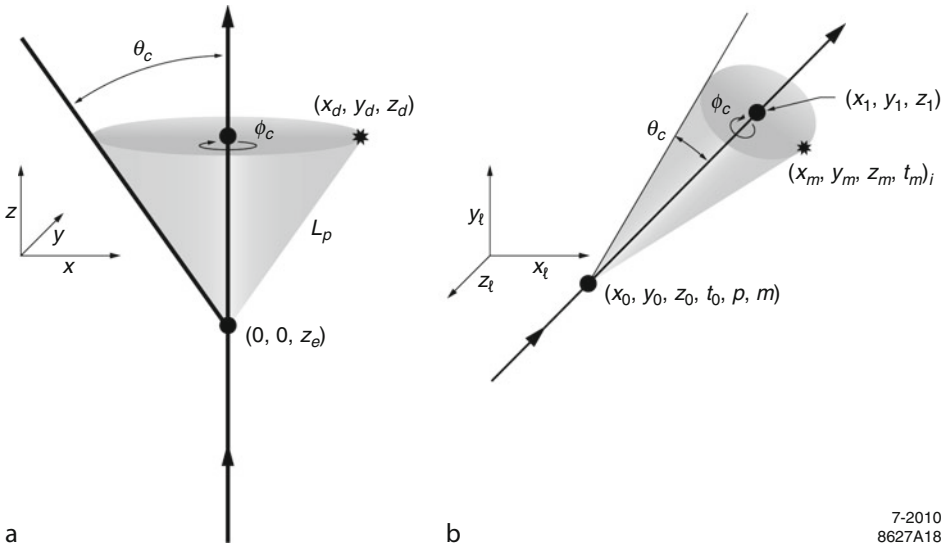


Fig. 2

Schematic of typical Cherenkov cones and reference frames: (a) with respect to the particle path, and (b) in the lab coordinate system

where the photon group velocity ($v_{\text{group}} = c/n_g$) must be used rather than the photon phase velocity ($v_{\text{phase}} = c/n$) since, in a dispersive medium, energy propagates at the photon group velocity. This is another way to understand why the conical Cherenkov radiation shell is not quite perpendicular to the Cherenkov propagation angle in a dispersive medium. In a real counter, the photon measurements are made in a lab frame such as that defined in *Fig. 2b*, perhaps after further reflections and focusing (not shown). The photon emission coordinates are not necessarily well known, but, if the photons are focused, it may not be necessary to know all emission coordinates well, in order to derive Cherenkov angular information. In any case, measured positions and times for photons must be transformed from the measurement frame such as *Fig. 2b* back to one like that of *Fig. 2a* typically using additional tracking and timing information, taken either from outside detectors, or from the correlations between the photons measured within the event itself.

The relationship between group and phase velocities, as a function of photon wavelength (λ), is usually derived in a simple one-dimensional picture (see, for example, Jackson 1962), and leads to the following relationship between the group and phase refractive indices:

$$n_g(\lambda) = n(\lambda) - \lambda \, \mathrm{d}n(\lambda)/\mathrm{d}\lambda.$$

$n_g(\lambda)$ is typically several percent larger than $n(\lambda)$ for photons in the visible and UV energy range and, more importantly for the resolution performance of a counter that uses time focusing (as discussed below), the dispersion of $n_g(\lambda)$ is also substantially greater.

3 Cherenkov Counters

Cherenkov counters are particle detection devices that utilize Cherenkov radiation. They are sometimes thought of as being useful mostly as particle identification (PID) detectors for accelerator physics experiments. PID detectors provide a measure of the mass of charged particles (and thus, determine their identity) by combining a measurement of the velocity of each charged particle made in the PID detector with a measurement of the particle's momentum made in a tracking chamber system that tracks the trajectory of the particle in a magnetic field, using

$$m = \frac{p}{\gamma \beta c}.$$

However, in practice, the use of Cherenkov counters extends over a broad range of applications including, for example, (1) fast particle counters in accelerator instrumentation; (2) hadronic PID in high energy physics particle detectors; (3) tracking detectors performing complete event reconstruction in neutrino astronomy; and (4) making quantitative radiation measurements in biology and medicine. Examples of applications from each category include: (1) the BaBar luminosity detector (Ecklund et al. 2001); (2) the hadronic PID detectors at the B factory detectors – DIRC in BaBar (Adam et al. 2005) and the aerogel threshold Cherenkov in Belle (Abashian et al. 2002); (3) large water Cherenkov counters such as Super-Kamiokande (Ashie et al. 2005) (see Chap. 14, "Neutrino Detectors"); and (4) quantitative measurements of beta particles in microfluidic chips (Cho et al. 2009).

Cherenkov counters contain two main elements: (1) a radiator through which the charged particle passes, and (2) a photodetector. As Cherenkov radiation is a weak source of photons, light collection and detection must be as efficient as possible. The refractive index n and the

particle's path length through the radiator L appear in the Cherenkov relations allowing the tuning of these quantities for particular applications.

Cherenkov detectors utilize one or more of the properties of Cherenkov radiation discussed in ❯ Sect. 2: the prompt emission of a light pulse; the existence of a velocity threshold for radiation; and the dependence of the Cherenkov-cone half-angle θ_c and the number of emitted photons on the velocity of the particle and the refractive index of the medium.

In practical detectors, Cherenkov radiation is nearly always observed using a sensitive photon detector that converts individual photons into photoelectrons. The number of photoelectrons observed by a typical detector for a particle of unit charge is

$$N_{\mathrm{pe}} = 370L \int \epsilon(E) \sin^2 \theta_c(E)\, \mathrm{d}E,$$

where $\epsilon(E)$ is the energy dependence of the photon transducer, and the integral is taken over the detector bandwidth.

The quantities ϵ and θ_c are functions of the photon energy E. As the typical energy-dependent variation of the index of refraction is modest, a quantity called the Cherenkov detector quality factor N_0 can be defined as

$$N_0 = \frac{\alpha^2 z^2}{r_e\, m_e c^2} \int \epsilon\, \mathrm{d}E,$$

so that, taking $z = 1$ (the usual case in high energy physics),

$$N_{\mathrm{pe}} \approx L N_0 \langle \sin^2 \theta_c \rangle.$$

This definition of the quality factor N_0 is not universal, nor, indeed, very useful for those common situations where ϵ factorizes as $\epsilon = \epsilon_{\mathrm{coll}}\epsilon_{\mathrm{det}}$ with the geometrical photon collection efficiency (ϵ_{coll}) varying substantially for different tracks while the photon detector efficiency (ϵ_{det}) remains nearly track independent. In this case, it can be useful to explicitly remove (ϵ_{coll}) from the definition of N_0. A typical value of N_0 for a photomultiplier detection system working in the visible and near UV, and collecting most of the Cherenkov light, is about 100 cm^{-1}. Practical counters, utilizing a variety of different photodetectors, have values ranging between about 30 and 180 cm^{-1}.

In theory, in a nondispersive medium, the shock-cone wavefront is arbitrarily thin, so that the light-pulse duration is a delta function. In practice, since all media are dispersive, light collection systems are not synchronous, and all photon detectors have finite time resolution; the observed pulse duration will be finite as will the pulse rise time. A short pulse with a very rapid rise time could lead to the development of very fast counting devices that would have many applications. Recent experiments have been able to demonstrate a time resolution of 14×10^{-12} s, which is dominated by the photon detector timing response of the Micro-Channel Plate Photomultiplier Tubes (MCP-PMTs, see also ❯ Chap. 13, "Photon Detectors") that are used to detect the Cherenkov photons (Va'vra et al. 2009). This remains a very active area for R&D.

3.1 Cherenkov Counter Components: Radiators

There are many transparent radiators available, ranging from light gases to dense glasses, which allow counters to be designed to cover an extremely wide range of particle momenta. ❯ *Table 1* compares the Cherenkov threshold gamma (γ_t) for a number of different radiator types with differing indices of refraction. In addition to refractive index, the choice requires consideration of

Table 1
Threshold and radiator length (R.L.) required to obtain 10 photoelectrons for a $\beta = 1$ particle for a number of different radiator materials assuming the typical photodetector described in the text with $N_0 = 100$

	Index of refraction	$\gamma_t\beta_t$	R.L. (cm)
He (gas)	1.000035	119.75	1429.
Ne (gas)	1.000067	86.4	746.
N_2 (gas)	1.00030	40.8	167.
C_5F_{12} (gas)	1.0017	17.1	29.5
Aerogel (low density)	1.007	8.4	7.22
Aerogel (high density)	1.13	1.90	0.46
Argon (liquid)	1.23	1.39	0.29
C_6F_{14} (liquid)	1.28	1.25	0.26
H_2O (liquid)	1.34	1.12	0.23
SiO_2 (solid)	1.47	0.93	0.19
LiF (solid)	1.50	0.89	0.18
Diamond (solid)	2.417	0.45	0.12

factors such as material density, radiation length and radiation hardness (see also ❯ Chap. 22, "Radiation Damage Effects"), transmission bandwidth, absorption length, chromatic dispersion, optical workability (for solids), availability, and cost. Tables giving the properties of a variety of commonly used radiator materials can be found or are referenced in the Particle Data Group Reviews (Amsler et al. 2008).

When the momenta of particles to be identified are high, the refractive index must be set close to 1, so that the photon yield per unit length is low and a long particle path in the radiator is required as shown in ❯ *Table 1*. The gap in refractive index that has traditionally existed between gases and liquid or solid materials has been partially closed with transparent silica aerogels with indices that range between about 1.007 and 1.13.

3.2 Cherenkov Counter Components: Detectors

Cherenkov counters became a practical technique for particle detection following the invention and development of the photomultiplier tube (PMT) (Jelley 1958) over 50 years ago. Even today, photon detectors for Cherenkov counters are challenging as they must detect single photons with high efficiency and little noise. Very fast timing resolution is essential in time-imaging counters, and is useful in any case to reject background. Good segmentation in space may be needed to obtain adequate angular resolution in ring imaging counters (RICH, see below), and is also useful to reject backgrounds. On the other hand, much of the costs scale with the number of pixels and must be strictly controlled. Detectors continue as an active arena for R&D (RICH Workshop Series).

There are several distinct types of photon detectors in use or proposed (see also ❯ Chap. 13, "Photon Detectors").

Vacuum photon detectors include dynode-based PMTs, micro-channel plate (MCP) PMTs, hybrid PMTs (HPMTs), etc. All use a photocathode in vacuum but with different techniques for obtaining gain. A wide variety of photocathodes are available that are sensitive to wavelengths

from the UV cutoff of the window material (LiF cuts off around 100 nm) to the near IR. They have an illustrious history in Cherenkov detectors, as most successful counters used PMTs until the 1980s. They are still very widely used, with many opportunities for further development. As a class, they are very sensitive, versatile, fast, and have high gain and low noise. Many are also quite robust in operation. Many types are commercially available, but all are difficult to produce and develop in a normal laboratory without a large infusion of capital and expertise. The usual types are quite sensitive to magnetic fields, but some will work in fields of over 1 T in the appropriate direction. Development continues and highly pixelated fast types have become available commercially recently.

Gaseous detectors (see also Chap. 11, "Gaseous Detectors") can provide inexpensive coverage of a large photon collection area with good point resolution. The single photoelectrons are typically read out with proportional chambers and/or time projection chambers. (Recently built instruments may use other devices like the gaseous electron multiplier (GEM).) Gaseous (tetrakis-dimethylaminoethylene (TMAE) or Tri-ethyl-amine (TEA)) and solid (CsI) photocathodes have been employed. Operational characteristics can be especially challenging. Since the photocathodes work near the UV-window cutoff, the number of photoelectrons is modest, and there is substantial chromatic dispersion. Performance at high luminosities is limited depending in detail on the photocathode and readout. Devices with TMAE photocathodes are quite slow, but TEA and CsI devices can be moderately fast. They can be used in a magnetic field but are not an option for a time imaging RICH.

Solid-state photon detectors (see also Chap. 16, "Semiconductor Counters") are an intriguing possibility for certain future applications. They are very compact, highly segmented, fast, and efficient. Major challenges are noise performance, radiation resistance, and cost.

4 Counter Types

Cherenkov counters may be classified as either imaging or threshold types, depending on whether they do or do not make use of Cherenkov angle (θ_c) information. Imaging counters may be used to track particles as well as identify them. The use of very fast photodetectors such as micro-channel plate PMTs (MCP-PMTs) also potentially allows very fast Cherenkov-based time-of-flight (TOF) detectors of either class (RICH Workshop Series).

4.1 Threshold Counters

Threshold Cherenkov detectors (Litt and Meunier 1973), in their simplest form, make a yes/no decision based on whether the particle is above or below the Cherenkov threshold velocity $\beta_t = 1/n$. A straightforward enhancement of such detectors uses the number of observed photoelectrons (or a calibrated pulse height) to discriminate between species or to set probabilities for each particle species (Bartlett et al. 1987). This strategy can increase the momentum range of particle separation by a modest amount (to a momentum some 20% above the threshold momentum of the heavier particle in a typical case).

Careful designs give $\langle \epsilon_{\text{coll}} \rangle \gtrsim 90\%$. For a photomultiplier with a typical bialkali cathode, $\int \epsilon_{\text{det}}\, \mathrm{d}E \approx 0.27$ eV, so that

$$N_{\text{pe}}/L \approx 90\ \text{cm}^{-1} \langle \sin^2 \theta_c \rangle \qquad (\text{i.e., } N_0 = 90\ \text{cm}^{-1}).$$

Suppose, for example, that n is chosen so that the threshold for species a is p_t, that is, at this momentum species a has velocity $\beta_a = 1/n$. A second, lighter, species b with the same momentum has velocity β_b, so $\cos\theta_c = \beta_a/\beta_b$, and

$$N_{\mathrm{pe}}/L \approx 90\ \mathrm{cm}^{-1}\,\frac{m_a^2 - m_b^2}{p_t^2/c^2 + m_a^2}.$$

For K/π separation at $p = p_t = 1$ GeV/c, $N_{\mathrm{pe}}/L \approx 16\ \mathrm{cm}^{-1}$ and at $p = p_t = 5$ GeV/c, $N_{\mathrm{pe}}/L \approx 0.8\ \mathrm{cm}^{-1}$ for π's while, by design, $N_{\mathrm{pe}} = 0$ for K's.

For limited path lengths, N_{pe} will usually be small. The overall efficiency of the device is controlled by Poisson fluctuations, which can be especially critical for separation of species where one particle type is dominant. Moreover, the effective number of photoelectrons is often less than the average number calculated above due to additional equivalent noise (ENF) from amplification statistics in the photodetector (Amsler et al. 2008).

It is common to design for at least 10 photoelectrons for the high-velocity particle in order to obtain a robust counter. As rejection of the particle that is below threshold depends on not seeing a signal, electronic and other background noise can be important. Physics sources of light production for the below-threshold particle, such as decay to an above-threshold particle or the production of delta rays in the radiator, often limit the separation attainable, and need to be carefully considered.

4.2 Imaging Counters

Imaging counters make the most powerful use of the information available by measuring the ring-correlated angles of emission of the individual Cherenkov photons. Since low-energy photon detectors can measure only the position (and, perhaps, a precise detection time) of the individual Cherenkov photons (not the angles directly), the photons must be "imaged" onto a detector so that their angles can be derived (Ratcliff 2003). It is helpful to consider two quite distinct imaging device types.

The first device type, the correlated Cherenkov Tracking Calorimeter (see also ❯ Chap. 20, "Calorimeters"), uses Cherenkov imaging as a relatively inexpensive technique to instrument a very large volume of material for particle detection as is necessary to search for or study very rare processes such as neutron decay or neutrino interactions. These are complete experimental detectors with the capability to track particles, vertex events, identify charged particles, measure energies via calorimetry, reject backgrounds, and self trigger (Amsler et al. 2008). The photodetectors are distributed either on the surface or throughout the volume of a very large radiator of a kiloton or more. The radiator might be either an optically transparent natural material (such as glacial ice, or deep-sea or lake water) or large tanks of purified material such as mineral oil or water. The photodetectors are usually large PMTs.

The reconstruction makes use of all available (space, time) information for each measured photon (x_m, y_m, z_m, t_m) as shown in ❯ *Fig. 2* above. Since, in this instance, there is no tracking information available from outside detectors, and the photon emission points are unknown at the outset, the number and locations of tracks must be iteratively derived by combining the measured information from the photons with the constraint that Cherenkov photons emerge from source tracks, acting as line sources for the Cherenkov radiation, at a constant polar angle θ_c. The reconstruction methodology is quite complex, not only because events can contain many tracks, with different lengths and vertices, but also because, in practice, even a single particle can shower and

produce many overlapping rings (RICH Workshop Series). However, once reconstructed, these showers provide useful separation between particle species, since non-showering particles such as muons, pions, and protons produce sharp rings, while showering particles such as electrons and photons produce diffuse rings. The number of photons observed provides a measure of the particle energy. The energy for showering particles is essentially linear with the observed photon number, but the relationship for more massive particles is complex. Careful energy calibration is essential. Examples of this detector type are described more extensively in Sect. 5.

The second device type, generically referred to as a RICH counter, contains separable radiator and photon detector regions (Séquinot and Ypsilantis 1977) (see also Chap. 6, "Particle Identification"). The detector and its optics constitute a "camera." Typically the camera optics map the Cherenkov cone onto (a portion of) a distorted "circle" at the photodetector. Though the imaging process is generally analogous to familiar imaging techniques used in telescopes and other optical instruments, there is a somewhat bewildering variety of methods used in a wide variety of counter types with different names. Some of the imaging methods used include (1) focusing by a lens; (2) proximity focusing (i.e., focusing by limiting the emission region of the radiation); and (3) focusing through an aperture (a pinhole). In addition, the prompt Cherenkov emission coupled with the speed of modern photon detectors allows the use of (4) time imaging, a method, which is little used in conventional imaging technology.

It should be noted that there are additional Cherenkov imaging device types that partially bridge the gap between these two, but that are not always called RICH detectors. As a particular example, the imaging atmospheric Cherenkov telescopes use the atmosphere as a very large radiator to convert high-energy gamma rays into electromagnetic showers. These are observed by a ground-level Cherenkov telescope, which measures the position and energy of the shower origin, and provides some information about the type of particle initiating the shower (RICH Workshop Series).

Typical RICH detectors are usually components of a larger detector in an accelerator physics experiment. In a simple model of such a RICH, the fractional error on the particle velocity (δ_β) is given by

$$\delta_\beta = \frac{\sigma_\beta}{\beta} = \tan\theta_c \, \sigma(\theta_c) \, ,$$

where

$$\sigma(\theta_c) = \frac{\langle \sigma(\theta_i) \rangle}{\sqrt{N_{\mathrm{pe}}}} \oplus C,$$

and $\langle \sigma(\theta_i) \rangle$ is the average single-photoelectron resolution, as defined by the optics, detector resolution, and the intrinsic chromaticity spread of the radiator index of refraction averaged over the photon detection bandwidth. C combines a number of other contributions to resolution including, (1) correlated terms such as tracking, alignment, and multiple scattering, (2) hit ambiguities, (3) background hits from random sources, and (4) hits coming from other tracks. The actual separation performance is also limited by physics effects such as decays in flight and particle interactions in the material of the detector. In many practical cases, the performance is limited by these effects.

For a $\beta \approx 1$ particle of momentum (p) well above threshold entering a radiator with index of refraction (n), the number of σ separation (N_σ) between particles of mass m_1 and m_2 is approximately

$$N_\sigma \approx \frac{\left|m_1^2 - m_2^2\right|}{2p^2\sigma(\theta_c)\sqrt{n^2-1}}.$$

In practical counters, the angular-resolution term $\sigma(\theta_c)$ varies between about 0.1 and 5 mrad depending on the size, radiator, and photodetector type of the particular counter. The range of momenta over which a particular counter can separate particle species extends from the point at which the number of photons emitted becomes sufficient for the counter to operate efficiently as a threshold device (~20% above the threshold for the lighter species) to the value in the imaging region given by the equation above. For example, for $\sigma(\theta_c) = 2$ mrad, a fused-silica radiator ($n = 1.47$), or a fluorocarbon gas radiator (C_5F_{12}, $n = 1.0017$), would separate π/K's from the threshold region starting around 0.15 (3) GeV/c through the imaging region up to about 4.2 (18) GeV/c at better than 3σ.

Many different imaging counters have been built during the last several decades (RICH Workshop Series). Among the earliest examples of this class of counters are the very-limited-acceptance Differential Cherenkov detectors, designed for particle selection in high-momentum beam lines. These devices use optical focusing and/or geometrical masking to select particles having velocities in a specified region. With careful design, a velocity resolution of $\sigma_\beta/\beta \approx 10^{-4}$ to 10^{-5} can be obtained (Litt and Meunier 1973).

Practical multi-track RICH counters in accelerator particle physics detectors are a more recent development, which have had substantial impact as PID detector in particle physics during the last two decades. RICH counters have used a variety of different radiators, imaging arrangements, and photon detectors that will be described in more detail below.

5 Examples of Cherenkov Counters

Cherenkov counters play an important role in many modern experiments in particle physics, nuclear physics, and particle astrophysics. This section presents a brief review of Cherenkov detectors used in current or past experiments as well as those planned in the near future. Rather than attempting a comprehensive review, a few select detectors will be described that can be considered as representative of a class of Cherenkov counter, with a focus on counters which may be either historically significant or state of the art.

5.1 Accelerator-Based Particle Identification Detectors

The typical use of Cherenkov counters in detectors at accelerators is to identify and thus to separate hadronic particle types. A closely related use is to separate electrons from hadrons in a hadron-rich environment such as a heavy-ion collider like RHIC (Tserruya 2006).

5.1.1 Threshold Cherenkov Counters

Among the earliest applications of Cherenkov counters for hadronic PID were threshold counters. The Aerogel Cherenkov Counter (ACC) of the Belle experiment (Sumiyoshi et al. 1999) is one of the most complex threshold counters to date. The Belle detector at the asymmetric KEKB e^+e^- collider studies the decays of particles produced on or near the $\Upsilon(4S)$ resonance. For the primary physics goal of Belle, the measurement of CP violation, excellent pion/kaon separation in hadronic decays of B mesons is crucial.

The radiator for the Belle ACC is silica aerogel. Fine-mesh PMTs are used for photon detection in the 1.5 T magnetic field of the Belle solenoid. Due to the asymmetric beam energies, and the resulting correlation between particle momentum and polar angle of the decay products from B mesons, the Cherenkov thresholds for different particle species vary as a function of the polar angle. Therefore, the best separation is obtained by carefully selecting the refractive indices in each kinematic region so that most pions produce Cherenkov radiation while most kaons are below Cherenkov threshold. The resulting optimized refractive indices of the aerogel go from $n = 1.028$ in the backward part of the barrel up to $n = 1.01$ in the forward region of the barrel ACC, and $n = 1.03$ in the endcap ACC.

During more than 10 years of successful operation, the Belle ACC achieved a kaon identification efficiency of up to 90% with a pion misidentification probability as low as 6% (Iijima et al. 2000).

5.1.2 Imaging Cherenkov Counters: RICH

Imaging RICH (see also Chap. 6, "Particle Identification") counters are sometimes classified by "generations" that differ based on historical timing, performance, design, and photodetection techniques.

Prototypical examples of first-generation RICH counters are those used in the DELPHI and SLD detectors at the LEP and SLC Z factory e^+e^- colliders (RICH Workshop Series). In both cases the RICH was designed to efficiently identify charged particles with momenta from about 0.25 to 20 GeV/c. This large momentum range required the use of two types of radiators, liquid (C_6F_{14}, $n = 1.276$) and gas (C_5F_{12}, $n = 1.0017$), the former being proximity imaged with the latter using mirrors. The phototransducers are a TPC/wire-chamber combination. They are made sensitive to photons by doping the TPC gas (usually, ethane/methane) with ~0.05% TMAE (tetrakis(dimethylamino)ethylene). Great attention to detail is required, (1) to avoid absorbing the UV photons to which TMAE is sensitive, (2) to avoid absorbing the single photoelectrons as they drift in the long TPC, and (3) to keep the chemically active TMAE vapor from interacting with materials in the system. In spite of their unforgiving operational characteristics, these counters attained good $e/\pi/K/p$ separation over the wide momentum ranges during several years of operation at LEP and SLC (Joram et al. 1999; Va'vra et al. 1999). Related but smaller-acceptance devices include the OMEGA RICH at the CERN SPS, and the RICH in the balloon-borne CAPRICE detector (RICH Workshop Series).

Later-generation counters generally operate at much higher rates, with more detection channels, than the first-generation detectors just described. They also utilize faster, more forgiving photon detectors, covering different photon detection bandwidths. Radiator choices have broadened to include materials such as lithium fluoride, fused silica, and aerogel.

Vacuum-based photodetection systems (e.g., single- or multi-anode PMTs, MCP-PMTs, or hybrid photodiodes (HPD)) have become increasingly common. They handle high rates, and can be used with a wide choice of radiators. Examples include (1) the SELEX RICH at Fermilab, which mirror focuses the Cherenkov photons from a neon radiator onto a camera array made of ~2,000 PMTs to separate hadrons over a wide momentum range (to well above 200 GeV/c for heavy hadrons) (Engelfried et al. 2003); (2) the HERMES RICH at HERA, which mirror focuses photons from C_4F_{10} (n = 1.00137) and aerogel (n = 1.0304) radiators within the same volume onto a PMT camera array to separate hadrons in the momentum range from 2 to 15 GeV/c (De Leo 2008); and (3) the LHCb detector at the LHC (Harnew 2008). It uses two separate counters. One volume, like HERMES, contains two radiators (aerogel and C_4F_{10}) while the second volume contains CF_4. Photons are mirror focused onto detector arrays of HPDs to cover a π/K separation momentum range between 1 and 150 GeV/c.

Other fast detection systems that use solid cesium-iodide (CsI) photocathodes or TEA doping in proportional chambers are useful with certain radiator types and geometries. Examples include (1) the CLEO-III RICH at CESR that uses a LiF radiator with TEA-doped proportional chambers (Sia 2005); (2) the ALICE detector at the LHC that uses proximity-focused liquid (C_6F_{14}) radiators and solid CsI photocathodes (Molnar et al. 2008) (similar photodetectors have been used for several years by the HADES and COMPASS detectors), and the hadron blind detector (HBD) in the PHENIX detector at RHIC that couples a low-index CF_4 radiator to a photodetector based on electron multiplier (GEM) chambers with reflective CsI photocathodes (Tserruya 2006).

A DIRC (Detection [of] Internally Reflected Cherenkov [light]) is a distinctive, compact RICH subtype first used in the BaBar detector (Adam et al. 2005). A DIRC "inverts" the usual RICH principle for use of light from the radiator by collecting and imaging the total internally reflected light rather than the transmitted light. It utilizes the optical material of the radiator in two ways, simultaneously; first as a Cherenkov radiator, and second, as a light pipe. The magnitudes of the photon angles are preserved during transport by the flat, rectangular-cross-section radiators, allowing the photons to be efficiently transported to a detector outside the path of the particle where they may be imaged in up to three independent dimensions (the usual two in space and, due to the long photon paths lengths, one in time). Because the index of refraction in the radiator is large ($n \sim 1.47$ for fused silica), the momentum range with good π/K separation goes up to 4–5 GeV/c although it is plausible, but difficult, to extend it up to about 10 GeV/c with an improved design.

The BaBar experiment at the asymmetric PEP-II e^+e^- collider studied CP violation in $\Upsilon(4S)$ decays. Excellent pion/kaon separation for particle momenta up to 4 GeV/c was required. The BaBar DIRC used 4.9 m long, rectangular bars made from synthetic fused silica as radiator and light guide. The photons were imaged via a "pin-hole" through an expansion region filled with 6,000 L of purified water onto an array of 10,752 densely packed photomultiplier tubes placed at a distance of about 1.2 m from the bar end. During more than 8 years of operation, the BaBar DIRC achieved π/K separation of 2.5 standard deviations or more up to the 4 GeV/c momentum. For a pion identification rate around 85%, the DIRC provided a kaon misidentification rate well below 1% up to 3 GeV/c (Adam et al. 2005).

New DIRC detectors are being developed that take advantage of the new, very fast, pixelated photodetectors becoming available, such as flat panel PMTs and MCP-PMTs. They typically utilize either time imaging or mirror-focused optics, or both, leading not only to a precision measurement of the Cherenkov angle, but in some cases, to a precise measurement of the

particle's time of flight, and/or for the correction of the chromatic dispersion in the radiator. Examples include: (1) the time-of-propagation (TOP) counter being developed for the Belle II upgrade at KEKB, which emphasizes precision timing for both Cherenkov imaging and TOF (Inami et al. 2008); (2) the fully three-dimensional imaging FDIRC for the SuperB detector at the Italian SuperB collider, which uses precision timing not only for improving the angle reconstruction and TOF, but also to correct the chromatic dispersion (Schwiening et al. 2008); and (3) the barrel and endcap DIRCs being developed for the PANDA detector at FAIR that use focusing optics and fast timing (Föhl et al. 2008).

5.2 Astroparticle Physics

A diverse range of Cherenkov counters is found in astroparticle physics experiments that study neutrinos, proton decays, γ rays, and charged cosmic rays. From neutrino detectors located deep underground to anti-matter searches at the International Space Station, the range includes large arrays of air shower detectors and strings of photon detectors submersed in deep lakes or the polar ice cap, imaging devices like the stereoscopic air Cherenkov telescopes and large-volume tracking calorimetric devices used in the study of nucleon decays or neutrino interactions.

5.2.1 Underground Neutrino Detectors

The study of neutrinos requires large detector volumes to improve the likelihood of observing a neutrino interaction inside the active volume. Purified water or heavy water is often used as an inexpensive Cherenkov radiator with large hemispherical PMTs as photon detectors. The radiator has to be shielded from muons produced in the atmosphere by high-energy cosmic radiation, which can be realized by placing the detector deep underground.

The first massive Cherenkov imaging detector of this type was the 10 kt IMB (Becker-Szendy et al. 1993), which began operations in an Ohio, USA, salt mine in 1982, motivated both by the search for neutrino oscillations and proton decay. The even larger Super-Kamiokande (Super-K) experiment (Fukuda et al. 2003) is placed in the Kamioka Mozumi mine in Japan at 1,000 m depth. The active volume is 50,000 t of pure water in a cylindrical stainless steel tank, 39 m in diameter and 41 m in height. The detector is sensitive to decay products from nucleon decay; neutrinos from the sun, the atmosphere, and extra-terrestrial sources; as well as cosmic rays. To distinguish these contributions, the Super-K detector is divided into an inner fiducial volume of 22,000 t, viewed with some 11,000 large 51 cm PMTs and a 28,000 t outer zone, viewed by some 1,900 20 cm PMTs.

A neutrino can interact with the electrons in the water target and transfer enough energy that the electrons will produce Cherenkov radiation. The number of photons produced provides a measure of the energy of the Cherenkov-light-emitting particles. Differences in the responses of the inner and outer detector as well as the sharpness of the ring image and track lengths are used to separate electrons from muons and thus, for instance, events caused by neutrinos produced in the earth beneath the detector from muons produced by cosmic rays in the atmosphere.

The Super-K experiment has been in operation since 1996 and has made measurements of neutrino oscillations and provided stringent limits to the proton lifetime. The detector has been

upgraded several times and, since 2009, is also used for detecting man-made neutrinos as part of the T2K experiment to study long-baseline neutrino oscillations (Trung Le: Le (2009)).

Another recent detector of this type is the SNO detector located at a depth of 2,000 m in the Creighton mine near Sudbury, Canada (Boger et al. 2000). SNO uses an acrylic spherical fiducial vessel filled with 1,000 t of heavy water, situated within a 30 m barrel-shaped cavity filled with normal water and the photon detectors. The SNO device is unique in that it is sensitive to all three neutrino types – thereby providing precise measurements of the rates and flavors of solar neutrinos that reach Earth.

5.2.2 Neutrino Detectors in Natural Water or Ice

When instrumented with sensitive photon detectors, naturally occurring water in the sea, lakes, or ice allows the creation of neutrino experiments with active volumes several orders of magnitude larger than those used by the underground water Cherenkov detectors. Strings of photon detectors are suspended in the active volume to form a three-dimensional matrix of space and time coordinates used to reconstruct the Cherenkov image and reject backgrounds.

The DUMAND detector proposed for the deep ocean near Hawaii was the earliest of these detector concepts (Roberts 1992), and though significant prototyping occurred, a complete detector was never built. ANTARES (Astronomy with a Neutrino Telescope and Abyss environmental RESearch) is the first neutrino telescope constructed in the deep sea (Bertin et al. 2009). It is located in the Mediterranean Sea 40 km off the coast of Southern France at a depth of 2,400 m. Twelve lines of photon detectors are anchored to the sea floor, each comprising 75 large (25 cm diameter) PMTs, looking downward at an angle of 45° to be sensitive to Cherenkov light from upward-going muon tracks produced in interactions of extraterrestrial neutrinos after traversing the Earth. The PMTs are contained in optical modules, 43 cm diameter glass pressure spheres, where they are shielded from the Earth's magnetic field by a mu-metal cage. Each line is divided into 25 stories, 14.5 m apart, with three optical modules at each story. At a pitch of 60–70 m, the lines cover an area of 0.1 km^2 with the PMTs suspended between 100 and 450 m above the sea floor. The spacing between the optical modules is driven by the transparency of the water at 2,400 m depth (Aguilar et al. 2005). Installation of the lines started in 2006 and was completed in 2008.

The observed background of 0.1 kHz cm^{-2} is dominated by bioluminescence from microorganisms, radioactive decay of ^{40}K, and short bursts of bioluminescence from macroorganisms. Readout and triggering are based on an all-data-to-shore concept. The PMT signals are digitized and sent via multiplexed gigabit links to a computer farm on shore, where software reduces the trigger rate to typically 5–10 Hz.

Analysis of the data based on five lines of photon detectors has shown that the flux of atmospheric muons agrees with the expectation from Monte Carlo simulation (Aguilar et al. 2010).

ANTARES is a first step toward the construction of a Cherenkov detector with an active volume of a cubic kilometer in the Mediterranean Sea as part of the KM3NeT project (Bagley et al.). The data obtained by ANTARES and KM3NeT are complementary to the results obtained by the AMANDA (Wischnewski et al. 1999) and IceCube (Achterberg et al. 2006) experiments, located at the South Pole, since ANTARES and KM3Net are sensitive to neutrino sources in the southern hemisphere while the Antarctic experiments will cover sources in the northern hemisphere.

5.2.3 Imaging Air Cherenkov Telescopes

In order to study very-high-energy cosmic rays, the atmosphere can be used as the Cherenkov radiator, while the radiation is imaged onto single-photon-sensitive detector planes of telescopes with large parabolic mirrors. Such ground-based Imaging Air Cherenkov Telescopes (IACTs) (see also ❯ Chap. 23, "Astrophysics and Space Instrumentation") can be sensitive to an energy range from a few GeV to more than 100 TeV. A γ ray interacting in the atmosphere will produce an air shower of secondary particles at an elevation of 10–15 km. The shower particles produce Cherenkov light and the resulting light cone will cover a disk with a radius of 100–120 m at the ground level with an intensity of 10–100 photons m^{-2}. A telescope looking up at the night sky will be able to detect the Cherenkov light and measure the intensity, orientation, and shape of the air shower, which are related to the primary energy and direction of the γ ray. Among the pioneering detectors of this type was the 10 m WHIPPLE telescope located at over 2,000 m in elevation in Arizona, USA (Krenrich et al. 1998). Substantially improved performance can be obtained with systems of several telescopes that provide multiple views of the same air shower.

A number of these stereoscopic telescope systems are now in operation, including VERITAS (Holder et al. 2006), H.E.S.S. (Bernlöhr et al. 2003), MAGIC (Cortina et al. 2009), and CANGAROO (Kubo et al. 2004). As an example, the H.E.S.S. (High Energy Stereoscopic System) experiment is an array of four IACTs with 13 m diameter telescopes situated in the Khomas Highland in Namibia at an altitude of 1,800 m above sea level and has been in operation since 2003. Each telescope consists of 382 round mirror facets of 60 cm diameter, made of aluminized glass with a quartz coating, focusing the light on a focal plane equipped with 960 PMTs. The combination of the images recorded by the four telescopes allows stereoscopic viewing of the air shower. This improves the sensitivity, the angular and energy resolution of the experiment, and provides better background rejection and dead time.

H.E.S.S. has performed surveys of the galactic plane, discovering 70 new sources of very-high-energy ($E > 1$ TeV) γ rays (Chaves et al. 2009). An upgrade of the H.E.S.S. experiment is under way. The mirrors of the four IACT telescopes will be refurbished by applying new coatings and a fifth telescope will be added at the center of the array. At 600 m^2 mirror area, this telescope will more than double the active area of the experiment, becoming the largest Cherenkov telescope to date. It is expected to be ready for operation in 2012.

6 Conclusions

Cherenkov light was first observed, although not immediately understood, more than 100 years ago, and fully explicated experimentally and understood theoretically within classical electromagnetic theory some three to four decades later. Today, more than 50 years after the first Cherenkov counters became operational, new detectors that exploit the special characteristics of Cherenkov radiation continue to be developed.

Important properties of the radiation that are utilized in varying ways by these devices include the rapid emission of a sharp light pulse; the existence of a velocity threshold for radiation; the direct dependence between the track length and the number of photons emitted; and the dependence of the Cherenkov cone half-angle θ_c and the number of emitted photons on the velocity of the particle and the refractive index of the medium. Moreover, useful numbers of photons are emitted in the visible and UV photon range where optical materials and large natural radiators, such as water and the atmosphere, are transparent, and where highly efficient photodetectors have been developed.

Cherenkov detectors have benefited substantially from numerous technological advances over the last half-century. The first photomultiplier tubes, which allowed efficient photon counting, were especially crucial to the early adoption of Cherenkov devices and continue to be developed. Modern PMTs are much more efficient, have better single-photon counting characteristics, come in many different sizes and shapes, and include a number of very fast, pixelated types. Additional photodetection options include gas detectors and solid-state devices such as Geiger-mode APDs. Radiator options have expanded due to the development of modern materials and industrial applications, often for high-technology purposes. These include very transparent fused silica, fluorocarbons, aerogels with a refractive index that can be tuned to a wide range of applications, and the ability to clean large volumes of water.

Cherenkov detectors are now found in a wide variety of unique applications throughout physics, astrophysics, and biomedicine, with more powerful, and/or larger devices continuing to be developed and implemented. Particular examples include the many detectors at particle accelerators that rely on powerful imaging detectors for hadronic particle identification, the large water Cherenkov detectors used for neutrino detection both for astrophysics and accelerator studies, and the imaging air Cherenkov telescopes used to study very-high-energy γ rays in cosmic radiation.

Acknowledgment

Work supported in part by the US Department of Energy under contract number DE-AC02-76SF00515.

References

Abashian A et al (2002) The Belle detector. Nucl Instrum Methods A 479:117

Achterberg A et al (2006) First year performance of the IceCube neutrino telescope. Astropart Phys 26:155

Adam I et al (2005) The DIRC particle identification system for the BABAR experiment. Nucl Instrum Methods A 538:281

Aguilar JA et al (2005) Transmission of light in deep sea water at the site of the ANTARES neutrino telescope. Astropart Phys 23:131

Aguilar JA et al (2010) Measurement of the atmospheric muon flux with a 4 GeV threshold in the ANTARES neutrino telescope. Astropart Phys 33:86

Amsler C et al (2008) Particle Data Group. Phys Lett B 667:1

Ashie Y et al (2005) Measurement of atmospheric neutrino oscillation parameters by Super-Kamiokande I. Phys Rev D 71:112005

Bagley P et al. KM3NeT conceptual design report. http://www.km3net.org. ISBN 978-90-6488-031-5

Bartlett D et al (1987) Performance of the Cherenkov counters in the fermilab tagged photon spectrometer facility. Nucl Instrum Methods A 260:55

Becker-Szendy R et al (1993) IMB-3: a large water Cherenkov detector for nucleon decay and neutrino interactions. Nucl Instrum Methods A 324:363

Bernlöhr K, Carrol O, Cornils R et al (2003) The optical system of the H.E.S.S. imaging atmospheric Cherenkov telescopes, Part I: layout and components of the system. Astropart Phys 20:111; Funk S, Hermann G, Hinton J et al (2004) The trigger system of the H.E.S.S. telescope array. Astropart Phys 22:285

Bertin V et al (2009) Status and first results of the ANTARES neutrino telescope. Nucl Instrum Methods A 604:136

Boger J et al (2000) The Sudbury Neutrino Observatory. Nucl Instrum Methods A 449:172

Chaves RCG et al (2009) Extending the H.E.S.S. Galactic Plane Survey. In: Proceedings of the 31st international cosmic ray conference,

Lodz. http://www.mpi-hd.mpg.de/hfm/HESS/. arXiv:0907.0768v1 [astro-ph.HE]
Cherenkov P (1934) Visible emission of clean liquids by action of gamma radiation. Dokl Akad Nauk SSSR 2:451
Cho JS et al (2009) Cerenkov radiation imaging as a method for quantitative measurements of beta particles in a microfluidic chip. Phys Med Biol 54:6757
Cortina J et al (2009) Technical performance of the MAGIC telescopes. In: Proceedings of the 31st international cosmic ray conference, Lodz. http://magic.mppmu.mpg.de/. arXiv:0907.1211v1 [astro-ph.IM]
Curie E (2001) Madame Curie. Da Capo, Cambridge, MA, 444 pp
De Leo R (2008) Long-term operational experience with the HERMES aerogel RICH detector. Nucl Instrum Methods A 595:19
Ecklund S et al (2001) A fast luminosity monitor system for PEP II. Nucl Instrum Methods A 463:68
Engelfried J et al (2003) SELEX RICH performance and physics results. Nucl Instrum Methods A 502:285
Föhl K et al (2008) The DIRC detectors of the PANDA experiment at FAIR. Nucl Instrum Methods A 585:88
Frank I, Tamm I (1937) Coherent visible radiation of fast electrons passing through matter. Dokl Akad Nauk 14:109
Fukuda Y et al (2003) The Super-Kamiokande detector. Nucl Instrum Methods A 501:418
Harnew N (2008) An overview of the status of the LHCb RICH detectors. Nucl Instrum Methods A 595:31
Holder J et al (2006) The first VERITAS telescope. Astropart Phys 25:391
Iijima T et al (2000) Aerogel Cherenkov counter for the Belle detector. Nucl Instrum Methods A 453:321; Nakano E (2002) Belle PID. Nucl Instrum Methods A 494:402
Inami K et al (2008) Development of a TOP counter for the Super B factory. Nucl Instrum Methods A 585:96
Jackson JD (1962) Classical electrodynamics, 1st edn. Wiley, Hoboken
Jelley JV (1958) Cherenkov radiation. Pergamon, New York
Joram C et al (1999) Operation, optimisation, and performance of the DEL-PHI RICH detectors. Nucl Instrum Methods A 433:47
Krenrich F et al (1998) Stereoscopic observations of gamma rays at the Whipple observatory. Astropart Phys 8:213
Kubo H et al (2004) Status of the CANGAROO-III Project. New Astron Rev 48:323
Le T (2009) Overview of the T2K long baseline neutrino oscillation experiment. arXiv:0910.4211 [hep-ex]; http://jnusrv01.kek.jp
Litt J, Meunier R (1973) Cerenkov counter technique in high-energy physics. Annu Rev Nucl Sci 23:1
Molnar L et al (2008) The ALICE HMPID detector ready for collisions at the LHC. Nucl Instrum Methods A 595:27
Motz H, Schiff LI (1953) Cerenkov radiation in a dispersive medium. Am J Phys 21:258
Ratcliff B (2003) Imaging rings in Ring Imaging Cherenkov counters. Nucl Instrum Methods A 502:211
RICH Workshop Series: Nucl Instrum Methods A 343:1 (1994); Nucl Instrum Methods A 371:1 (1996); Nucl Instrum Methods A 433:1 (1999); Nucl Instrum Methods A 502:1 (2003); Nucl Instrum Methods A 553:1 (2005); Nucl Instrum Methods A 595:1 (2008)
Roberts A (1992) The birth of high-energy neutrino astronomy: a personal history of the DUMAND project. Rev Mod Phys 64:259
Schwiening J et al (2008) Status of the fast Focusing DIRC (fDIRC). Nucl Instrum Methods A 585:104
Séquinot J, Ypsilantis T (1977) Photoionization and Cherenkov Ring Imaging. Nucl Instrum Methods A 142:377
Sia R (2005) Performance of the LiF-TEA ring imaging Cherenkov detector at CLEO. Nucl Instrum Methods A 553:323
Sumiyoshi T et al (1999) Silica aerogel Cherenkov counter for the KEK B-factory experiment. Nucl Instrum Methods A 433:385
Tamm I (1939) Radiation emitted by uniformly moving electrons. J Phys USSR 1:439
Tserruya I (2006) Report on the Hadron Blind Detector for the PHENIX experiment. Nucl Instrum Methods A 563:333
Va'vra J et al (1999) Long-term operational experience with the barrel CRID at SLD. Nucl Instrum Methods A 433:59
Va'vra J et al (2009) Beam test of a Time-of-Flight detector prototype. Nucl Instrum Methods A 606:404
Wischnewski R et al (1999) The AMANDA Neutrino detector. Nucl Phys Proc Suppl 75A:312

Further Reading

Buckley J et al (2008) The status and future of ground-based TeV gamma-ray astronomy: a White Paper prepared for the Division of Astrophysics of the American Physical Society. arXiv:0810.0444v1 [astro-ph]

Cherenkova EP (2008) The discovery of the Cherenkov radiation. Nucl Instrum Methods A 595:8

Hallewell GD (2005) The status of Cherenkov detectors in astroparticle physics. Nucl Instrum Methods A 553:242

Križan P (2009) Advances in particle-identification concepts. J Instrum 4:P11017

Ratcliff B (2008) Advantages and limitations of the RICH technique for particle identification. Nucl Instrum Methods A 595:1

Review of Particle Properties: Amsler C et al, Particle Data Group (2008) Phys Lett B 667:1

Websdale DM (2008) Review of Cherenkov imaging devices in particle and nuclear physics experiments. Nucl Instrum Methods A 595:12

19 Muon Spectrometers

Thomas Hebbeker · Kerstin Hoepfner
RWTH Aachen University, Aachen, Germany

1 Introduction ***474***

2 General Considerations ***474***

3 Magnetic Spectrometers ***475***
3.1 Magnets 476
3.2 Tracking Detectors 477

4 Muon Detectors at Accelerator-Based Experiments ***478***
4.1 Drift-Tube Detectors 482
4.2 Resistive-Plate Chambers (RPC) 484
4.3 Multi-Wire Chambers 486

5 Muon Spectrometers for Cosmic Ray Measurements ***487***
5.1 Atmospheric Muon Detectors 489
5.2 Air Shower Detector Arrays 492

6 Muon Radiography ***493***

7 Conclusions ***494***

References ***495***

C. Grupen, I. Buvat (eds.), *Handbook of Particle Detection and Imaging*,
DOI 10.1007/978-3-642-13271-1_19, © Springer-Verlag Berlin Heidelberg 2012

Abstract: The detection of muons and the measurement of their momenta is an important task both in astroparticle physics and in elementary particle physics. In cosmic ray physics, the need for muon detectors is obvious since atmospheric showers consist mainly of muons when reaching the surface of the earth. In accelerator-based experiments, the muon plays a special role as a long-lived particle with only electromagnetic interactions; it can easily be identified, and muons provide in many theoretical models a characteristic signature for new physics. A muon spectrometer consists of a position-sensitive detector that records tracks of charged particles, and a magnetic field so that charge and momentum can be deduced from the track curvature. The identification of muons often relies on the large amount of absorber material in front of the muon detector, which allows only muons (and neutrinos) to pass. We first describe the general detector layout and discuss the related uncertainties. Then we present several examples of muon spectrometers that were or are successfully operated in accelerator physics or in cosmic ray physics. We include also muon detectors without a magnetic field. Finally, we report on the application of muon detectors outside particle and astroparticle physics.

1 Introduction

Muon spectrometers play a central role in particle and cosmic ray physics. Pioneering experiments used muon spectrometers for the measurement and identification of elementary particles. Almost all modern general-purpose detectors at electron–positron or proton colliders use muon spectrometers of multifunctional design with various magnetic field configurations.

2 General Considerations

Muons can be **detected** via their electromagnetic interactions in matter (see Chap. 1, "Interactions of Particles and Radiation with Matter"):

- In modern detectors, often the ionization of gas molecules (Blum et al. 2008) or the creation of electron–hole pairs in semiconductors (Spieler 2005) is exploited to recognize the passage of a muon and to measure at the same time space points along the trajectory. A resolution of 10 μm to some 100 μm can be achieved, while the detection efficiency is near 100%. Gaseous detectors have been used for particle detection in many variations, from Geiger–Müller counters to spark chambers and multi-channel drift chambers. In earlier experiments the ionization of fluids along a particle's trajectory was made visible in Wilson and bubble chambers.
- Also chemical processes can be exploited, notably in nuclear emulsions; this technique is still used today (OPERA Coll 2010), since it provides a three-dimensional spatial resolution of the order of a μm.
- Scintillators yield light through excitation and de-excitation of molecules, see Chap. 15, "Scintillation Counters". In some applications like veto counters, only a very crude position measurement is provided, while modern scintillating fiber detectors allow for resolutions of better than 100 μm.

- Cherenkov and transition radiation are also suitable to detect relativistic muons, see Chap. 18, "Cherenkov Counters". Examples are water tanks of air shower observatories like Auger (Pierre Auger Coll 2004) or the transition radiation tracker (ATLAS Coll 2008) in the ATLAS detector.

Since we focus here on spectrometers which need a precise tracking of muons, mainly gaseous and semiconductor detectors are considered in the following. We concentrate on relativistic muons and on detectors specializing on muon measurements. See Chap. 11, "Gaseous Detectors" and Chap. 12, "Tracking Detectors" for more details on these detection methods.

Relativistic muons can be *identified* in different ways. The most suitable property distinguishing muons from other long-lived charged particles is their relatively small energy loss in matter. The reach of electrons and hadrons is limited by the corresponding shower lengths of $\sim 10 X_0$ and $\sim 20 \Lambda$; for water, this translates into 4 and 17 m. Muons on the other hand lose in water only about

$$\frac{\mathrm{d}E}{\mathrm{d}x} \sim 0.2 \frac{\mathrm{GeV}}{\mathrm{m}} \tag{1}$$

through ionization and can penetrate several km for momenta in the TeV regime.

In accelerator experiments, the electromagnetic and hadronic calorimeters, and additional absorber material like iron of the magnet yoke, stop nearly all charged particles before reaching the outer muon chambers. Cosmic ray experiments are often carried out in underground caverns, thus the overburden provides the required shielding.

The *momentum* and simultaneous *charge* measurement of muons require the combination of a magnetic field and a tracking detector. We will discuss the resolution achievable with these muon spectrometers in Sect. 3.

The energy of a high-energy muon (but not its charge) can also be determined from the amount of multiple scattering when passing a material of thickness l with radiation length X_0 (Particle Data Group 2008):

$$\theta^{\mathrm{rms}} \approx \frac{14\,\mathrm{MeV}}{p} \sqrt{\frac{l}{X_0}}. \tag{2}$$

In this formula, θ^{rms} is the r.m.s. width of the scattering-angle distribution projected onto a plane containing the incoming particle direction.

For ultra-relativistic muons, the amount of bremsstrahlung accompanying the track is a measure of its energy. Above the critical energy of (Grupen and Shwartz 2008; Particle Data Group 2008)

$$E_c \approx \frac{25\,\mathrm{TeV}}{Z+1} \tag{3}$$

(for solids) bremsstrahlung becomes the dominant energy-loss mechanism for muons.

3 Magnetic Spectrometers

Inside a homogeneous magnetic field, a particle with unit charge follows a circular path with radius

$$R = \frac{p_\perp}{e\,B} = 3.3\,\mathrm{m} \cdot \frac{p_\perp/(\mathrm{GeV}/c)}{B/\mathrm{T}}, \tag{4}$$

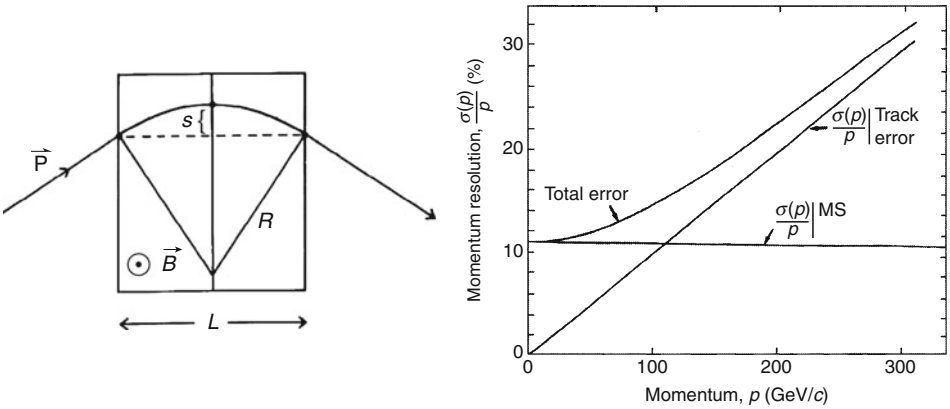


Fig. 1
***Left*: Sagitta measurement in magnetic spectrometer. *Right*: Momentum resolution as a function of momentum (for component perpendicular to magnetic field) (Grupen and Shwartz 2008)**

where $p_\perp$ is the momentum component perpendicular to the B field. The direction of the bending depends on the sign of the charge. Since momentum components parallel to the B field are not affected, the trajectory will be a helix in general.

Equation 4 is the basis for all magnetic spectrometers. In the following, we consider only the movement of the particle – muon or antimuon – in the plane perpendicular to the B field. A high-energy muon cannot be "trapped" in the magnetic field; it will enter the field region, follow a short arc of a circle with a large bending radius R, and leave the field in a slightly different direction, see *Fig. 1*. Often the sagitta s as defined in the figure is measured. To determine $p_\perp$ we need

- A strong and large magnet with a well-known field strength
- Several precise measurements of the particle's position along its trajectory inside and/or outside the magnetic field with (nondestructive) tracking detectors

And we have to keep the principal "adversary" under control:

- Multiple scattering in the materials traversed by the muon, faking magnetic bending.

3.1 Magnets

Dipoles, solenoids, and toroids are common geometries. For example, at the LHC, the ATLAS experiment uses a large air-core toroid (plus a smaller solenoid) and CMS a solenoid, see *Fig. 2*. LHCb with its fixed-target-like detector geometry relies on a dipole magnet.

Often superconducting coils are used, with B fields up to 4 T. The length of the field region can reach several meters and the stored energy up to 2.5 GJ (CMS Coll 2008). Magnetized iron

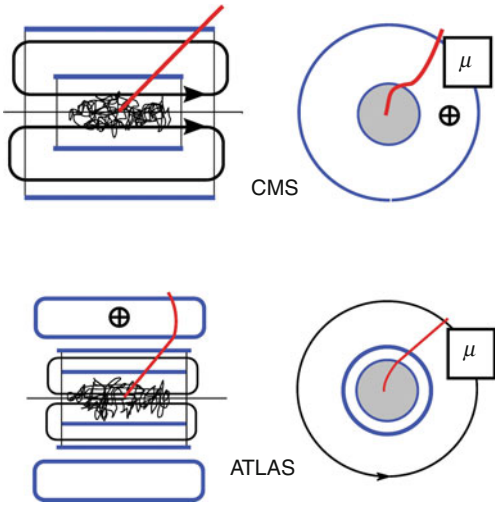


Fig. 2
***Top*: Bending in CMS solenoid, *bottom*: ATLAS toroid and solenoid (C. Mai, 2010, private communication)**

is sometimes used to increase the B field inside a conventional coil or to guide the field lines outside a solenoid (return yoke), see *Fig. 2* (CMS and ATLAS (via hadron calorimeter)).

3.2 Tracking Detectors

Space points along the muon's path can be measured precisely with silicon pixel or silicon strip trackers reaching a resolution of $\sigma_x \sim 20\,\mu\text{m}$ or drift chambers achieving typically $\sigma_x \sim 200\,\mu\text{m}$. The number of points N measured varies between 10 and 100 for most detectors. In addition, other constraints can be exploited, in particular the collision vertex, see *Fig. 2*. In Sect. 4, some tracking detectors are described in detail.

A nice summary of momentum measurements for different spectrometer geometries can be found in Grupen and Shwartz (2008). Here we discuss only the configuration as shown in *Fig. 1*, and assume that the N points were all measured in the magnetic volume of length L, with equidistant spacing and equal spatial resolution σ_x. The resulting momentum resolution $\sigma(p_\perp)^{\text{tracking}}$ was already calculated by R.L. Gluckstern about 50 years ago (Gluckstern 1963). For $N \gg 1$:

$$\frac{\sigma(p_\perp)^{\text{tracking}}}{p_\perp} = 89 \cdot 10^{-6} \cdot \frac{\sigma_x/\mu\text{m}}{B/\text{T} \cdot L^2/\text{m}^2 \cdot \sqrt{N+4}} \cdot \frac{p_\perp}{\text{GeV}/c}. \tag{5}$$

Example: In the CMS inner detector (CMS Coll 2008) (4 T solenoid plus silicon tracker of 1.2 m radius), a resolution of

$$\frac{\sigma(p_\perp)^{\text{tracking}}}{p_\perp} = 1.5 \cdot 10^{-4} \cdot \frac{p_\perp}{\text{GeV}/c} \tag{6}$$

can been reached. In addition, multiple scattering (MS) has to be included, see Eq. 2, so that the total momentum resolution, assuming the particle is moving perpendicular to the B field, is given by

$$\frac{\sigma(p)}{p} = \frac{\sigma(p)^{\text{tracking}}}{p} \oplus \frac{\sigma(p)^{\text{MS}}}{p}, \tag{7}$$

$$\frac{\sigma(p)^{\text{MS}}}{p} = \frac{0.055}{B/\text{T} \cdot \sqrt{L/\text{m}} \cdot \sqrt{X_0/\text{m}}}. \tag{8}$$

Here X_0 is the effective radiation length in the magnetic volume of length L. The relative importance of the two terms in Eq. 7 is illustrated schematically in *Fig. 1*. For very high momenta, the resolution is further degraded by bremsstrahlung, see Eq. 3.

Example: The muon spectrometer of the ATLAS experiment (ATLAS Coll 2008) consists of an air-core toroid, thus minimizing the MS term, and Monitored Drift Tube chambers (MDT) with a spatial resolution of better than 100 μm. The resulting momentum resolution for centrally produced muons can be parametrized as

$$\frac{\sigma(p)}{p} = 1.0 \cdot 10^{-4} \frac{p}{\text{GeV}/c} \oplus 0.02. \tag{9}$$

Including also the inner tracking systems inside the solenoid magnet the resolution is further improved, in particular at low momenta. In cosmic ray physics, often the maximum detectable momentum p_{mdm} is used to characterize the momentum resolution (Grupen and Shwartz 2008); it is defined as the momentum value for which the resolution equals its value: $\sigma(p_{\text{mdm}})/p_{\text{mdm}} = 1$.

4 Muon Detectors at Accelerator-Based Experiments

Muon spectrometers play a key role in accelerator-based experiments, since muons in the final state often constitute the "golden channel" in the search for new particles. Besides their main task of muon identification, precise muon momentum and charge measurements are required in order to allow for the reconstruction of invariant mass(es) and kinematics of the primary physics reaction. A popular example is the search for the Higgs boson decaying into four leptons via $H \to ZZ^{(*)} \to \ell^+\ell^-\ell^+\ell^-$ (with $\ell = \mu, e$), as seen in *Fig. 3*. Such events are reconstructed via the invariant mass of both Z bosons based on the measured lepton momenta, which are subsequently combined to reconstruct the Higgs invariant mass. Other potential new phenomena or particles, ranging from supersymmetry to extra dimensions to heavy vector bosons, are also searched for with muons in the final state.

To select interesting physics events, a muon level-1 trigger relies on muon identification and momentum determination. To provide the information within O(μs), detector response and readout have to be sufficiently fast. Rate capability and aging are a lesser problem than for inner detectors due to the lower occupancy as only muons should arrive. However, calorimeter leakage and punch through may occur (especially around pseudorapidity $\eta = 0$ where the particle's track length through the calorimeter is minimal), thus increasing the rates from O(1 Hz/cm^2) to O(10 Hz/cm^2). In muon spectrometers with iron, only the first station is affected as the iron acts as additional absorber.

For reactions such as the Higgs boson decaying into four muons, hermeticity and acceptance are key parameters as the efficiency goes with ϵ^4. In order not to align insensitive areas along a particle's path, stations are staggered as can be seen in the event display of *Fig. 3*.

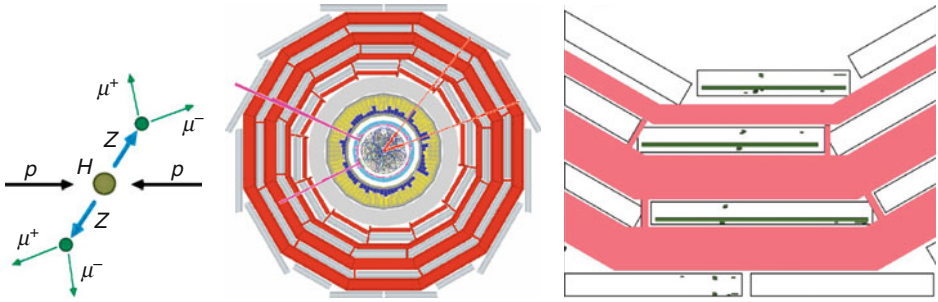


Fig. 3

Searching for the hypothetical Higgs boson, through its decay into four leptons ("golden channel"). On the *left*, an example for a simulated Higgs boson decaying into pairs of muons and electrons in the CMS detector is shown. The close-up on the *right* shows individual hits in the $r\phi$ projection in four consecutive muon stations allowing a precise stand-alone tracking. Horizontal lines represent the hits in the orthogonal *rz* projection which is not instrumented in the outermost station

An unavoidable exception is often the interface region between barrel and endcaps, resulting in a slightly reduced trigger efficiency for these pseudorapidities.

In this section, technology implementations for muon systems are discussed with the intent to illustrate their diversity on selected examples. Detectors at accelerators may be arranged in fixed-target geometry (layered structure perpendicular to the incident beam) or in cylindrical layers around a central collision point. The latter geometry is far more common at modern accelerators. Examples are: the Tevatron experiments CDF (CDF Coll 1996) and D0 (D0 Coll 2006), at CERN's LHC the ATLAS (ATLAS Coll 2008), ALICE (ALICE Coll 2008), and CMS (CMS Coll 2008) experiments, or the BELLE (BELLE Coll 2002) and BaBar (BaBar Coll 2002) detectors at the B-factories. The cylindrical layers around the interaction point (often referred to as "barrel" detectors) cover pseudorapidities up to $O(|\eta| \leq 1)$, and these are complemented by disks of "forward" detectors (also referred to as "endcap") extending the reach in pseudorapidity up to $O(|\eta| \leq 2.5)$, see ❯ *Fig. 11* for a typical example. Due to the lower center-of-mass energy, fixed-target experiments are rarely being built anymore, with the notable exception of neutrino experiments, OPERA (OPERA Coll 2010) in the CERN-to-Gran Sasso neutrino beam, and MINOS (MINOS Coll 2009) at the end of the long-baseline neutrino beam from Fermilab. Many earlier experiments around 1960–1980 were of fixed-target type, with the exception of e^+e^- machines, until the era of UA1 and UA2 at the CERN proton–antiproton collider. Although it is a colliding-beam experiment, LHCb (LHCb Coll 2008) is constructed in fixed-target geometry, since the strongly boosted B mesons, the subject of LHCb's research, fly mainly in the forward/backward direction with one of the hemispheres being instrumented. ❯ *Figure 4* shows the LHCb muon spectrometer in such a fixed-target arrangement, while ❯ *Fig. 5* displays the ATLAS muon spectrometer at the LHC in colliding-beam geometry. In LHCb, the active muon stations are interleaved with iron (also found in many other experiments such as D0, CMS, BaBar, Belle, and OPERA), an arrangement often implemented as it is convenient to insert the muon chambers in the return yoke of the magnet. In addition to the absorber function, the iron causes multiple scattering of the muons, thus limiting the overall resolution

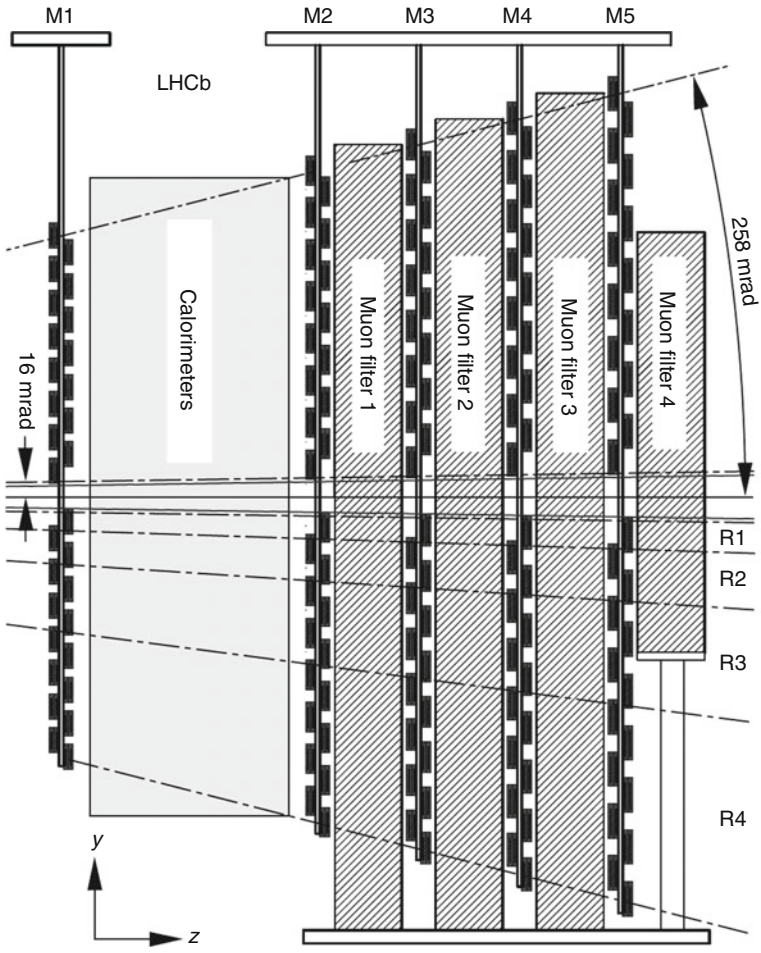


Fig. 4

The muon system of LHCb is built in a fixed-target geometry (LHCb Coll 2008) with increasing cell size as the distance from the interaction point increases. The muon chambers are interleaved with iron

(as discussed in Sect. 3). This is not the case for the ATLAS muon system with its air-core toroid, where the chamber resolution is actually the limiting parameter and, hence, has been maximized to yield the presently best stand-alone momentum resolution of a muon system.

Common to all detectors is the position of the muon system as the outermost subsystem, given that muons are the only charged particles passing several interaction lengths of material with little energy loss. Consequently, identification of muons based on their reach is the main task of the muon system. Before the start of the LHC era, a signal based on 1, 2, or 3 hits in muon detectors was sufficient for many experiments. Momentum measurement and tracking were provided by the inner tracker. Such an example is shown in *Fig. 6* in form of the D0 muon system (D0 Coll 1997, 2006) using layers of $1/2''$ thick scintillator planes with PMT readout

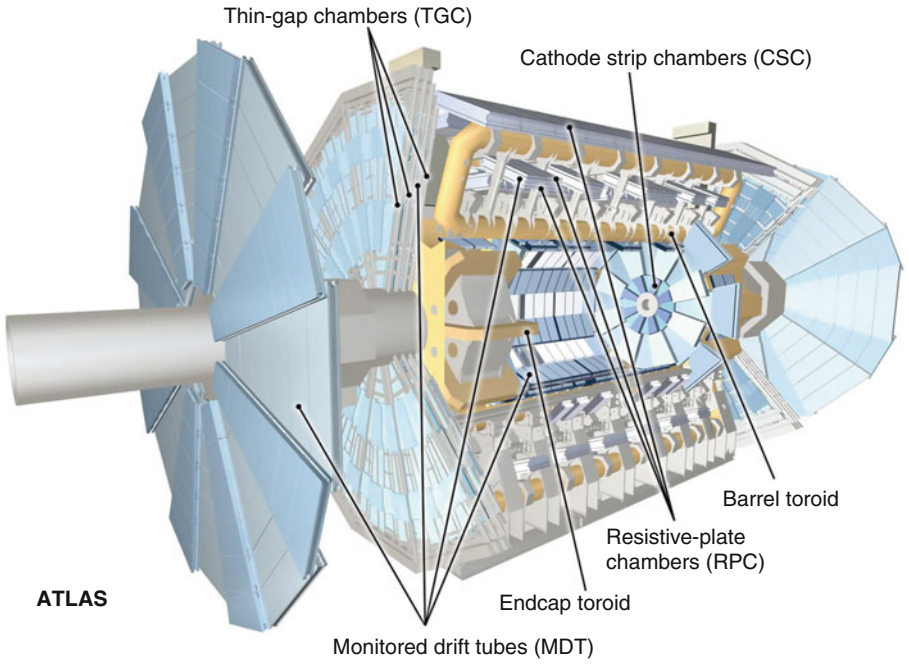


Fig. 5
The ATLAS detector with its muon system (ATLAS Coll 2008) exploiting several detection technologies based on gaseous detectors

along with layers of drift-tube chambers, either mini drift tubes or proportional drift tubes. Like in many systems, a muon from the interaction point can pass up to three stations. Below the detector, regions of poor coverage are present around the feet. Modern muon systems at the LHC are independent tracking detectors providing stand-alone track segments with up to O(50) muon hits, an example of which is shown in *Fig. 3*, of O(200) μm point resolution along with charge and momentum determination independently of the tracker.

As a consequence of their location at large radii, muon systems have to cover large areas of the order 10–100 m^2, limiting the technology choice to scintillators or gaseous detectors. Scintillator is sufficient if the identification of a muon as such is the only task of the muon system and has been used for the Tevatron detectors. To observe a track segment based on many hits and to make a precise momentum measurement, gaseous detectors have a large advantage as their cost per channel is moderate while being able to provide a track resolution better than 100 μm. Therefore, this has been the choice of all LHC experiments. As the oldest particle detection technology, a large variety of gaseous detectors have been developed, notably for applications in muon systems. The key types are discussed in the following sections.

The history of searches for new physics often required new detectors and increasing resolution to cope with the tiny cross sections. Detectors at the LHC with their muon systems being essentially stand-alone trackers are a good example of this development.

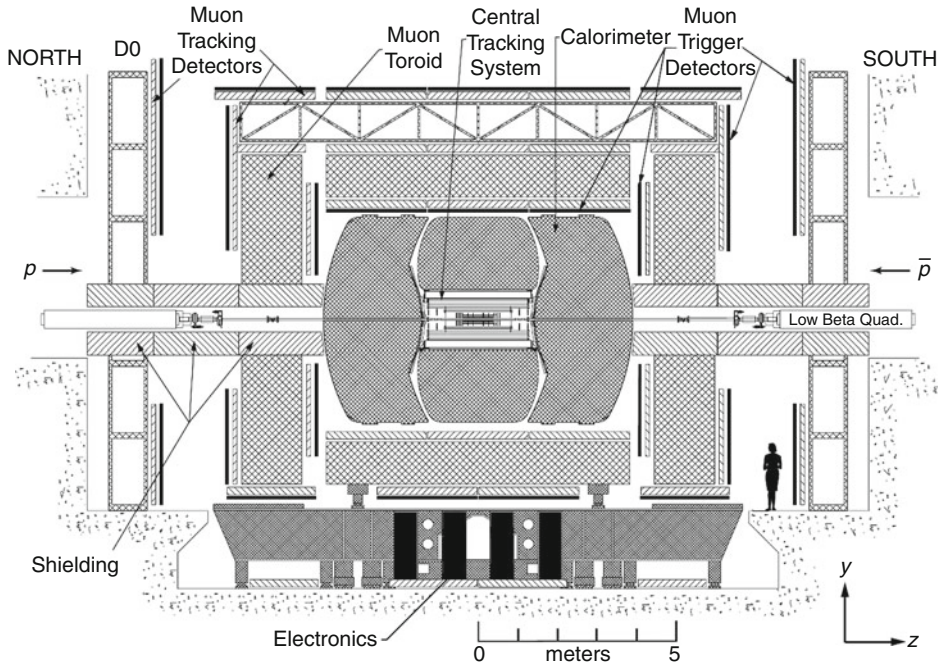


Fig. 6
The D0 detector (D0 Coll 2006), exploiting in the muon system proportional and mini-drift-tube chambers (D0 Coll 1997) and scintillators

4.1 Drift-Tube Detectors

Drift tubes with either round, hexagonal, or rectangular cross section and a central anode wire operating in avalanche or drift mode are very common. In muon spectrometers, tubes of O(cm) diameter are typically arranged in layers of O(1–5) m^2 area. Examples are depicted in *Fig. 7*. They are usually operated with a standard drift gas, often an Ar/CO_2 mixture (ATLAS, CMS), or CH_4-based mixtures (D0). Even relatively slow gas mixtures can be used given the very low occupancy in muon systems. For example, the mixture of 85% Ar with 15% CO_2 exhibits a drift velocity of ~55 μm/ns at nominal pressure and electric field of O(E = 2 kV/cm), yielding a maximal drift time for the CMS drift chambers of 380 ns (max. drift distance = 2 cm), thus integrating over 16 bunch crossings at the LHC.

All modern applications measure the drift time for electrons in the gas with respect to the time when the muon passed (time given by the accelerator clock or external trigger counters), a method pioneered by drift chambers, yielding a resolution which is about 50 times better than a pure "digital" readout where only the signal wire is identified. Such a resolution can be improved considerably when operating at overpressure, e.g., the point-resolution of the ATLAS drift tubes at 3 bar is about 80 μm (ATLAS Coll 2008) to be compared to 250 μm at nominal pressure. The ATLAS tubes are assembled and tested individually before being glued together to form either rectangular or even trapezoidal chambers (by varying the length of the individual drift tubes). Each drift-tube station measures the two-dimensional $r\phi$ projection of the muon track in the

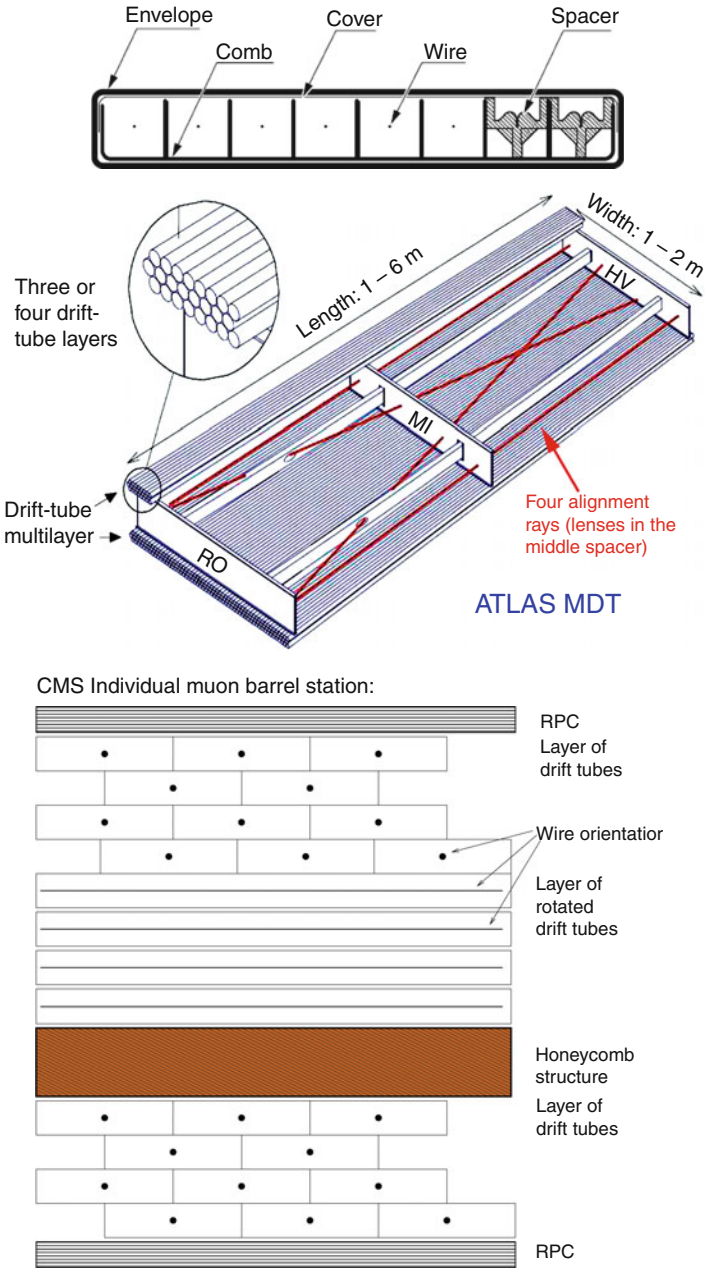


Fig. 7

Implementations of drift tubes in muon systems. *From the top*: D0 mini drift tubes (D0 Coll 1997); ATLAS monitored drift tubes (MDT) (ATLAS Coll 2008) combine two groups of three layers measuring the transverse projection (here shown opened up). CMS barrel drift-tube chambers (DT) (CMS Coll 2008) combine three superlayers of four individual layers and measure both projections

magnetic field of the air-core toroid of 0.9 T with a momentum resolution given by Eq. 9. The third dimension is determined by resistive-plate chambers, coupled to the drift chamber.

A similar tube geometry is implemented in the OPERA muon systems, where 3-cm drift tubes are used as well, although of 8-m length arranged in three large stations of parallel layers. Also OPERA measures the muon momentum through bending of the muon track with dipole magnets. Being a neutrino experiment, the occupancy is very low and multiple hits along the drift tube are practically absent. In general, multiple hits can be resolved with a resolution of O(cm) by determining the travel time of electrons along the anode wire or by using a multi-hit time-to-digital converter (TDC).

CMS uses $4 \times 1\,\mathrm{cm}^2$ drift cells with a 50-μm gold-plated steel wire, operated at ambient pressure. These cells are not pre-fabricated as individual tubes but by gluing together large aluminum plates with a set of spacers forming these rectangular cells. This so-called MIT design was first used for UA1 with cell sizes of $10 \times 3\,\mathrm{cm}^2$, almost one order of magnitude larger than the CMS cells, reflecting the increasing resolution requirement for muon systems. The CMS drift chambers are located in the iron return yoke for the inner solenoid magnet. The iron is magnetized ($B \sim 1.9$ T), but the field is largely contained in the iron and the intermediate gaps hosting the chambers are essentially field-free, except for the large-η regions. The muon track is bent when passing through the iron with its bending direction inverted in the middle of the track, see *Fig. 2*. Per cell a resolution of about 250 μm is achieved. Three of the four CMS muon barrel stations are arranged as three superlayers of four individual layers each. While two such superlayers (separated by 30 cm) measure the $r\phi$ projection, the rz projection is measured by a third superlayer which is rotated by 90° with respect to the other two (see *Fig. 7*). This yields an arrangement of hits as shown in *Fig. 3*. This way, every muon station allows to reconstruct a 3D track segment and the resistive-plate chambers provide a complementary, redundant measurement.

Drift tubes have to be constructed with a precision of O(100) μm, and especially the wire position in the center has to be precise. Wires are strung with a tension to avoid sagging. This force has to be held either by the tube wall (which should not flex) or by a dedicated support.

The drift of electrons to the anode wire is influenced by the magnetic field which exhibits a Lorentz force. The resulting deviations from the straight drift path change the drift time. In a magnetic environment, one should preferentially operate insensitive detectors such as RPCs or detectors with very short drift distances to minimize the impact of magnetic effects.

4.2 Resistive-Plate Chambers (RPC)

Resistive-plate chambers were developed as spark chambers which provide a large signal amplitude while being relatively simple in their construction. It is one of the few implementations of gaseous detectors without anode wires. A thin (2 mm) gas gap is enclosed between highly resistive plates (either Bakelite with $\rho \approx 10^9$–$10^{11}\,\Omega\,\mathrm{cm}$ or glass with $\rho \approx 10^{13}\,\Omega\,\mathrm{cm}$) covered with a conductive graphite coating on the outside. The movement of the charge in the gas-filled gap induces a signal outside the plates, which is subsequently picked up by external readout strips (or pads), usually crossing each other and thus providing a 2-dimensional readout. Together with the small thickness, a 3-dimensional space point is provided by an RPC station. *Figure 8* shows a double-gap RPC with a common readout strip enclosed between both chambers.

Due to the very short distance, the response time is short, of the order 4 ns, thus making RPCs a good choice for triggering chambers at high interaction rates, as is done for all four

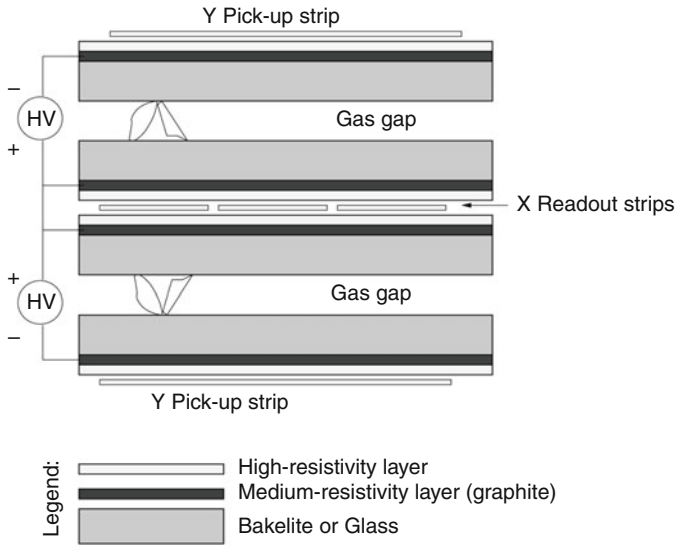


Fig. 8
Principle of a double-gap RPC

LHC experiments. The chamber design allows one to produce thin, large-area detector panels to cover hundreds of m^2 area and may even fit in thin gaps, as for example in the dipole magnets of the OPERA experiment (OPERA Coll 2010) or the instrumented flux return of BaBar (BaBar Coll 2002), as shown in *Fig. 9*, and Belle (BELLE Coll 2002). The achievable spatial resolution (O(cm–mm)) depends on the resistivity of the plates, which prevents the charge from spreading out, and the granularity of the readout strips. Depending on the operating voltage (usually of the order of 10 kV), the amplification varies between $O(10^5)$ in avalanche or proportional mode, for example, in ATLAS, CMS, STAR, and ALICE TOF, and $O(10^7)$ in streamer mode, for example, in BaBar, Belle, and ALICE. The high electric field strengths related to sparks may potentially damage the inner Bakelite coating, followed by a small but constant reduction of the efficiency, as seen by BaBar (Anulli et al. 2002, 2003, 2005). It appears that, besides the amplification, other quality parameters of the gaps also play a role, such as the surface coating and variations of the gap thickness. Given that the BaBar interaction rate is much lower than that at LHC, where all four experiments exploit RPCs, an intensive R&D program was initiated. As a source for such damage, roughness of the inner surface has been identified. Little bumps on the surface receive higher charges due to the smaller gap which may yield damage with time. To some extent, the surface roughness can be minimized by applying a thin film of oil. Another aspect is the amount of charge generated for the signal. By lowering the voltage, the efficiency drops slightly but also the damage potential is reduced. Therefore, the RPC chambers of the CMS muon system are operated at only 9.5 kV. The lower efficiency is compensated by combining two chambers as shown in *Fig. 8*.

ATLAS and CMS instrumented 3,650 and 8,000 m^2, respectively, in the muon barrel and the forward systems. Both RPC systems are operating in avalanche mode. While the CMS chambers

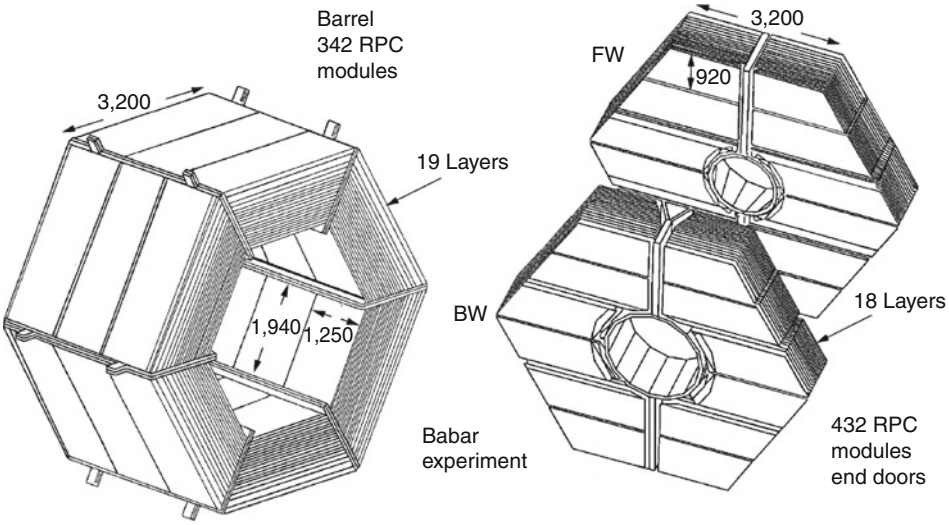


Fig. 9
RPCs in the BaBar detector (Anulli et al. 2003)

provide redundant information to the high-resolution muon drift and cathode strip chambers, the ATLAS RPCs provide the necessary third projection complementing the 2D information of the precision drift tubes.

4.3 Multi-Wire Chambers

Under the label of multi-wire proportional chambers, we want to summarize all implementations of gaseous detectors where several wires arranged as a plane share a gas volume. The important aspect here is the plane, to be distinct from drift chambers, where many (10^2–10^3) wires are arranged concentrically inside a gas-filled cylinder of O(m) diameter. Such drift chambers are used as central tracking chambers, but not for the outside muon systems. Also, the planar, multi-wire chambers have been used as inner trackers (e.g., HERA-B) but also find implementation in muon systems, for example, in LHCb (see *Fig. 4*) and the CMS forward muon system (see *Fig. 11*).

The conceptual design of multi-wire chambers is shown in *Fig. 10*. The anode wires are arranged in planar layers spaced by d = 1–2 mm. The minimal distance is limited by electrostatic forces, just as in drift chambers. If wires are read out individually, their separation distance determines the spatial resolution, but cost constraints may force the user to match several wires onto one readout channel. The layer of wires is enclosed between two cathodes and the gap is filled with an appropriate gas mixture for generating primary electrons and amplifying the signal near the anode wires, just like with drift tubes.

The anode signal is fast, determined by the drift velocity, but the positively charged ions drift about a factor of 1,000 times slower to the cathode. While the previously described detectors do not use the cathode signal, multi-wire chambers with a segmented cathode do, and this principle is sketched in *Fig. 10*. The charge spreads over several strips, and charge interpolation

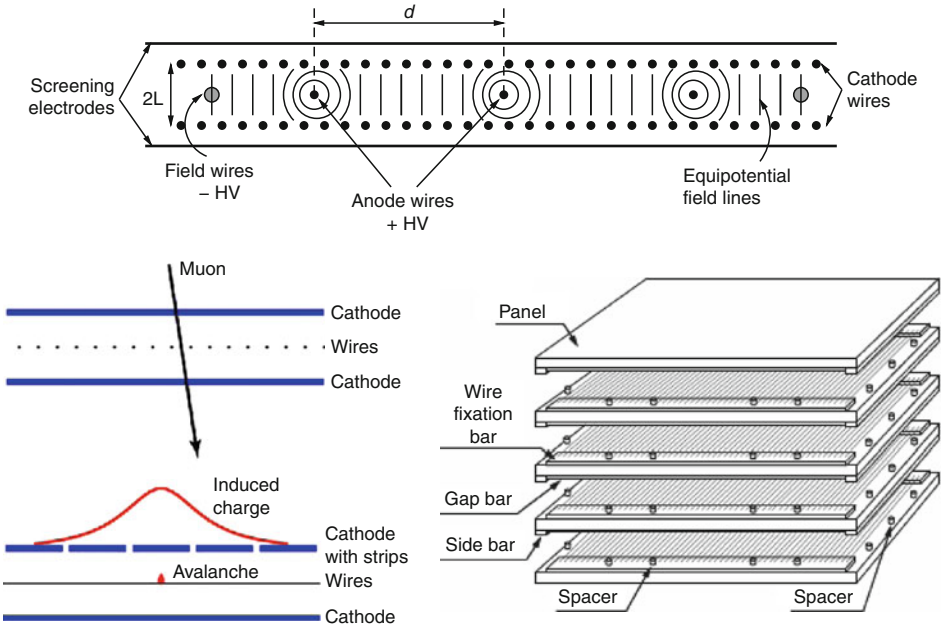


Fig. 10

In a multi-wire chamber, wires are arranged in planes and enclosed between cathodes with identical spacing. If the cathodes are segmented and read out in addition to the wires, the design is referred to as cathode strip chamber (CMS Coll 2008). The example shows a multi-wire proportional chamber from the LHCb muon system (LHCb Coll 2008)

provides a resolution better than the strip width divided by $\sqrt{12}$. This way a high-resolution 2D space point can be extracted while keeping the number of readout channels at a reasonable level. Combining the information of the cathode strips and anode wires provides a 3D space point. Such cathode strip chambers are the precision detector component of the CMS muon endcaps (*Fig. 11*) and instrument the regions of high pseudorapidity of the ATLAS muon endcaps (see region indicated in *Fig. 5*). Their main advantages in these forward regions are their capability to handle larger occupancies than drift tubes and to be rather insensitive to the magnetic field due to their short drift distances. Most of the ATLAS endcap region is instrumented with thin-gap chambers (TGC), multi-wire proportional chambers with a signal based on anode wires.

5 Muon Spectrometers for Cosmic Ray Measurements

Several multipurpose detectors, built for accelerator-based particle physics, have also been used to measure cosmic muons, see Chap. 24, "Indirect Detection of Cosmic Rays". Their large effective surfaces and/or their excellent spectrometers based on a huge magnetic volume allow for precise measurements of the flux as a function of momentum, charge, and direction. Recent examples are the extended LEP detectors Cosmo-Aleph (Grupen et al. 2008), Delphi

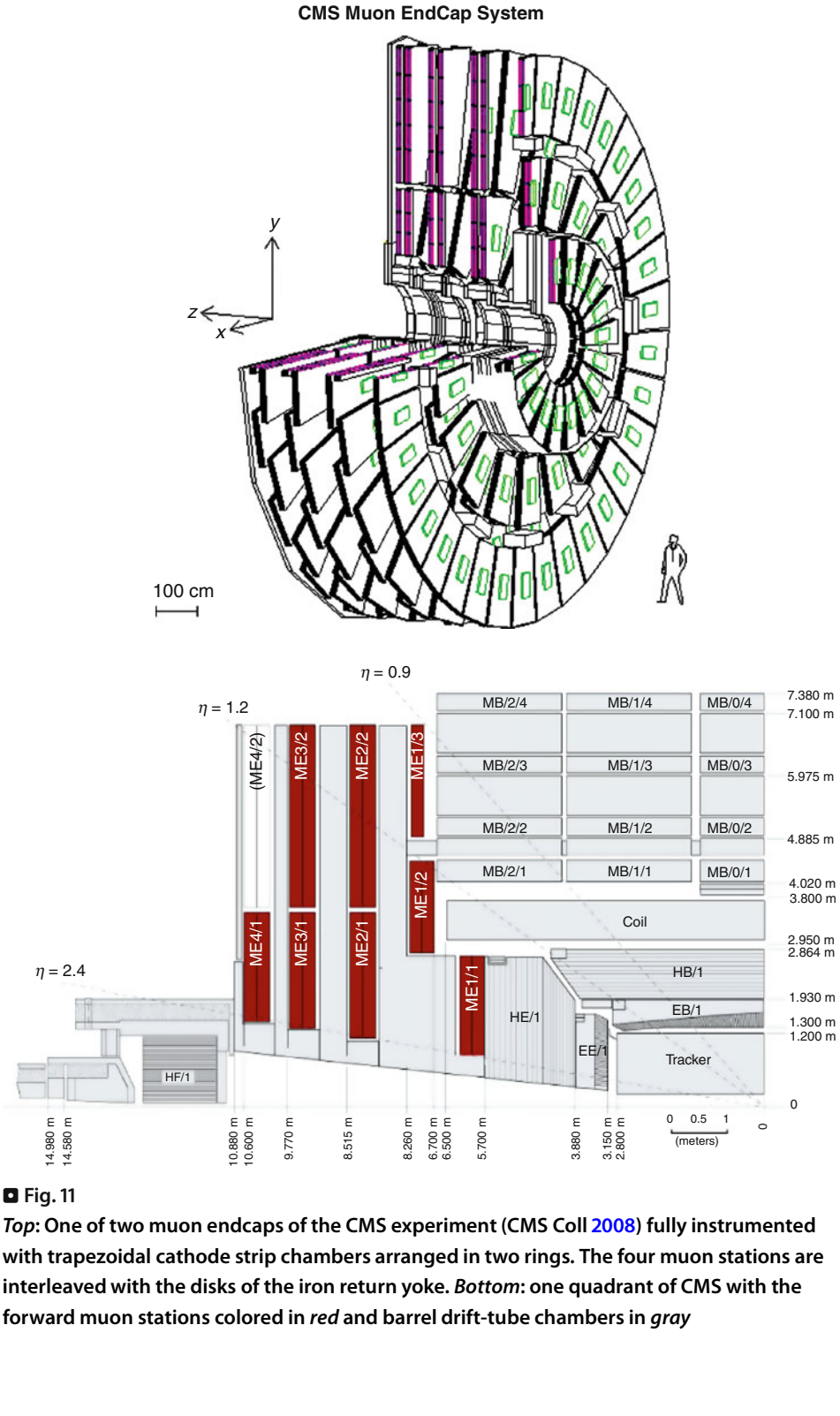


Fig. 11

Top*: One of two muon endcaps of the CMS experiment (CMS Coll 2008) fully instrumented with trapezoidal cathode strip chambers arranged in two rings. The four muon stations are interleaved with the disks of the iron return yoke. *Bottom*: one quadrant of CMS with the forward muon stations colored in *red* and barrel drift-tube chambers in *gray

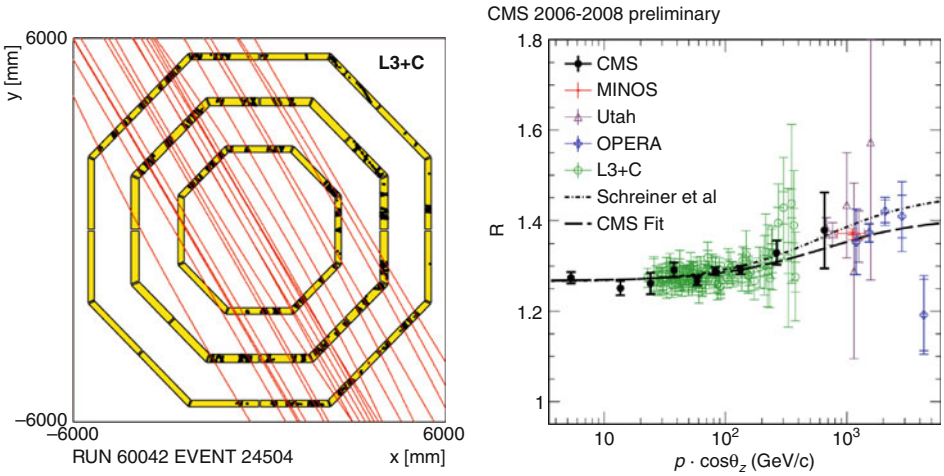


Fig. 12

***Left*: Muon-bundle event recorded by the L3+C detector (L3 Coll 2002). *Right*: Muon charge-ratio measurements as a function of vertical momentum component (CMS Coll 2010)**

(Travnicek and Ridky 2003), and L3+C (L3 Coll 2002), the underground neutrino detectors MINOS (MINOS Coll 2009) and OPERA (OPERA Coll 2010), and the CMS experiment (CMS Coll 2010) at the LHC collider. ❯ *Figure 12* shows a multi-muon event recorded in the L3+C detector. The deep underground detectors (at depths of a few km of water equivalent) combine the measurement of momentum and charge inside the detector with the angle-dependent energy loss in the overburden, thus allowing a charge and momentum measurement even for multi-TeV muons. ❯ *Figure 12* shows the recently measured ratio of the fluxes of positive to negative cosmic muons, as a function of the vertical momentum component, together with model fits (CMS Coll 2010).

Dedicated cosmic spectrometers are operating at the top of the atmosphere (balloons) or above (satellites) to measure the primary cosmic particles, nuclei, and electrons. Atmospheric muons produced as secondary particles in air showers can be detected at earth's surface or underground. Detectors are built to measure either individual muons or to reconstruct the whole air shower – by sampling the secondary particles on the ground, muons, and also electrons and hadrons, depending on shower energy, altitude, and zenith angle. In addition, cosmic neutrino experiments like ANTARES (ANTARES Coll 2010) or IceCube (IceCube Coll 2006) detect the muons created in charged-current neutrino interaction in liquid or frozen water.

In the following, we first present three examples for cosmic muon spectrometers and then we discuss briefly the measurement of the muon component of air showers.

5.1 Atmospheric Muon Detectors

The *Okayama* muon telescope (Yamashita et al. 1996) is shown in ❯ *Fig. 13*. It was set up in the year 1992 at the Okayama University in Japan at sea level. Its alt-azimuthal mount can – like an optical telescope – move the apparatus in all directions, in particular zenith angles from

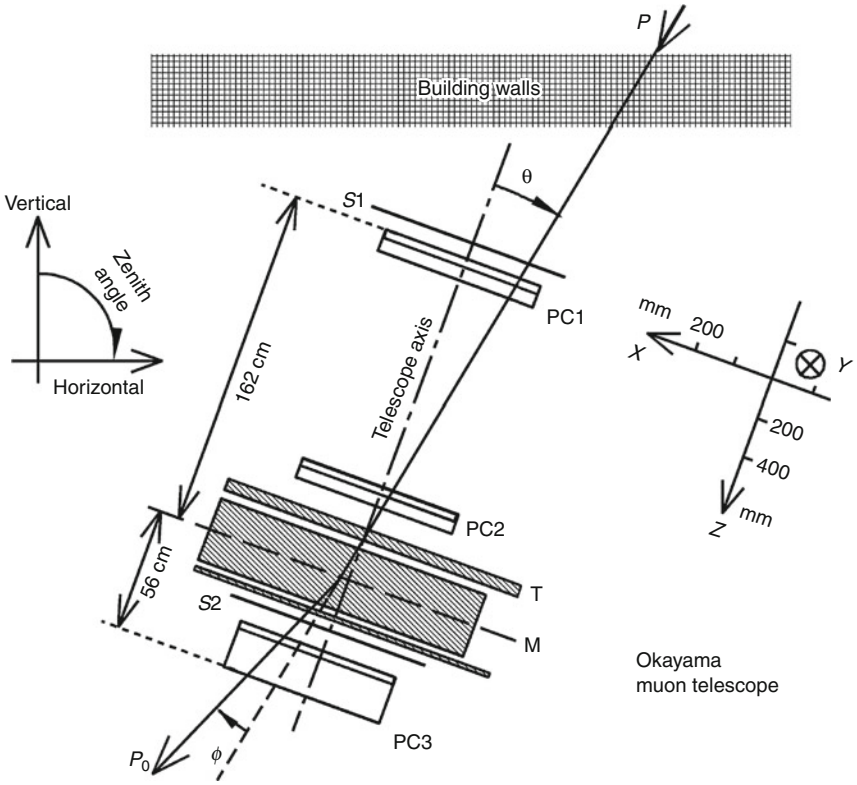


Fig. 13
The Okayama muon telescope (Tsuji et al. 2001)

0° to 80° can be accessed. The magnet consists of an iron cube with dimension 32 cm and a conventional coil. The magnetic field has a strength of 1.8 T. The geometrical acceptance is rather small, 75 cm^2 sr. Drift chambers (wall-less multi-wire proportional chambers, PC1, PC2, PC3 in the figure) measure the trajectory (both coordinates) with a resolution of 0.28 mm per point. Scintillators are used for triggering. The maximum detectable momentum p_{mdm} is 270 GeV/c. The Okayama telescope has been used to measure the flux of cosmic muons in the momentum range 1.5–250 GeV/c for a variety of zenith angles, and also the charge ratio (Tsuji et al. 1998). In addition, the azimuthal angular dependence was investigated, clearly showing the east–west effect for low-energy muons, caused by the geomagnetic field. Searches for point sources were unsuccessful.

The *MACRO* detector (Monopole, Astrophysics, Cosmic Ray Observatory) was installed in the Gran Sasso underground laboratory, at an average depth of 3.7 km (water equivalent) (MACRO Coll 2002). Data were taken from 1989 till end of 2000. The multipurpose detector consisted of six supermodules of dimensions 12.6 m (length) × 12 m (width) × 9.3 m (height). *Figure 14* shows a cross section. A muon traversing the detector from the top passes through three scintillator planes (top, middle, bottom), streamer tubes (at the top and in lower hemisphere), and track-etch detectors. This detector design was optimized for searches for magnetic

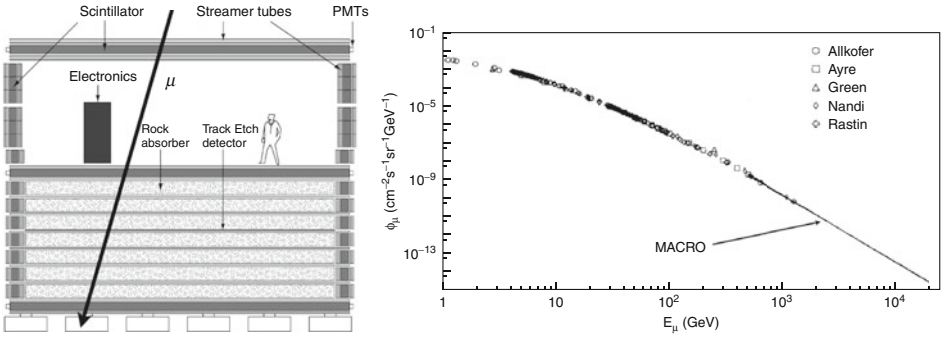


Fig. 14

***Left*: The MACRO detector, cross section (MACRO Coll 2002). *Right*: Near vertical momentum spectrum of atmospheric muons (MACRO Coll 1995)**

monopoles, slow and strongly ionizing particles. MACRO analyzed – among other topics – the flux of downgoing atmospheric muons, and measured also upgoing muons from neutrino interactions in the earth.

The direction of muons passing through the detector could be measured with a resolution of about 0.2°. The MACRO detector included neither an active magnet nor magnetized iron components. Nevertheless, the local momentum (but not the charge) of muons could be determined. For this purpose, transition radiation detectors (TRD), foam radiators plus proportional counters, were installed in the empty upper hemisphere of the MACRO detector. The TRDs measure $\gamma = E_\mu/(m_\mu c^2)$ so that MACRO was able to determine the local muon energy from 0.1 to 1 TeV. As an alternative method, the determination of the amount of multiple scattering in the rock absorber layers inside the detector had been proposed – it allows for momentum measurements up to 40 GeV. The local muon measurement can be combined with the calculable muon energy loss in the overburden to determine the momentum spectrum at the surface.

Even without any momentum measurement inside the MACRO detector the near vertical muon energy spectrum at the earth's surface could be determined, exploiting the complex topography of the surface, resulting in a variation of the depth from 3 to 7 km (water equivalent), depending on the muon direction: The underground muon intensity was measured for different angles, corresponding to different rock thicknesses, and with the help of a simple model for the energy dependence of the flux (power law $\Phi \sim E^{-\gamma}$), the surface muon spectrum could be fitted, in the range 0.5–20 TeV, see *Fig. 14*.

Normally, the *BESS* detector (Balloon-borne Experiment with a Superconducting Spectrometer) is airborne – several balloon flights were undertaken in the Antarctica from 1993 to 2008. In addition, measurements of atmospheric muons were performed at various heights, and also at ground level. The "BESS-TeV" spectrometer (Haino et al. 2004) is depicted in *Fig. 15*. The following components are relevant for precise muon momentum and charge measurements:

- The uniform magnetic field is provided by the superconducting solenoid of 1 m diameter and 1 m length, with a field strength of 1 T.

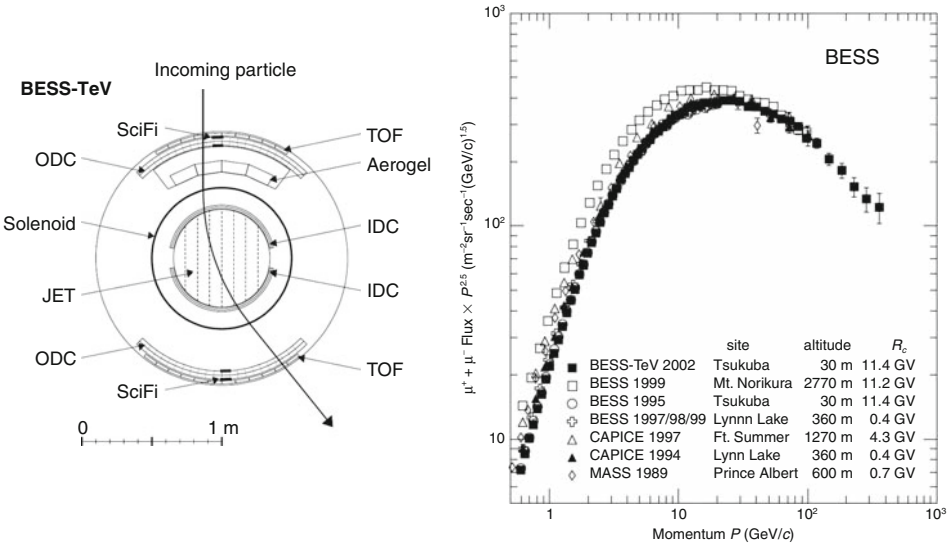


Fig. 15
***Left*: The BESS-TeV spectrometer, cross section (Haino et al. 2004). *Right*: Near vertical momentum spectrum of atmospheric muons (Haino et al. 2004)**

Three kinds of tracking devices measure the muon trajectory, over a distance of up to 1.7 m:

- The jet-type central drift chamber (JET)
- Inner drift chambers (IDC), inside the coil
- The outer drift chamber (ODC)

In total about 60 measurements are made, the typical spatial accuracy is 150 µm per point. The small scintillating fiber system (SciFi) was used for calibrating the ODC. The maximum detectable momentum p_{mdm} is 1.4 TeV/c. The measured atmospheric muon spectra (Haino et al. 2004) are shown in *Fig. 15*, for zenith angles smaller than 26°. They are complementary to the MACRO results in the TeV regime, see *Fig. 14*.

5.2 Air Shower Detector Arrays

The first air showers were measured by P. Auger et al. in the Swiss Alps using Geiger tubes at distances of up to 300 m (Auger et al. 1939). Some of the larger air shower detectors built later were equipped also with muon track detectors, covering a small fraction of the total array. An example is the KASCADE detector (KASCADE Coll 2003) with its underground setup of limited streamer tubes to measure the direction of muons, but not their momenta. The biggest air shower detector today, the 3,000 km^2 large Auger observatory (Pierre Auger Coll 2004), uses water tanks in which charged particles generate Cherenkov light, which is detected by photomultipliers. For inclined showers (zenith angle >60°), the muons are the dominant component at the earth's surface.

6 Muon Radiography

Muon detectors can also be used to localize absorbing material through the resulting reduction in the flux of cosmic muons or via multiple scattering of muons – we speak of "muon radiography." The possible applications reach from archaeology (Alvarez et al. 1970) and geology (Macedonio and Martini 2009) to the detection of smuggled nuclear weapons (Szeptycka and Szymanski 2009), see also ❯ Chap. 26, "Accelerator Mass Spectrometry and its Applications in Archaeology, Geology, and Environmental Research," ❯ Chap. 28, "Particle Detectors Used in Isotope Ratio Mass Spectrometry, with Applications in Geology, Environmental Science and Nuclear Forensics," and ❯ Chap. 25, "Technology for Border Security".

As a first example, we present the search for cavities in pyramids. This idea, illustrated in ❯ *Fig. 16*, was put forward by L. Alvarez who searched in the 1960s – unsuccessfully – for hidden chambers in the Chephren Pyramid in Egypt (Alvarez et al. 1970). He used a stack of 2 m^2 large spark chambers placed in a void beneath the pyramid and measured the counting rate as a function of direction. The typical counting rate in a $3^\circ \times 3^\circ$ bin was of the order of 1/h. No significant local maximum, as expected for reduced absorption, was found. Recently, the Pyramid of the Sun in Mexico has been selected for similar measurements using multi-wire proportional chambers (Alfaro et al. 2008).

Another application of muon detectors is the use of the 7 m^2 large drift-tube (DT) chambers designed for the CMS experiment to inspect transport containers from the outside (Benettoni et al. 2007; Pesente et al. 2009). ❯ *Figure 17* shows the principle: The size of the Coulomb scattering angle θ of a downward going cosmic muon is a measure of the amount of material. As can be seen from ❯ Eq. 2, the scattering angle depends on the size l of the object, its mass density ρ, and atomic number Z:

$$\theta \propto \frac{\sqrt{l\,\rho\,Z}}{p}. \tag{10}$$

A typical projected scattering angle for iron and l = 10 cm, p = 1 GeV is 30 mrad, to be compared with the angular resolution of about 1 mrad. The momentum is not measured, but since the

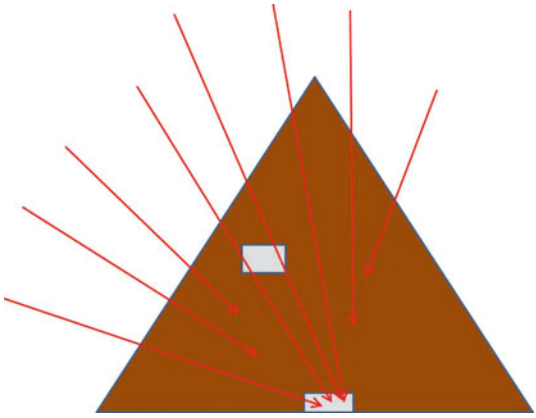


Fig. 16
Principle of muon radiography of pyramids



Fig. 17
***Left*: Principle of muon radiography with CMS-type muon chambers (Benettoni et al. 2007; Pesente et al. 2009). *Right*: Radiography result with CMS-type chambers (Benettoni et al. 2007; Pesente et al. 2009). A lead and an iron block are detected**

cosmic momentum spectrum is known, a statistical analysis can reveal the enclosed material. *Figure 17* shows the result obtained for a test setup with a lead and a bigger iron block placed in between the chambers, obtained in 1 h of data taking. The color scale is a measure of the scattering angle.

Also the moon has been seen via particle radiography – this is again an astroparticle application of muon detectors. The absorption of primary cosmic particles causes a reduction of the atmospheric muon flux in the direction of the moon – after correcting for geomagnetic effects. In the last couple of years, for example, L3+C (L3 Coll 2005), ARGO-YBJ (ARGO-YBJ Coll 2008), IceCube (Boersma et al. 2009), ANTARES (Distefano 2009), and MACRO (MACRO Coll 2002) have observed the moon shadow. These measurements can be translated into limits on the cosmic antiproton flux (reversed bending in the earth's magnetic field), and they allow to determine the detector's angular resolution, since the moon's diameter is only 0.5°.

7 Conclusions

Muon spectrometers have played a central role in particle and cosmic ray physics since the discovery of atmospheric muons in form of charged-particle tracks bending in the magnetic field of a cloud chamber in the year 1936. Today it is a "must" for all big particle detectors at lepton or hadron colliders to include sophisticated muon detectors embedded in strong magnetic fields. Muons are on the one hand easy to detect, on the other hand they provide clear signatures for many processes involving new physics. Due to their long range in matter, muons have also become interesting tools for various radiography applications, measuring the thickness of materials not directly accessible. By now we have seen an enormous variety of muon spectrometer types, combining different detection methods with various magnetic field configurations.

References

Alfaro R et al (2008) Searching for possible hidden chambers in the pyramid of the sun. Proceedings of the 30th International Cosmic Ray Conference, Mexico City

ALICE Coll (2008) The ALICE experiement at the CERN LHC. J Instrum 3:S08001

Alvarez L et al (1970) Search for hidden chambers in the pyramids. Science 167:832

ANTARES Coll (2010) Measurement of the atmospheric muon flux with a 4 GeV threshold in the ANTARES neutrino telescope. Astropart Phys 33:86

Anulli F et al (2002) The BaBar instrumented flux return performance: lessons learned. Nucl Instrum Methods A 494:455

Anulli F et al (2003) Mechanisms affecting performance of the BaBar resistive plate chambers and searches for remediation. Nucl Instrum Methods A508:128

Anulli F et al (2005) Performance of second generation BaBar flux resistive plate chamber. Nucl Instrum Methods A 552:276

ARGO-YBJ Coll, Wang Y et al (2008) Preliminary results of the Moon shadow using ARGO-YBJ detector. Nucl Phys Proc Suppl 175:551

ATLAS Coll (2008) The ATLAS Experiment at the CERN LHC. J Instrum 3:S08003

Auger P et al (1939) Extensive cosmic-ray showers. Rev Mod Phys 11:288

BaBar Coll (2002) The BaBar detector. Nucl Instrum Methods A479:1

BELLE Coll (2002) The Belle detector. Nucl Instrum Methods A479:117–232

Benettoni M et al (2007) Muon radiography with the CMS muon barrel chambers. Proceedings of the 2007 IEEE nuclear science symposium, Honolulu, Hawaii

Blum W, Riegler W, Rolandi L (2005) Particle detection with drift chambers. Springer, Berlin, Germany

Boersma DJ et al for the Icecube Collaboration (2009) Moon shadow observation by ice-cube. Proceedings of the 31st international cosmic ray conference, Lodz, Poland

CDF Coll (1996) The CDF II Technical Design Report. Fermilab-Pub-96-390-E, November 1996

CMS Coll (2008) The CMS experiment at the CERN LHC. J Instrum 3:S08004

CMS Coll (2010) Measurement of the charge asymmetry of atmospheric muons with the CMS detector. Physics Analysis Summary CMS-PAS-MUO-10-001

D0 Coll (2006) The upgraded D0 Detector. Nucl Instrum Methods A565:463–537

D0 Coll, Baldin B et al (1997) Technical design of the central muon system. D0 Note 3365, 29 March 1997

Distefano C for the Antqres Coll (2009) Detection of the moon shadow with the ANTARES neutrino telescope. International workshop on very large volume neutrinotelescopes, Athens, Greece

Gluckstern RL (1963) Uncertainties in track momentum and direction due to multiple scattering and measurement errors. Nucl Instrum Methods 24:381

Grupen C, Shwartz B (2008) Particle detectors. Cambridge University Press, Cambridge, New York

Grupen C et al (2008) Cosmic ray results from the CosmoALEPH experiment. Nucl Phys B 175–176:286

Haino S et al (2004) Measurements of primary and atmospheric cosmic-ray spectra with the BESS-TeV spectrometer. Phys Lett B 594:35

IceCube Coll (2006) First year performance of the IceCube neutrino telescope. Astropart Phys 26:155

KASCADE Coll (2003) The cosmic-ray experiment KASCADE. Nucl Instrum Methods A 513:490

L3 Coll (2002) The L3+C detector, a unique tool-set to study cosmic rays. Nucl Instrum Methods A 488:209

L3 Coll, Achard P et al (2005) Measurement of the shadowing of high-energy cosmic rays by the Moon: A search for TeV-energy antiprotons. Astropart Phys 23:411

LHCb Coll (2008) The LHCb detector at the LHC. J Instrum 3:S08005

Macedonio G, Martini M (2009) Motivations for muon radiography of active volcanoes. Earth Planets Space 61:1 and references therein

MACRO Coll (1995) Vertical muon intensity measured with MACRO at the Gran Sasso Laboratory. Phys Rev D 52:3793

MACRO Coll (2002) The MACRO detector at Gran Sasso. Nucl Instrum Methods A 486:663

MINOS Coll (2009) Measurement of the atmospheric muon charge ratio at TeV energies with the MINOS detector. Phys Rev D 76:052003

OPERA Coll (2010) Measurement of the atmospheric muon charge ratio with the OPERA detector. arXiv:1003.1907v1, 9 March 2010

Particle Data Group, Amsler C et al (2008) Review of particle physics. Phys Lett B 667:1, and references therein

Pesente S et al (2009) First results on material identification and imaging with a large-volume muon tomography prototype. Nucl Instrum Methods A 604:738

Pierre Auger Coll (2004) Properties and performance of the prototype instrument for the Pierre Auger Observatory. Nucl Instrum Methods Phys Res A 523:50

Spieler H (2005) Semiconductor systems. Oxford University Press, New York

Szeptycka M, Szymanski P (2009) Remarks on myon radiography. In: Begun V, Jenkovszky LL, Polanski A (eds) Progress in high energy physics and nuclear safety. Springer, Dordrecht, pp 353–362

Travnicek P, Ridky J (2003) Cosmic multi-muon bundles measured at DELPHI. Nucl Phys B 122:285

Tsuji S et al (1998) Measurements of muon at sea level. J Phys G Nucl Part Phys 24:1805

Tsuji S et al (2001) Atmospheric muon measurements II: zenith angular dependence. Proceedings of the 27th International Cosmic Ray Conference, Hamburg, Germany

Yamashita Y et al (1996) An altazimuthal counter telescope with a magnet spectrometer tracing Cygnus X-3. Nucl Instrum Methods A 374:245

20 Calorimeters

Richard Wigmans
Texas Tech University, Lubbock, TX, USA

1 Introduction ***498***

2 Functions and Properties of Calorimeters ***499***
2.1 Calorimeters in Modern Particle Physics Experiments 499
2.2 Important Calorimeter Properties 500

3 Calorimeter Types ***503***
3.1 Electromagnetic Calorimeters 503
3.2 Hadron Calorimeters 507
3.3 Cryogenic Calorimeters 512
3.4 Natural Calorimeters 513

4 Concluding Remarks ***515***

References ***516***

Further Reading ***517***

C. Grupen, I. Buvat (eds.), *Handbook of Particle Detection and Imaging*,
DOI 10.1007/978-3-642-13271-1_20, © Springer-Verlag Berlin Heidelberg 2012

Abstract: Calorimeters were originally developed as instruments that measure energy deposits through changes in temperature. In nuclear and particle physics, this class of instruments is used to measure the properties of particles carrying energies ranging from a small fraction of 1 eV to 10^{20} eV or more. And these properties are not limited to the energy carried by these particles, but concern the entire four-vector, including the particle mass and type. In many modern experiments, large calorimeter systems play a central role. In this chapter, we review this role, and the important calorimeter properties deriving from it. We also give an overview of the most common types of calorimeters, classified according to the type of particles and the energy range for which they are intended.

1 Introduction

The term "calorimetry" (literally: *Heat measurement*) has its origin in thermodynamics. One *calorie* is defined as the amount of energy needed to increase the temperature of 1 g of water by 1°C. Many readers probably remember the demonstrations in high school in which the specific heat of some material was measured, or (on sunny days) the solar constant. Calorimeters were the thermally isolated boxes containing the substance under study. A thermometer provided the experimental information.

Modern, highly sophisticated versions of these instruments are in use in nuclear weapons laboratories, where they are used for the assay of fissionable material. For example, ^{239}Pu produces heat at a rate of 2 mW g^{-1}. Calorimetry can provide an accurate measurement of the amount of plutonium in a sample, in a noninvasive manner.

In nuclear and particle physics, calorimetry refers to the detection of particles, and measurement of their properties, through total absorption in an instrument called a *calorimeter*. Calorimeters exist in a wide variety, but they all have the common feature that the measurement process through which the particle properties are determined is *destructive*. Unlike, for example, wire chambers that measure a particle's properties by tracking it in a magnetic field, the particles are no longer available for inspection by other devices once the calorimeter is done with them. The only exception to this rule concerns muons. The fact that these particles may penetrate the substantial amount of matter represented by a calorimeter without losing much of their energy is actually an important ingredient for their identification as muons. Other particles (neutrinos and particles hypothesized in the context of Supersymmetry) do not leave any trace in a calorimeter, or in any other detector component. Yet, calorimeters are also crucial tools for recognizing these particles, and measuring their properties.

In the absorption process, almost all the particle's energy is eventually converted into heat, hence the term calorimetry. However, the units of the energy involved in this process are typically very different from the thermodynamic ones. The most energetic particles in modern accelerator experiments are measured in units of TeV (1 TeV = 10^{12} eV = 1,000 GeV), whereas one calorie (4.18 J) is equivalent to about 10^7 TeV. The rise in temperature of the particle detector is thus, for all practical purposes, negligible, and therefore other ways to measure the deposited energy are employed. These methods are typically based on the measurable effects of atomic or molecular excitation (ionization charge, scintillation light), or on collective effects such as the production of Čerenkov light or sound in the absorbing medium. There is, however, one class of particle detectors in which thermal effects are being exploited, e.g., the transition from the superconducting to the normally conducting material phase. Such detectors are used in the search for very specific phenomena in which miniscule amounts of energy are transferred.

In the following, we first describe the functions that calorimeters typically fulfill in modern experiments, and the relevant calorimeter properties deriving from that. Then, we give an overview of the most common types of calorimeters used in such experiments.

2 Functions and Properties of Calorimeters

2.1 Calorimeters in Modern Particle Physics Experiments

Calorimeters measure the energy released in the absorption of (sub)nuclear particles entering them. They generate signals that make it possible to quantify that energy. However, typically these signals provide also other information about these particles, and about the event in which they were produced. The signals from a properly instrumented absorber may be used to measure the entire four-vector of the particles. By analyzing the energy deposit pattern, the direction of the particle can be measured. The mass of the showering particle can be determined in a variety of ways, of which we mention:

- The E/p *method*, in which the energy measured in the calorimeter is compared with the momentum measured with a tracker in a magnetic field. This method only works for charged particles and relatively low energies.
- By analyzing the *energy deposit profile*. This method is frequently used to identify electrons. Especially in calorimeters with high-Z absorber material, em showers are much more shallow and concentrated around the shower axis than hadronic showers. This feature is also exploited in *preshower detectors* (Fig. 1a).
- By measuring the *time structure* of the calorimeter signals. An example of electron identification on the basis of this method is shown in Fig. 1b.

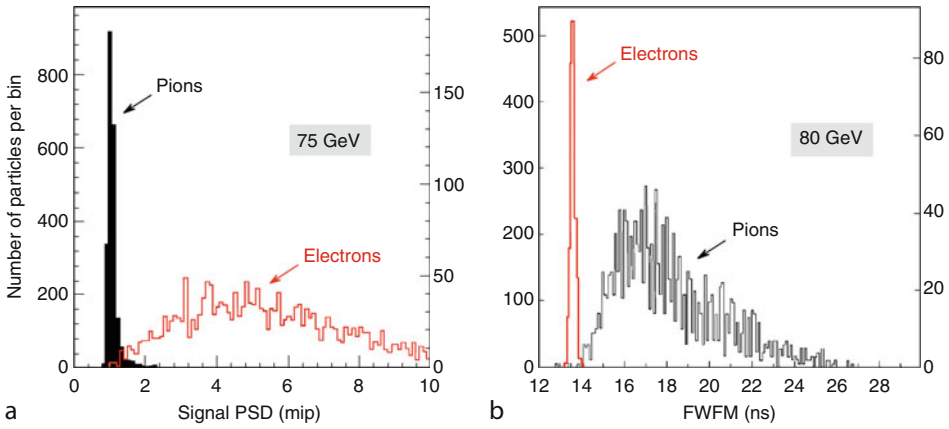


Fig. 1

Two different methods to distinguish electrons from pions in a calorimeter. Diagram (a) shows the signal distributions in a $2X_0$ thick preshower detector (PSD). In diagram (b), the distributions of the duration of the pulse (here defined as the full width at one-fifth of the maximum amplitude) is plotted. In both diagrams, the left-hand scale refers to the narrow distribution, the right-hand scale to the broad one

Apart from these methods, which are geared toward identifying electrons, calorimeters are also used to identify muons and neutrinos. High-energy muons usually deposit only a small fraction of their energy in the calorimeter and produce signals in downstream detectors. Neutrinos typically do not interact at all in the calorimeter. If an energetic neutrino is produced in a colliding-beam experiment, this phenomenon will lead to an imbalance between the energies deposited in any two hemispheres into which a 4π detector can be split. Such an imbalance, usually referred to as *missing transverse energy*, led to the discovery of the W boson (Arnison et al. 1983).

The latter is an example of the *energy-flow* information a calorimeter system can provide. Other examples of such information concern the *total transverse energy* and the production of *hadronic jets* in the measured events. Since this information is often directly related to the physics goals of the experiment, and since it can be obtained extremely fast, calorimeters usually play a crucial role in the trigger scheme, through which interesting events are selected and retained for further inspection off-line.

2.2 Important Calorimeter Properties

The calorimeter's properties should be commensurate with the role it has to play in the experiment. Relevant properties in this context are:

- The *energy resolution*. Just as for any other detector, the precision of the measured information determines the quality of the scientific studies that can be performed with it. The importance of the energy resolution is illustrated in ❯ *Fig. 2*, for two very different instruments used in energy ranges differing by seven orders of magnitude.

  The energy resolution is determined by *fluctuations* in the absorption process. If these fluctuations are stochastic, i.e., obey Poisson statistics, then the relative energy resolution should *improve* with increasing energy as

$$\frac{\sigma}{E} = \frac{a}{\sqrt{E}}. \tag{1}$$

  This is an attractive feature, which distinguishes calorimeters from almost all other particle detectors. However, often non-stochastic fluctuations play a role as well, and these tend to dominate the calorimeter performance in the high-energy range.
- The *size*. An example of the non-stochastic effects mentioned above is the fluctuations in leakage that occur when the calorimeter is not large enough to contain the showers. The effects of shower leakage on the energy resolution depend on the average containment level, which itself varies with energy. Therefore, there is no simple expression such as ❯ Eq. 1 to describe these effects for a given detector. The effects also depend very strongly on the *type of leakage*. ❯ *Figure 3* shows that the resolution is much more sensitive to the effects of longitudinally escaping shower particles than to lateral non-containment. Even more consequential are the effects of *albedo*, i.e., shower leakage through the detector's front face, but the energy involved in this process is usually negligible in calorimeters intended for the GeV domain and up. A calorimeter should thus first and foremost be sufficiently deep to contain the highest-energy showers at the desired level. As a rule of thumb, the contribution of longitudinal leakage fluctuations to the energy resolution is about the same as the average

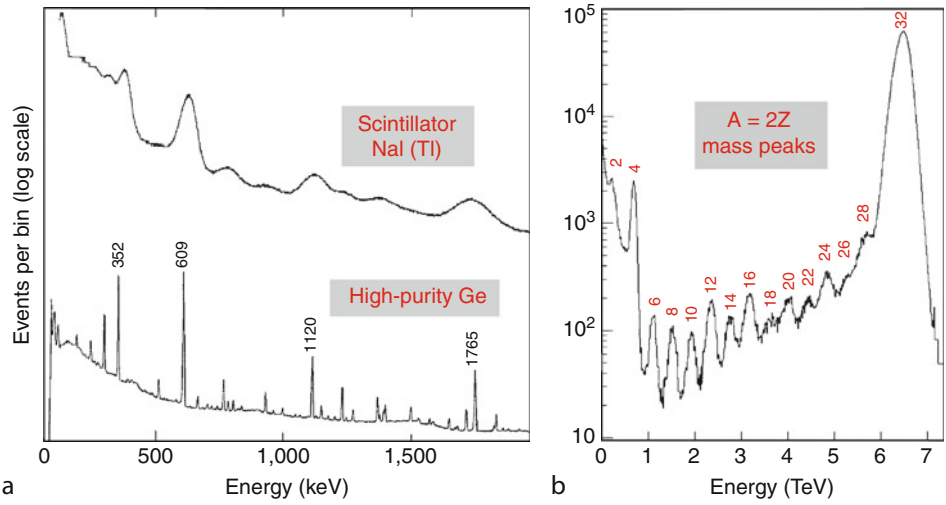


Fig. 2

Signal distributions measured with calorimeters. (a) shows the spectrum of nuclear γ rays from the decay of natural uranium ore, measured either with a NaI(Tl) scintillator, or with a high-purity germanium crystal (Wigmans 2000). (b) shows a distribution of the total energy carried by heavy ions accelerated to 200 GeV/nucleon dumped in a uranium/plastic-scintillator calorimeter (Young et al. 1989)

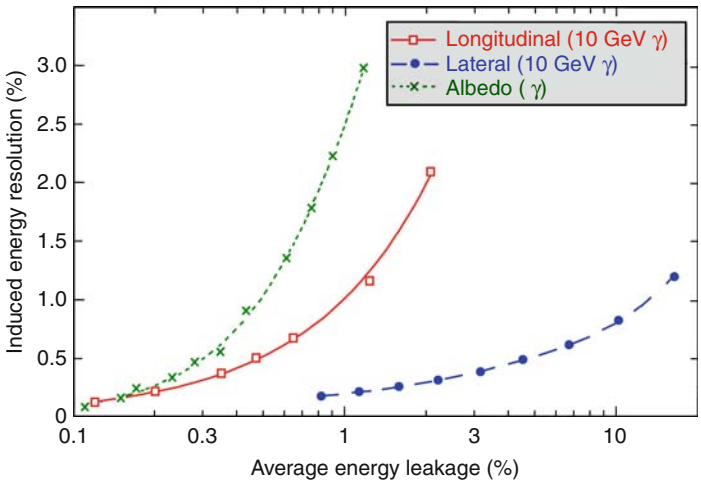


Fig. 3

A comparison of the effects caused by different types of shower leakage on the energy resolution for electrons and photons. Shown are the induced energy resolutions resulting from longitudinal and lateral leakage and from albedo, as a function of the energy fraction carried by the shower particles escaping from the detector. Results from EGS4 Monte Carlo calculations (Wigmans 2000)

level of non-containment. If the average longitudinal containment is 95%, one should thus not expect energy resolutions better than 5%.

In Sect. 3, the relationship between the absorber depth and the average level of shower containment is discussed, both for electromagnetic and hadronic showers.

- The *signal speed*. Especially in experiments in which the event rates are very high, and the calorimeter is crucial for selecting the small fraction of potentially interesting events, signal speed is crucial. This is particularly true for experiments at CERN's Large Hadron Collider, where this fraction amounts to only ~10^{-7}, at the nominal 1 GHz event rate. A high signal speed can be obtained either by using an intrinsically fast signal-generating mechanism, such as Čerenkov radiation, or by electronic means, i.e., fast signal shaping, in the case of a slower mechanism (e.g., drifting ionization charge in noble liquids). A slow signal-generating mechanism may also lead to detection inefficiencies. For example, wire chambers operating in the Geiger mode need some time to recover after recording a passing shower particle, and are insensitive to other particles during that period.
- *Hermeticity*. If one wants to use the calorimeter for accurate measurements of energy-flow variables such as missing (transverse) energy, it is very important that the entire phase space be covered. In colliding-beam experiments, this means that the calorimeter has to cover the entire 4π solid angle around the interaction vertex. This requirement creates a conflict with the necessity to transport detector signals to the outside world. For this reason, a number of new readout schemes have been developed over the years. Some of the schemes used for calorimeters based on light detection are shown in Fig. 4a. The classical "sandwich" structure (Fig. 4a) does not allow for a very hermetic coverage. Better opportunities are offered

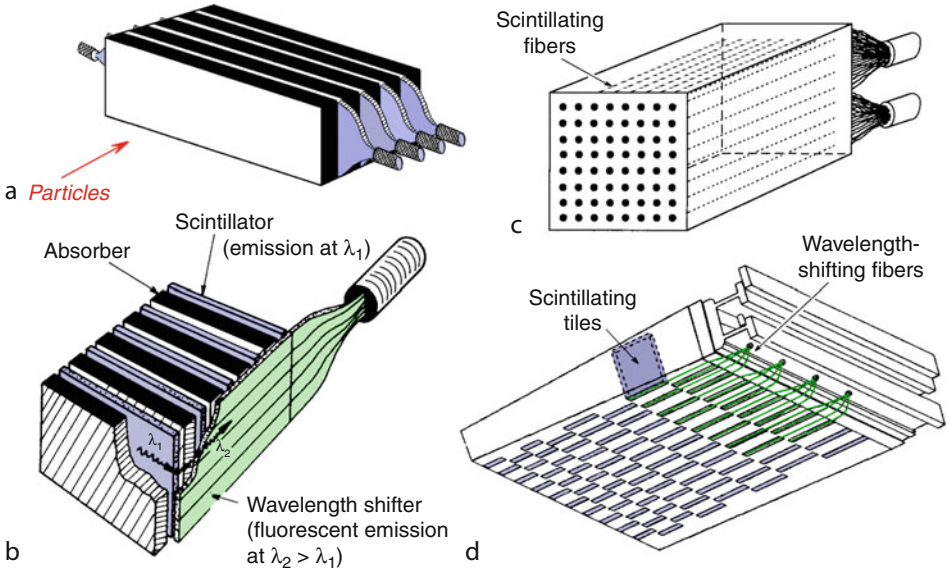


Fig. 4

Schematic representation of frequently used readout techniques for scintillation calorimeters: scintillator plates individually coupled to PMTs (a). Scintillator plates read out by WLS plates (b). Scintillating fibers coupled via a light guide to a PMT (c). Longitudinally oriented scintillator tiles read out by WLS fibers (d)

by using wavelength-shifting plates (b) or fibers (d), or longitudinally oriented scintillating fibers (c). For calorimeters based on direct charge collection, hermeticity is in general somewhat easier to achieve, although the cryogenic requirements for calorimeters based on noble liquids tend to create their own set of problems.

3 Calorimeter Types

One frequently distinguishes between *homogeneous* and *sampling* calorimeters. In a homogeneous calorimeter, the entire detector volume is sensitive to the particles and may contribute to the generated signals. In a sampling calorimeter, the functions of particle absorption and signal generation are exercised by *different* materials, called the *passive* and *active medium*, respectively. The passive medium is usually a high-density material, such as iron, copper, lead, or uranium. The active medium generates the light or charge that forms the basis for the signals from such a calorimeter.

In some non-accelerator experiments, the calorimeter is also the *source* that generates the particles to be detected. As examples, we mention large water Čerenkov counters built to search for proton decay and the high-purity ^{76}Ge crystals (Günther et al. 1997) or the ^{136}Xe liquid (LePort et al. 2007) used to study $\beta\beta$ decay.

3.1 Electromagnetic Calorimeters

Electromagnetic (em) calorimeters are specifically intended for the detection of energetic electrons and γs, but produce usually also signals when traversed by other types of particles. They are used over a very wide energy range, from the semiconductor crystals that measure X-rays down to a few keV to shower counters such as AGILE, PAMELA, and FERMI, which orbit the Earth on satellites in search for electrons, positrons, and γs with energies >10 TeV (Atwood et al. 2009).

Because of the peculiarities of em shower development (see ❯ Chap. 1, "Interactions of Particles and Radiation with Matter"), these calorimeters don't need to be very deep, especially when high-Z absorber material is used. For example, when 100 GeV electrons enter a block of lead, ~90% of their energy is deposited in only 4 kg of material.

As was shown in ❯ Sect. 2.2, good energy resolution can only be achieved when the shower is, on average, sufficiently contained. ❯ *Figure 5a* shows the required depth of the calorimeter, as a function of the electron energy, needed for 99% longitudinal containment. The figure shows this depth requirement, needed for energy resolutions of 1%, for calorimeters using Pb, Sn, Cu, and Al as absorber material. The fact that the four curves are not identical indicates that the radiation length (X_0), which is commonly used to describe the longitudinal development of em showers (see ❯ Chap. 1, "Interactions of Particles and Radiation with Matter"), is not a perfect scaling variable. It should also be noted that γ-induced showers require about one radiation length more material in order to be contained at a certain level than do electron showers of the same energy (Wigmans and Zeyrek 2002).

The effects of lateral shower leakage on the energy resolution are much smaller than for longitudinal leakage. According to ❯ *Fig. 3*, as much as 10% leakage can be tolerated before the induced energy resolution resulting from leakage *fluctuations* exceeds 1%. ❯ *Figure 5b* shows the average lateral leakage fraction as a function of the radius of an infinitely deep calorimeter

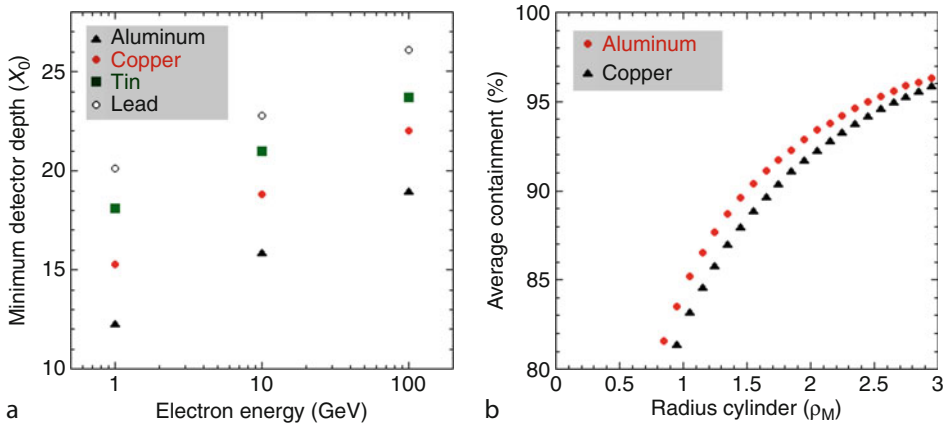


Fig. 5

Size requirements for electromagnetic shower containment. The depth of a calorimeter needed to contain electron showers, on average, at the 99% level, as a function of the electron energy. Results are given for four different absorber media (a). Average lateral containment of electron-induced showers in a copper- and an aluminum-based calorimeter, as a function of the radius (expressed in Molière radii, ρ_M) of an infinitely deep cylinder around the shower axis (b). From Wigmans (2000)

centered on the shower axis. It turns out that a radius of $2\rho_M$ is more than adequate for 90% lateral containment. This number is almost independent of the shower energy and the absorber material, in stark contrast with the depth requirements. The figure also indicates that em showers tend to have considerable radial tails. In order to capture 99% instead of 90% of the shower energy, the detector mass has to be increased by an order of magnitude.

By far the best energy resolutions have been obtained with large semiconductor crystals, and in particular high-purity germanium. These are the detectors of choice in nuclear γ-ray spectroscopy, and routinely obtain resolutions (σ/E) of 0.1% in the 1 MeV energy range. An example is shown in *Fig. 2a*, which also makes a comparison with the next best class of detectors, scintillating crystals. The latter are often the detectors of choice in experiments involving γ rays in the energy range from 1 to 20 GeV, which they measure with energy resolutions of the order of 1%. Excellent performance in this energy range has also been reported for liquid-krypton and -xenon detectors, which are bright (UV) scintillators. Other homogeneous detectors of em showers are based on Čerenkov light, in particular lead glass. Very large water Čerenkov calorimeters (e.g., SuperKamiokande, Fukuda et al. 2003) should also be mentioned in this category.

Table 1 lists materials commonly used as homogeneous calorimeters in particle physics experiments and some of their relevant properties. Not mentioned in this table is the light yield of these materials. The for practical purposes relevant strength of the signals depends sensitively on the effects of self-absorption and other factors that cause light losses, as well as on the quantum efficiency of the detector that converts the light into electrical pulses. These effects may vary strongly for different detectors based on materials listed in the table. Of these materials, NaI(Tl) is the brightest source of photons: ≈50,000 per MeV deposited energy. CsI(Tl) and the scintillating liquids generate light at the same order of magnitude. The light yield in BGO is one order of magnitude less, while $PbWO_4$ generates three orders of magnitude less light than the

Table 1
Relevant properties of a variety of light-based detectors that are used as homogeneous electromagnetic calorimeters in particle physics experiments

Detector	ρ (g cm^{-3})	X_0 (cm)	λ_{int} (cm)	λ_{emis} (nm)	τ (ns)	σ/E at 1 GeV
Scintillating crystals						
NaI(Tl)	3.67	2.59	41.4	410	230	2.7%
CsI(Tl)	4.51	1.85	37.0	560	1,300	2.7%
BGO	7.13	1.12	21.8	480	300	2.1%
$PbWO_4$	8.30	0.89	18.0	420	10	3.1%
Scintillating liquids						
LKr	2.41	4.7	61	147	~85	1.9%
LXe	2.95	2.77	57	174	~30	1.6%
Č calorimeters						
Lead glass	4.06	2.5	33	λ^{-2}	0	5%
Water	1.0	36.1	84.9	λ^{-2}	0	2.6%

brightest scintillators. The light yield deriving from the Čerenkov mechanism is four orders of magnitude less than that of sodium iodide, and photoelectron statistics tends to dominate the electromagnetic energy resolution of detectors based on this mechanism. As illustrated by the last column of the table, light yield is not the dominant limiting factor for the energy resolution of the mentioned scintillators, at least not in the GeV domain.

Sampling calorimeters, which are typically much cheaper, become competitive at higher energies. In properly designed instruments of this type, the energy resolution is determined by *sampling fluctuations*. These represent fluctuations in the number of different shower particles that contribute to the calorimeter signals, convoluted with fluctuations in the amount of energy deposited by individual shower particles in the active calorimeter layers. They depend both on the *sampling fraction*, which is determined by the ratio of active and passive material, and on the *sampling frequency*, determined by the number of different sampling elements in the region where the showers develop. Sampling fluctuations are stochastic and their contribution to the energy resolution is thus described by Eq. 1, with

$$a = 0.027\sqrt{d/f_{\text{samp}}} \tag{2}$$

in which d represents the thickness of individual active sampling layers (in mm), and f_{samp} the sampling fraction for minimum-ionizing particles (*mip*s). This expression describes data obtained with a large variety of different (non-gaseous) sampling calorimeters reasonably well (*Fig. 6*).

Table 2 lists some characteristics of a representative selection of sampling calorimeters used in particle physics experiments. Above 100 GeV, the resolution of all calorimeters mentioned above is ~1%, and systematic factors, such as stability of the electronic components, the effects of light attenuation, or temperature variations of the light yield, tend to dominate the performance.

An important characteristic of these sampling calorimeters is that the sampling fraction, i.e., the fraction of the energy that is deposited in the active calorimeter layers, *decreases* as the shower develops. This is because the spectrum of the γs that produce the shower particles that constitute the signals (electrons, positrons) becomes much softer. In the early stages,

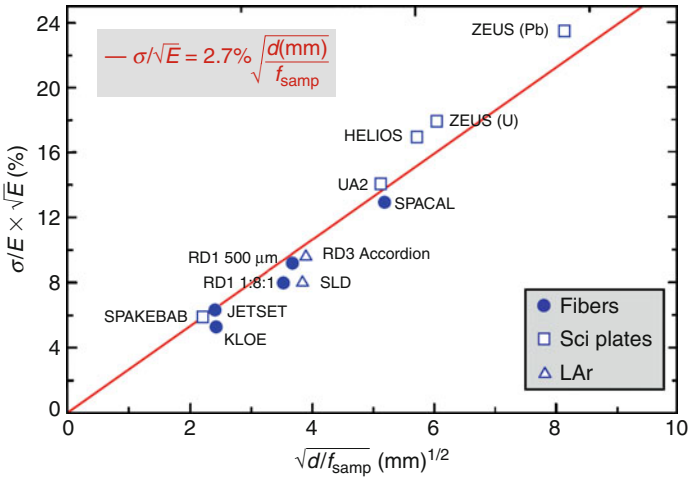


Fig. 6

The em energy resolution of non-gaseous sampling calorimeters as a function of the parameter $(d/f_{samp})^{1/2}$, in which d is the thickness of an active sampling layer (e.g., the diameter of a fiber or the thickness of a liquid-argon gap), and f_{samp} the sampling fraction for mips (E is expressed in GeV) (Livan et al. 1995)

Table 2

A representative selection of electromagnetic sampling calorimeters used in past and present particle physics experiments. The energy E is expressed in GeV, the sampling fraction f_{samp} refers to minimum-ionizing particles. For the energy resolution, only the $E^{-1/2}$ scaling term, which dominates in the practically important energy range for these experiments, is listed

Experiment	Calorimeter structure	X_0 (cm)	f_{samp}	σ/E
KLOE (Frascati)	Pb/fibers	1.6	17%	$4.7\%/\sqrt{E}$
ZEUS (DESY)	^{238}U/scintillator	0.7	9%	$18\%/\sqrt{E}$
NA31 (CERN)	Pb/LAr	1.5	23%	$7.5\%/\sqrt{E}$
ALEPH (LEP)	Pb/gas	3–6	0.01–0.02%	$18\%/\sqrt{E}$
SLD (SLAC)	Pb/LAr	1.3	19%	$8\%/\sqrt{E}$

these γs mainly convert into energetic e^+e^- pairs, but beyond the shower maximum, Compton scattering and photoelectron production become more and more dominant (see Chap. 1, "Interactions of Particles and Radiation with Matter"). Because of the Z dependence of the latter processes (Z^5 for the photoelectric effect), the soft γs almost exclusively react in the high-Z absorber material (lead or uranium), and the resulting electrons are thus much less efficiently sampled by the signal-producing calorimeter layers than the e^+e^- pairs that dominate the early part of the shower. This phenomenon may cause large, nontrivial problems for the calibration of longitudinally segmented calorimeters (Albrow et al. 2002; Cervelli et al. 2002; Aharrouche et al. 2006; Wigmans 2006).

3.2 Hadron Calorimeters

The energy range covered by hadron calorimeters is in principle even larger than that for em ones. Calorimetric techniques are used to detect thermal neutrons, which have kinetic energies of a small fraction of 1 eV, to the highest-energy particles observed in nature, which reach the Earth from outer space as cosmic rays carrying 10^{20} eV or more. In accelerator-based particle physics experiments, hadron calorimeters are typically used to detect protons, pions, kaons, and fragmenting quarks and gluons (commonly referred to as *jets*) with energies in the GeV–TeV range. In this subsection, we mainly discuss the latter instruments.

The development of hadronic cascades in dense matter differs in essential ways from that of electromagnetic ones, with important consequences for calorimetry. Hadronic showers consist of two distinctly different components:

1. An *electromagnetic* component; π^0s and ηs generated in the absorption process decay into γs which develop em showers.
2. A *non-electromagnetic* component, which combines essentially everything else that takes place in the absorption process.

For the purpose of calorimetry, the main difference between these components is that some fraction of the energy contained in the non-em component does *not* contribute to the signals. This *invisible energy*, which mainly consists of the binding energy of nucleons released in the numerous nuclear reactions, may represent up to 40% of the total non-em energy, with large event-to-event fluctuations.

The appropriate length scale of hadronic showers is the nuclear interaction length (λ_{int}), which is typically much larger (up to 30 times for high-Z materials) than the radiation length. Many experiments make use of this fact to distinguish between electrons and hadrons on the basis of the energy deposit profile in their calorimeter system. ❯ *Figure 1a* shows an example of this. Since the ratio λ_{int}/X_0 is proportional to Z, particle identification on this basis works best for high-Z absorber materials. Lead and depleted uranium are therefore popular choices for the absorber material in preshower detectors and the first section of a longitudinally segmented calorimeter, which is therefore commonly referred to as the *electromagnetic section*.

Just as for the detection of em showers, high-resolution hadron calorimetry requires an average longitudinal containment better than 99%. In iron and materials with similar Z, which are most frequently used for hadron calorimeters, 99% longitudinal containment requires a thickness ranging from $5\lambda_{\text{int}}$ at 20 GeV to $8\lambda_{\text{int}}$ at 150 GeV. ❯ *Figure 7a* shows how these numbers change when the containment requirement is relaxed to 95%. (❯ *Figure 7a* is based on data from the CDHS experiment (Abramowicz et al. 1981). More recent data (Adragna et al. 2010) suggest that iron-based calorimeters have to be significantly deeper to contain 99% of the shower energy. However, the (ATLAS) detector used for the latter studies had a much coarser longitudinal sampling (1.5 λ, vs 0.3–0.7 λ for CDHS), and was mainly focused on leakage beyond great depths.) Just as for em showers, lateral shower containment is less critical for the energy resolution than longitudinal containment, albeit to a much lesser extent than suggested by ❯ *Fig. 3*. ❯ *Figure 7b* shows that 95% lateral containment requires a radius of 1.5–2λ_{int}. Interestingly, the average lateral shower leakage fraction from a given calorimeter *decreases* at increasing energy. This is the result of increased π^0 production. The large event-to-event fluctuations in π^0 production, which have no equivalent in em shower development, are also responsible for the fact that the hadronic energy resolution is much more sensitive to the effects of side leakage than

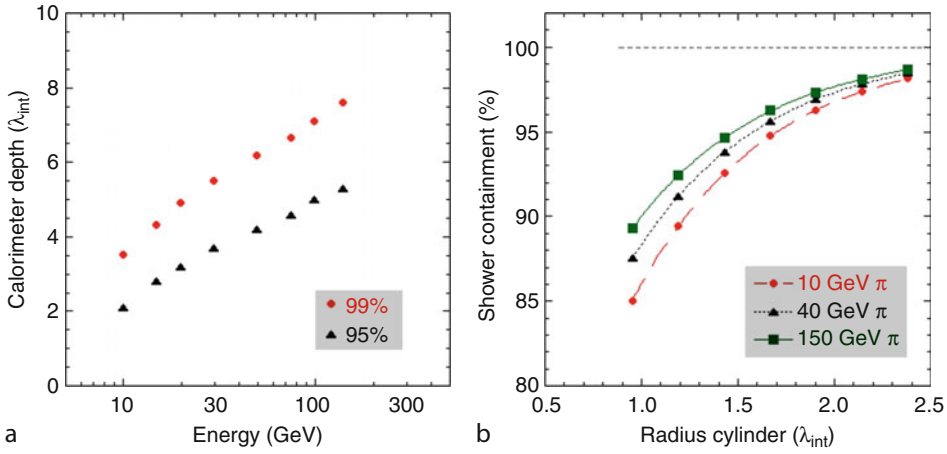


Fig. 7

Size requirements for hadronic shower containment. The depth of an iron-based calorimeter needed to contain pion showers, on average, at the 95% or 99% level, as a function of the pion energy (a). Average lateral containment of pion-induced showers in a lead-based calorimeter as a function of the radius of an infinitely deep cylinder around the shower axis, for three different pion energies (b)

the em energy resolution. Hadronic energy resolutions of 1% thus require not only longitudinal shower containment at the 99% level, but also lateral containment of 98% or better.

Energetic π^0s may be produced throughout the absorber volume, and not exclusively in the em calorimeter section. They lead to local regions of highly concentrated energy deposit. Therefore, there is no such thing as a "typical hadronic shower profile" (Green 1994). This feature affects not only the shower containment requirements, but also the calibration of longitudinally segmented calorimeters (Albrow et al. 2002; Cervelli et al. 2002; Aharrouche et al. 2006; Wigmans 2006), and the applicability of *Particle Flow Analysis* techniques (Adloff et al. 2007; TESLA 2001) in which one tries to improve the quality of calorimetric energy measurements of jets with an upstream tracker, which can measure the momenta of the charged jet constituents with great precision.

The properties of the em shower component have also important consequences for the *energy resolution*, the signal *linearity*, and the *response function*. The average fraction of the total shower energy contained in the em component has been measured to increase with energy following a power law (Acosta et al. 1992; Akchurin et al. 1997), confirming an induction argument made to that effect (Gabriel et al. 1994):

$$\langle f_{\mathrm{em}} \rangle = 1 - \left[E/E_0 \right]^{k-1}, \tag{3}$$

where E_0 is a material-dependent constant related to the average multiplicity in hadronic interactions (varying from 0.7 GeV to 1.3 GeV for π-induced reactions on Cu and Pb, respectively), and $k \sim 0.82$. For proton-induced reactions, $\langle f_{\mathrm{em}} \rangle$ is typically considerably smaller, as a result of baryon-number conservation in the shower development (Akchurin et al. 1998).

Let us define the calorimeter *response* as the conversion efficiency from deposited energy to generated signal, and normalize it to electrons. The responses of a given calorimeter to the em and non-em hadronic shower components, e and h, are usually not the same, as a result of invisible energy and a variety of other effects. Such calorimeters are called *non-compensating* ($e/h \neq 1$). Since their response to pions, $\langle f_{\mathrm{em}} \rangle + \left[1 - \langle f_{\mathrm{em}} \rangle\right] h/e$, is energy dependent (Eq. 3), they are intrinsically nonlinear.

Event-to-event fluctuations in f_{em} are large and non-Poissonian. If $e/h \neq 1$, these fluctuations tend to dominate the hadronic energy resolution and their asymmetric distribution characteristics are reflected in the response function (*Fig. 9a*). It is often assumed that the effect of non-compensation on the energy resolution is energy independent ("constant term"). This is incorrect. The measured effects of *fluctuations* in f_{em} (*Fig. 8a*) can be described by a term that is very similar to the one used for its energy dependence (Eq. 3). This term should be added in quadrature to the $E^{-1/2}$ scaling term which accounts for all Poissonian fluctuations:

$$\frac{\sigma}{E} = \frac{a_1}{\sqrt{E}} \oplus a_2 \left[\left(\frac{E}{E_0} \right)^{l-1} \right], \tag{4}$$

where the parameter $a_2 = |1 - h/e|$ is determined by the degree of non-compensation (Groom 2007), and $l \sim 0.72$. Equation 4 is represented by the solid curve in *Fig. 8b*, for parameter values that are typical for many calorimeters. In the energy range covered by the current generation of test beams, i.e., up to 400 GeV, it runs almost parallel to the dotted line, which represents a single stochastic term with a somewhat larger coefficient ($a_1 = 0.55$ instead of 0.50).

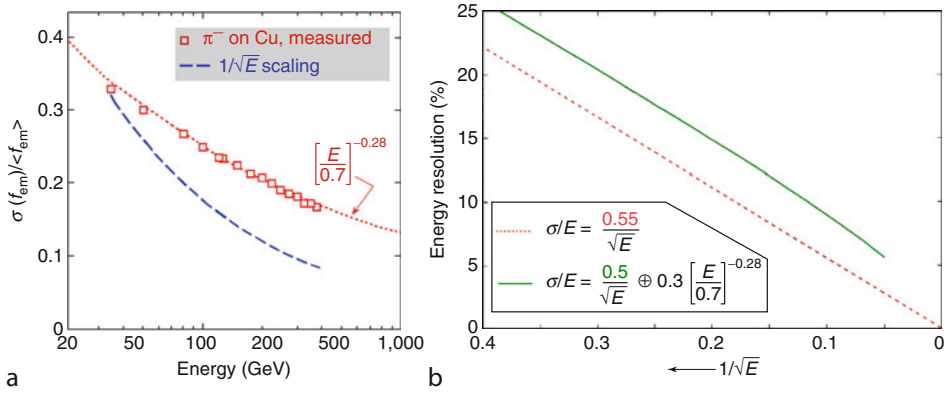


Fig. 8

Energy dependence of the fractional width of the f_{em} distribution (with energies given in GeV). Shown are the results of measurements (Akchurin et al. 1997) and the expected dependence for fluctuations governed by Poisson statistics (a). The hadronic energy resolution calculated for a typical non-compensating calorimeter in the energy regime up to 400 GeV (the *solid line*), and calculated with a sole stochastic term with a slightly larger scaling constant (b) (Wigmans 2000)

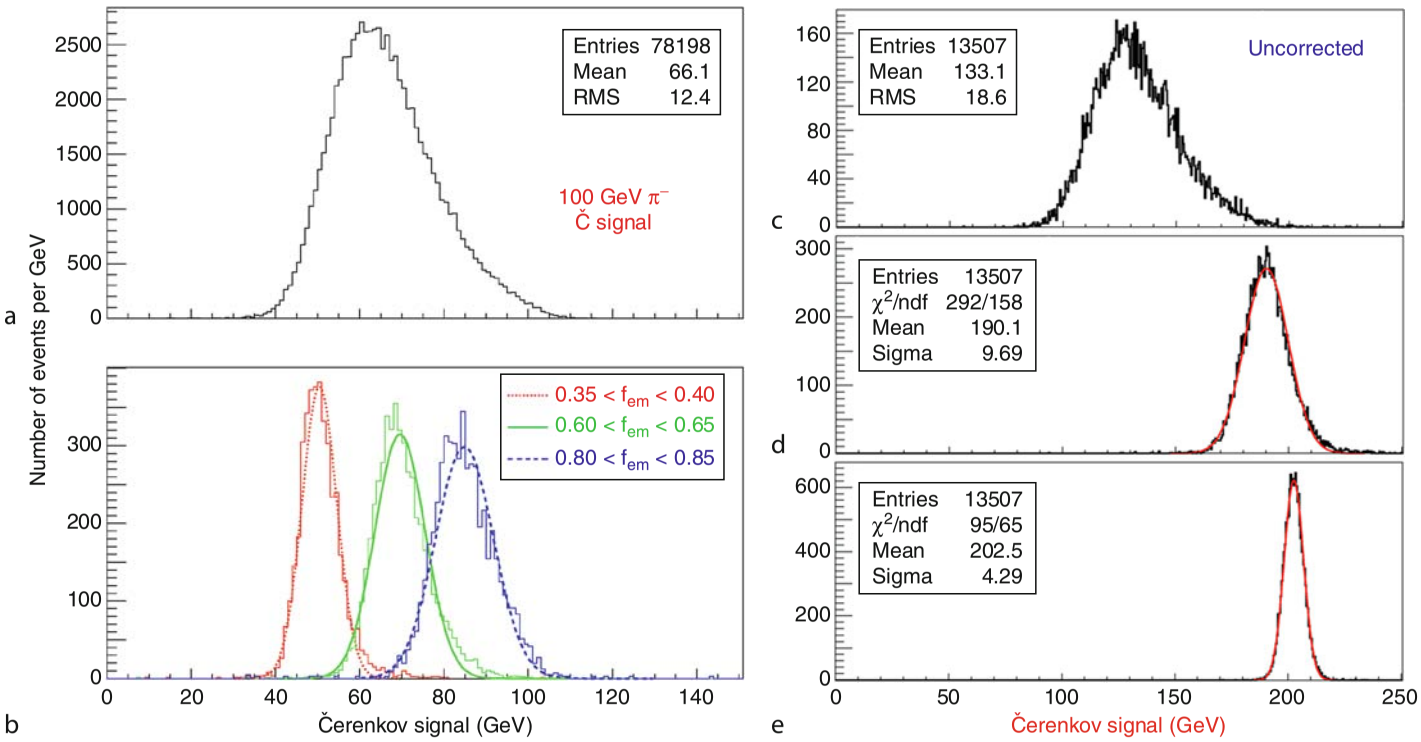


Fig. 9

Čerenkov signal distribution for 100 GeV π^- (a) and distributions for subsamples of events selected on the basis of the measured f_{em} value (b). Signal distributions for high-multiplicity 200 GeV "jets" in the DREAM calorimeter before (c) and after (d) corrections as described in the text were applied. In diagram (e), energy constraints were used, which eliminated the effects of lateral shower leakage fluctuations that dominate the resolution in (d) (Akchurin et al. 2005)

Table 3

Hadron calorimeters used in some particle physics experiments. If the calorimeter system has an em section of different composition, its structure is given as well. For 4π calorimeters, the depth is given at pseudorapidity 0. The energy E is in GeV

Experiment	Calorimeter structure	Depth	Resolution σ/E
CDF (FNAL)	(Pb+Fe)/scintillator	$5.3\lambda_{\text{int}}$	$50\%/\sqrt{E} \oplus 3\%$
D0 (FNAL)	^{238}U/liquid-argon	$7.3\lambda_{\text{int}}$	$45\%/\sqrt{E} \oplus 1.3/E \oplus 4\%$
ATLAS (CERN)	Pb/LAr + Fe/scintillator	$8\lambda_{\text{int}}$	$47\%/\sqrt{E} + 2.2\%$
CMS (CERN)	$PbWO_4$ + brass/scintillator	$7\lambda_{\text{int}}$	$100\%/\sqrt{E} + 5\%$
MINOS (FNAL)	Fe/scintillator	$75\lambda_{\text{int}}$	$55\%/\sqrt{E}$ (MC)

This solves an old mystery, since it means that experimental data might also be described by an expression of the type

$$\frac{\sigma}{E} = \frac{a_1}{\sqrt{E}} + a_2, \tag{5}$$

i.e., a *linear sum* of a stochastic term and a constant term. Many sets of experimental hadronic energy-resolution data exhibit exactly this characteristic. This is illustrated in *Table 3*, which lists properties of some calorimeter systems used in the present generation of Particle Physics experiments.

High-quality hadron calorimetry is deemed essential for the success of experiments at a future Linear e^+e^- Collider. The mentioned effects of non-compensation on resolution, linearity, and line shape, as well as the associated calibration problems (Abdullin et al. 2009) are absent in compensating calorimeters ($e/h = 1.0$). Compensation can be achieved in sampling calorimeters with high-Z absorber material and hydrogenous active material. It requires a very specific sampling fraction, so that the response to shower neutrons is boosted by the precise factor needed to equalize e and h. For example, in Pb/scintillating-plastic structures, this sampling fraction is ~2% for showers (Bernardi et al. 1987; Acosta et al. 1991; Suzuki et al. 1999). This small sampling fraction sets a lower limit on the contribution of sampling fluctuations, while the need to efficiently detect MeV-type neutrons requires signal integration over a relatively large volume during at least 30 ns. Yet, calorimeters of this type currently hold the world record for hadronic energy resolution ($\sigma/E \sim 30\%/\sqrt{E}$, Acosta et al. (1991)).

The Dual-Readout approach aims to achieve the advantages of compensation without these disadvantages. The energy carried by the non-em shower component is mostly deposited by nonrelativistic shower particles (protons), and therefore does not contribute to the signals of a Čerenkov calorimeter. By measuring simultaneously dE/dx and the Čerenkov light generated in the shower absorption process, one can determine f_{em} *event by event* and thus eliminate (the effects of) its fluctuations. The correct hadron energy can be determined from a combination of both signals.

This principle was first experimentally demonstrated by the DREAM Collaboration (Akchurin et al. 2005), with a Cu/fiber calorimeter. Scintillating fibers measured dE/dx, quartz fibers the Čerenkov light. The response ratio of these two signals is related to f_{em} as

$$\frac{Q}{S} = \frac{f_{\text{em}} + 0.21\,(1 - f_{\text{em}})}{f_{\text{em}} + 0.77\,(1 - f_{\text{em}})}, \tag{6}$$

where 0.21 and 0.77 represent the h/e ratios of the Čerenkov and scintillator calorimeter structures, respectively. The hadron energy can be derived directly from the two signals (Groom 2007):

$$E = \frac{\chi \cdot S - Q}{\chi - 1}, \quad \text{with} \quad \chi = \frac{\left[1 - (h/e)_Q\right]}{\left[1 - (h/e)_S\right]} = 3.43 \text{ and } E \text{ in GeV}. \tag{7}$$

The merits of this method are illustrated in ❯ *Fig. 9*, which shows that the energy resolution improved, the signal distribution became much more Gaussian, and, most importantly, the hadronic energy was correctly reproduced in this way. This was true both for single pions as well as for jets. It was shown that similar results can also be obtained with high-Z crystal calorimeters ($PbWO_4$, BGO), whose signals can be separated into scintillation and Čerenkov components (Akchurin et al. 2007a,b).

3.3 Cryogenic Calorimeters

There is a class of highly specialized detectors that employ calorimetric methods to study a series of very specific phenomena in the boundary area between particle physics and astrophysics: dark matter, solar neutrinos, magnetic monopoles, nuclear double β decay, etc. All these issues require precise measurements of small energy deposits. In order to achieve that goal, the mentioned detectors exploit phenomena that play a role at temperatures close to zero, in the few-mK to 1 K range. These phenomena include:

1. The fact that some elementary excitations require very little energy. For example, Cooper pairs in superconductors have binding energies in the μeV–meV range and may be broken by phonon absorption.
2. The fact that the specific heat for dielectric crystals and for superconductors decreases to very small values at these low temperatures.
3. The fact that thermal noise in the detectors and the associated electronics becomes very small.
4. The fact that some materials exhibit specific behavior (e.g., change in magnetization, latent heat release) that may provide detector signals.

Most of the devices that have been proposed in this context are still in the early phases of the R&D process. In many cases, this R&D involves fundamental research in solid-state physics and materials science. Among the devices that have reached the prototype stage are

- *Bolometers*, which are based on principle (2). These are calorimeters in the true sense of the word, since the energy deposit of particles (in an insulating crystal at very low temperature) is measured with a resistive thermometer. Detectors with masses in excess of 1 g have been developed, and sensitivity to energy deposits of a fraction of 1 eV has been demonstrated.
- *Superconducting Tunnel Junctions*, in which the quasi-particles and -holes (Cooper pairs) excited by incident radiation tunnel through a thin layer separating two superconducting materials.
- *Superheated Superconducting Granules*, which are based on the fact that certain type-I superconductors can exhibit metastable states, in which the material remains superconducting in

external magnetic fields exceeding the critical field. These detectors are usually prepared as a colloid of small (diameter 1–100 μm) metallic granules suspended in a dielectric matrix (e.g., paraffin). Heat deposited by an interacting particle may drive one of several granules from the superconducting to the normal state. The resulting change in magnetic flux (disappearance of the Meissner effect) may be recorded by a pickup coil.

There is a considerable amount of effort going into the development of these and many related, similarly ingenious devices. However, this highly specialized work falls somewhat outside the scope of this book. The interested reader is referred to reviews of this field that can be found in Twerenbold (1996), Pretzl (2004), Enss (2005).

3.4 Natural Calorimeters

All calorimeters discussed in the previous sections were man-made. In this section, we look into efforts to use our natural environment as a calorimeter. The driving force behind these efforts is the opportunity to create a very large instrument in this way. Typical detector volumes are measured in units of km^3, i.e., at least four orders of magnitude larger than SuperKamiokande (Fukuda et al. 2003), which has one of the largest instrumented volumes of any man-made calorimeter. Such large volumes are needed to achieve the scientific goals of the experiments, which usually focus on the study of very rare natural phenomena. Examples of such phenomena include the absorption of extremely high-energy protons, αs, or heavier atomic nuclei of extraterrestrial origin in the Earth's atmosphere and interactions of extra-galactic neutrinos in the Earth itself.

Almost all natural calorimeters are based on light as the source of experimental information. In one noteworthy exception, radio signals are exploited. Usually, the Čerenkov mechanism is the source of the signals, especially in detectors where sea water or Arctic ice serves as the absorber medium. Čerenkov light is usually also an important source of experimental information in detectors using the Earth's atmosphere as a calorimeter. In some experiments of the latter type, scintillation light is used as well.

The technique to measure the Čerenkov light signals produced by extremely high-energy neutrinos of cosmic origin in water was pioneered in Lake Baikal (Siberia) by a consortium of Russian and German scientists (http://baikalweb.jinr.ru/). A telescope consisting of a lattice of 200 large PMTs spread over a large open volume at a depth of ~1 km looks for upward-traveling muons produced in neutrino interactions in water in the vicinity of the detector. Since the angle between the parent neutrino and the muon produced in the interaction is very small, high-resolution astronomy is in principle possible. The direction of the particle can be inferred from the measured arrival times and amplitudes of the Čerenkov photons observed in the various PMTs. The same technique is applied, on a much larger scale, in the Antarctic ice near the Amundsen–Scott South Pole station. The IceCube experiment (Klein et al. 2009) has currently installed 4,740 PMTs covering a volume of ~1 km^3 at a depth ranging from 1,450 to 2,450 m below the surface (*Fig. 10*).

Another structure of similar dimensions is planned for installation somewhere in the Mediterranean. Exploratory work in that context has been carried out by the ANTARES (http://antares.in2p3.fr/), NEMO (http://nemoweb.lns.infn.it/), and NESTOR (http://www.nestor.noa.gr/) Collaborations. Some advantages of ice over water in a lake or sea include the absence of light-emitting organisms and underwater currents that may jeopardize the integrity of the

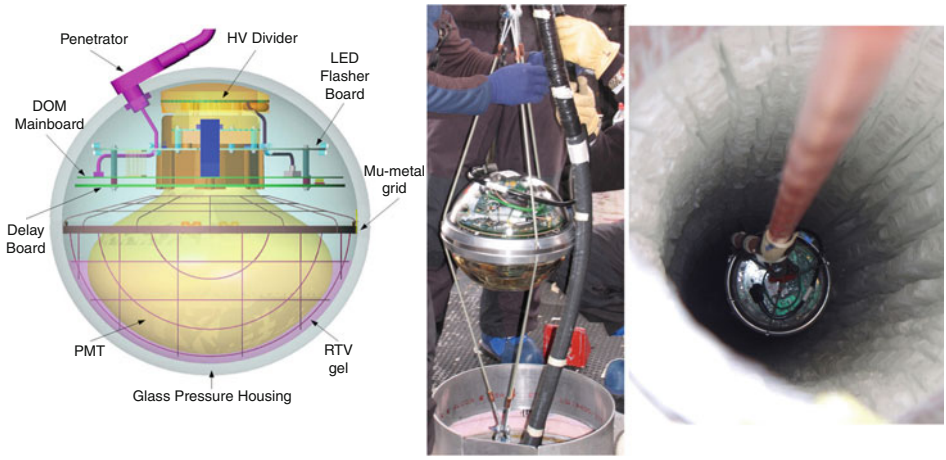


Fig. 10

One of the Digital Optical Modules used in the IceCube experiment. Shown are a schematic drawing of this module (*left*), the real thing assembled (*center*), and being lowered to its final position in the ice (*right*)

detector structure. On the other hand, light scattering by air bubbles trapped in ice may limit the possibilities to reconstruct the direction of the incoming particles.

The operation of detectors of this type is of course very different from those in accelerator laboratories. Whereas muons are usually referred to as "minimum ionizing particles" (mips) in accelerator-based experiments, IceCube uses the non-mip nature of these particles to calibrate the energy scale of their detector, exploiting the fact that the specific energy loss $(\mathrm{d}E/\mathrm{d}x)$ depends logarithmically on the muon energy in the region of interest (TeV–EeV). The angular resolution of their instrument ($<0.5^\circ$) is measured using the shadow of the Moon, which measurably affects the rate of down-going atmospheric muons. With this kind of resolution, one might hope to detect point sources of extraterrestrial neutrinos.

The Antarctic ice cap is also the source of signals for the ANITA experiment (http://www.phys.hawaii.edu/~anita/web/), which aims to detect the radio component of the coherent Čerenkov signals produced as a result of the charge asymmetry in high-energy em showers, the so-called Askaryan effect (Askaryan 1961). This mechanism may also be exploited in other media that are transparent to such radio signals, for example, large rock-salt formations (Gorham and Saltzberg 2002).

Experiments using the Earth's atmosphere as a calorimeter are primarily looking for extensive air showers caused by very-high-energy charged cosmic particles. At sea level, the atmosphere represents an absorber with a total thickness of ~11λ, or ~28X_0, enough to absorb even the highest-energy particles to a very large extent. The Čerenkov light produced in the absorption process is relatively easy to detect, *provided that it is emitted in the direction of the telescope that is looking for it* (◘ *Fig. 11*). Because of the very small refraction index of air, the Čerenkov angle is very small and only energetic shower particles produced in the early stages of the shower development emit Čerenkov light. As a result, this light is highly collimated, a shower starting at the typical altitude of 10 km produces a light cone at ground level with a radius



Fig. 11
One of the Čerenkov telescopes used in the HEGRA experiment at La Palma. Photo courtesy K. Bernlohr

of only ~100 m. Therefore, modern experiments looking for hadronic showers complement the Čerenkov telescopes with additional detectors, looking for (isotropically emitted) fluorescent light (produced by transitions in molecular nitrogen and in N_2^+ ions), for muons from decaying shower particles (π, K), and/or shower particles themselves. Examples of such experiments include AUGER (http://www.auger.org/) and KASKADE-Grande (http://www-ik.fzk.de/KASCADE/) (hadronic showers) and VERITAS (http://veritas.sao.arizona.edu/) and HESS (http://www.mpi-hd.mpg.de/hfm/HESS/) (em showers). The angular resolutions of these experiments are even better than those obtained with IceCube. Apart from the shadowing effect of the Moon (Hoffman et al. 1999), one can also use the signals from some known point sources of γ rays for this purpose. The signals from the strongest of these sources, the Crab nebula, are also a valuable tool for the energy calibration. Variations of the order of 20% between the different experiments that use this technique (Aharonian et al. 2000) are indicative for the relative uncertainty in the energy scale.

4 Concluding Remarks

Calorimeters are instruments for measuring energy. The history of physics in general, and of nuclear and particle physics in particular, is filled with examples proving that measurement precision pays off. A better, more accurate instrument allows more precise measurements. More precise measurement results make it possible to discover new phenomena, and/or to better understand old ones.

The history of calorimetry itself illustrates this process in a nutshell. Calorimeters were originally invented as crude, cheap instruments for some specialized applications (e.g., detection of neutrino interactions). Initially, their performance was often perceived as somewhat mysterious by their users. Only after the physics on which calorimeters are based was understood in detail did it become possible to develop these detectors into the precision instruments that they are nowadays and which form the centerpiece of many modern experiments in particle physics.

In nuclear spectroscopy, the advent of germanium-based solid-state detectors with their unprecedented energy resolution caused a revolution in the 1960s (see *Fig. 2*). We are now at the brink of an era in which calorimetry will allow the measurement of fragmenting quarks and gluons and other elementary particles with nuclear-spectroscopic precision. And calorimeters are also making their mark in the relatively new field of astroparticle physics, where innovative low-temperature devices and ever larger instrumented volumes are used to explore new frontiers. If nature is kind to us, a new world might open up as a result of all this.

References

Abdullin S et al (2009) The CMS barrel calorimeter response to particle beams from 2 GeV/c to 350 GeV/c. Eur Phys J C 60:359

Abramowicz H et al (1981) The response and resolution of an iron scintillator calorimeter for hadronic and electromagnetic showers between 10 GeV and 140 GeV. Nucl Instrum Methods 180:429

Acosta D et al (1991) Electron, pion and multiparticle detection with a lead/scintillating-fiber calorimeter. Nucl Instrum Methods A 308:481

Acosta D et al (1992) Lateral shower profles in a scintillating fiber calorimeter. Nucl Instrum Methods A 316:184

Adloff C et al, CALICE Collaboration (2007) CALICE report to the R&D review panel. arXiv:0707.1245 [physics.ins-det]

Adragna P et al (2010) Measurement of pion and proton response and longitudinal shower profiles up to 20 nuclear interaction lengths with the ATLAS Tile calorimeter. Nucl Instrum Methods A 615:158

Aharonian FA et al (2000) The energy spectrum of TeV gamma-rays from the Crab nebula as measured by the HEGRA system of imaging air Cherenkov telescopes. Astrophys J 539:317

Aharrouche M et al (2006) Energy linearity and resolution of the ATLAS electromagnetic barrel calorimeter in an electron test-beam. Nucl Instrum Methods A 568:601

Akchurin N et al (1997) Beam test results from a fine-sampling quartz fiber calorimeter for the detection of electrons, photons and hadrons. Nucl Instrum Methods A 399:202

Akchurin N et al (1998) On the difference between proton and pion showers and their signals in a non-compensating calorimeter. Nucl Instrum Methods A 408:380

Akchurin N et al (2005) Hadron and jet detection with a Dual-Readout Calorimeter. Nucl Instrum Methods A 537:537

Akchurin N et al (2007a) Comparison of high-energy hadronic shower profiles measured with scintillation and Čerenkov light. Nucl Instrum Methods A 584:273

Akchurin N et al (2007b) Separation of crystal signals into scintillation and Čerenkov components. Nucl Instrum Methods A 595:359

Albrow M et al (2002) Intercalibration of the longitudinal segments of a calorimeter system. Nucl Instrum Methods A 487:381

Arnison G et al (1983) Experimental observation of isolated large transverse energy electrons with associated missing energy at $\sqrt{s}$ = 540 GeV. Phys Lett B 122:103

Askaryan G (1961) Radio-frequency emission and currents from showers and muons produced in a medium by a beam of high-energy neutrinos. Sov Phys – J Exp Theor Phys 14:441

Atwood WB et al (2009) The Large Area Telescope on the Fermi Gamma-ray Space Telescope Mission. Astrophys J 697:1071

Bernardi E et al (1987) Performance of a compensating lead-scintillator hadronic calorimeter. Nucl Instrum Methods A 262:229

Cervelli F et al (2002) A reduced-scale e.m. calorimeter prototype for the AMS-02 experiment. Nucl Instrum Methods A 490:132

Enss Chr (ed) (2005) Cryogenic particle detection. Topics in applied physics 99. Springer, Berlin
Fukuda S et al (2003) The Super-Kamiokande detector. Nucl Instrum Methods A 501:418
Gabriel TA et al (1994) Energy dependence of hadronic activity. Nucl Instrum Methods A 338:336
Gorham P et al (2002) Measurements of the suitability of large rock salt formations for radio detection of high-energy neutrinos. Nucl Instrum Meth A 490:476
Green D (1994) Selected Topics in Sampling Calorimetry. In: Menzione A, Scribano A (eds) Proceedings of the fourth international conference on calorimetry in high energy physics, La Biodola, Italy. World Scientific, Singapore, p 1
Groom DE (2007) Energy flow in a hadronic cascade: Application to hadron calorimetry. Nucl Instrum Methods A 572:633
Günther M et al (1997) Heidelberg-Moscow $\beta\beta$ experiment with ^{76}Ge: Full setup with five detectors. Phys Rev D 55:54
Hoffman CM, Sinnis C, Fleury P, Punch M (1999) Gamma-ray astronomy at high energies. Rev Mod Phys 71:897
Klein SR et al (2009) IceCube: A Cubic Kilometer Radiation Detector. IEEE Trans Nucl Sci 56:1141
LePort F et al (2007) A liquid xenon ionization chamber in an all-fluoropolymer vessel. Nucl Instrum Methods A 578:409
Livan M, Vercesi V, Wigmans R (1995) Scintillating-fibre calorimetry. CERN Yellow Report, CERN 95-02, Genéve, Switzerland
Pretzl K (2004) Dark matter searches with cryogenic detectors. In: Spooner NJC, Kudryavtsev V (eds) Proceedings of the fourth international workshop on the identification of dark matter. World Scientific, Singapore, p 205
Suzuki T et al (1999) A systematic measurement of energy resolution and e/π ratio of a lead/plastic-scintillator sampling calorimeter. Nucl Instrum Methods A 432:48
TESLA (2001) Technical Design Report. Report DESY 2001-011, DESY, Hamburg, Germany
Twerenbold D (1996) Cryogenic particle detectors. Rep Prog Phys 59:349
Wigmans R (2000) Calorimetry, energy measurement in particle physics. International series of monographs on physics, vol 107. Oxford University Press, Oxford
Wigmans R (2006) On the calibration of segmented calorimeters. In: Proceedings of the twelfth international conference on calorimetry in high energy physics, Chicago, 2006. AIP Conference Proceedings 867, 90
Wigmans R, Zeyrek MT (2002) On the differences between calorimetric detection of electrons and photons. Nucl Instrum Methods A 485:385
Young GR et al (1989) The zero-degree calorimeter for the relativistic heavy-ion experiment WA80 at CERN. Nucl Instrum Methods A 279:503

Further Reading

Clay R, Dawson B (1997) Cosmic bullets: high energy particles in astrophysics. Perseus, Cambridge
Fabjan C, Gianotti F (2003) Calorimetry for particle physics. Rev Mod Phys 75:1243
Wigmans R (2000) Calorimetry, energy measurement in particle physics. International series of monographs on physics, vol 107. Oxford University Press, Oxford
Enss Chr (ed) (2005) Cryogenic particle detection. Topics in applied physics 99. Springer, Berlin

21 New Solid State Detectors

Christoph J. Ilgner
Helmholtz-Zentrum Dresden-Rossendorf, Dresden, Germany

1 Radiation Environment at Contemporary Hadron Accelerators 520

2 Artificial Diamond as a Sensor Material 521
2.1 Chemical Vapor Deposition (CVD) Diamond 521
2.1.1 Production of Artificial Diamond 521
2.2 Diamonds as Solid State Detectors 523
2.3 Charge Collection in Polycrystalline CVD Diamonds 523
2.4 Radiation Effects 524
2.4.1 The Beam-Conditions Monitor of the LHCb Experiment: An Application Example 525

3 Cadmium Telluride and Cadmium Zinc Telluride as Sensor Materials . . . 529

4 New Passive Thermoluminescence Detectors 530

5 Conclusions 532

Acknowledgments 532

Suppliers of Equipment 532

References 532

C. Grupen, I. Buvat (eds.), *Handbook of Particle Detection and Imaging*,
DOI 10.1007/978-3-642-13271-1_21, © Springer-Verlag Berlin Heidelberg 2012

Abstract: Contemporary particle accelerators for fundamental research in particle physics like Fermilab's Tevatron and CERN's Large Hadron Collider (LHC) provide researchers with higher and higher luminosities. This sets the pace for the need for radiation-hard detector materials for both beamline instrumentation and the physics experiments themselves.

Silicon pixel and silicon microstrip detectors are well-developed devices for tracking applications in these high-energy physics experiments. However, these detectors are expected to reach the end of their lifetime within a few years due to their exposure to harsh radiation, of which the yearly level amounts to up to several 10^{14} hadrons/cm^2 during the foreseen 10 years of operation in the case of LHC experiments.

In order to protect sensitive experimental devices from adverse beam conditions, chemical vapor deposition (CVD) diamond, an artificially generated diamond material, is more and more being used in systems called Beam Condition Monitors (BCM). The radiation level these sensors are exposed to is even higher than in the case of position-sensitive tracking detectors. An example are the CVD diamond sensors of the BCM of the LHCb experiment at CERN, which is meant to withstand 10^{15} hadrons/cm^2 during 10 years.

Preparation of CVD diamond sensors for BCM applications is discussed in detail, together with the properties of this new material as a candidate for position-sensitive devices in high-energy physics experiments, addressing also operational questions like the appearance of erratic dark currents in polycrystalline diamond bulks. Other new materials for position-sensitive devices such as CdZnTe and CdTe are discussed as well and compared to the well-established silicon, together with a compilation of their properties relevant to particle detection.

Recent advances in the field of passive radiation monitors, where thermoluminescent sensors made from lithium fluoride now cover a dynamic range from several μGy up to 10^5 Gy, are also discussed briefly.

1 Radiation Environment at Contemporary Hadron Accelerators

Particle physics experiments installed at high-luminosity hadron accelerators such as the Large Hadron collider (LHC) at CERN, the European Organization for Nuclear Research, need to cope with possible adverse conditions of the hadron beams they are using, resulting in exposure of their sensitive detection equipment to high levels of ionizing radiation. In the case of the LHC, these are particularly hadronic showers from misaligned beams hitting structure material, or failures of components upon particle injection into the LHC from its pre-accelerator chain. In such a case, beam particles would interact with matter, causing a significant buildup in ionizing-radiation dose which potentially destroys sensitive detector components.

A prominent example is the Vertex Locator of the LHCb experiment installed at the LHC (Bates et al. 2006). This detector is one out of two movable devices along the LHC ring; it consists of 42 silicon detector modules which are being moved close to the LHC beams during operation in order to do *B* tagging, i.e., the identification of *b*- and *c*-hadron decays, on which the physics analysis of the LHCb experiment will focus.

That is why a system called Beam-Conditions Monitor (BCM) has been installed, consisting of sensors around the beam pipe, which is meant to trigger a beam abort from the LHC in case of adverse beam conditions. At the location of the innermost sensors of this BCM, which have a surface of 10 mm × 10 mm out of which 8 mm × 8 mm are active, already under normal operation of the LHC, the particle flux is expected to be 1 particle/(collision · cm^2) ≈ $5 \cdot 10^{15}$ particles/(10 years · cm^2).

2 Artificial Diamond as a Sensor Material

In order to cope with this flux in a way that the lifetime of the BCM extends at least over ten years, polycrystalline CVD diamond has been used as the sensor material for this application.

Diamond as a sensor material features the following properties:

- Low atomic number ($Z = 6$) can be considered a tissue-equivalent sensor material, thus low absorption of radiation
- Relative radiation hardness as compared to other sensor materials
- Wide band gap of 5.47 eV, so virtually no thermally generated noise
- High mobility of both electrons and holes
- Low capacitance
- Low leakage currents
- High thermal conductivity
- Operation at ambient or even higher temperatures, no need for cooling

2.1 Chemical Vapor Deposition (CVD) Diamond

Chemical vapor deposition (CVD) diamond detectors are more and more becoming considerable alternatives to silicon detectors due to their high radiation hardness. Nevertheless, their operation differs in various respects from that of the well-established silicon devices.

2.1.1 Production of Artificial Diamond

The production principle of CVD diamond in a plasma-enhanced process with hydrogen and methane as the reactants is as follows: The substrate surface in the CVD process is hydrogen-saturated while the surrounding gas contains methane (CH_4). The highly active hydrogen atoms are able to break the C–H bonds of the methane and create H–H bonds. The resulting free space is then filled by carbon atoms. Until 1962, it was expected that graphite is formed (Eversole et al. 1962), but in fact, a diamond structure builds up instead. During the chemical vapor deposition process, the diamond grows from a substrate of Pt(1 1 1) (Shintani 1996) or silicon oxide at a rate of several µm per hour. The CVD process is shown in ❯ *Fig. 1*, from which it can be concluded that the presence of hydrogen is essential.

During the carbon deposition process, grains of different sizes are formed: ❯ *Figure 2* shows a side view of a polycrystalline diamond as it grows in the chemical vapor deposition process.

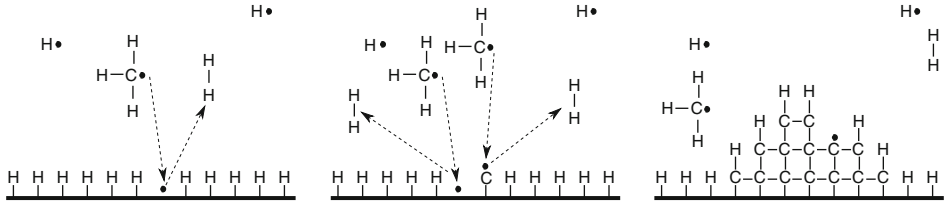


Fig. 1

The typical processes during chemical vapor deposition of carbon: The surface (Pt(1 1 1), for instance) is saturated with hydrogen, but reacts with hydrogen radicals from the gas phase (*left*). Carbon atoms are added at spots where no hydrogen is present anymore (*center*). The process continues, forming a carbon bulk in the form of the diamond lattice (*right*) (Sauerbrey 2009)

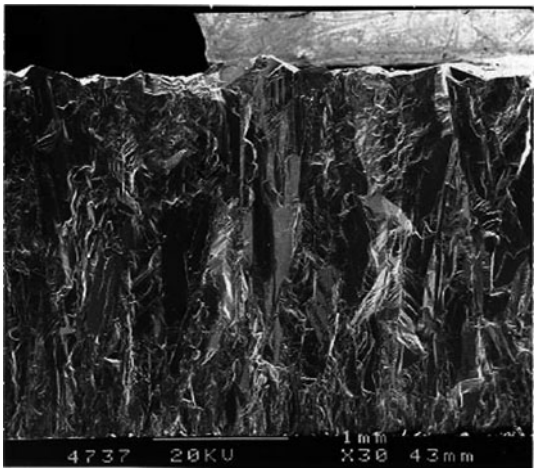


Fig. 2

A polycrystalline CVD diamond seen from the side. The structure originating from the growing process from bottom to top is clearly visible (Meier 1999)

The grains are smaller at the bottom, close to the substrate, and larger toward the top. The different grain sizes are considered to be due to impurities interfering with the growing process close to the substrate.

With respect to this, for applications in particle detection, single-crystal diamond is of course best suited. Grown on a diamond substrate, it is available already in surface sizes of about 3 cm × 3 cm, at significantly higher costs. The material commonly being used for beam monitors at particle accelerators is called polycrystalline. Here, the effective grain size is enlarged by mechanical removal of material from the substrate side. Typically, the grain size is then on the order of 1 µm or larger.

The grain size is crucial for the use of diamonds as particle detectors, since the detection principle is based on ionization of lattice atoms along the trajectory of a crossing particle. In the band model, electrons are shifted from the valence band into the conduction band. Then,

diamond, which by definition is an insulator due to its large band gap, can therefore be considered a semiconductor where it is possible to supply electrons to the conduction band through the interaction with ionizing particles.

2.2 Diamonds as Solid State Detectors

The basic working principle of particle detection with solid state detectors is the creation of electron–hole pairs in a sensitive volume with an external electric field applied, which is followed by an amplification stage. The intrinsic charge-carrier density inside the sensitive volume is low. If semiconductors like silicon or germanium are used as detectors, the sensitive volume is the depleted zone in an asymmetrically doped, reversely biased diode.

Also in the case of (CVD or natural) diamond, the intrinsic charge-carrier density inside the sensitive volume is low. The energy loss of charged particles passing through the detector is given by the Bethe–Bloch formula, based on which a restricted energy loss can be calculated, which takes energy loss by escaping secondary particles into account and represents the energy deposited in the bulk. This restricted energy loss generates free charge by producing electron–hole pairs, which then move to the contact electrodes. The collected charge is proportional to the energy deposition by ionization in the detector volume.

Due to the larger band gap (5.47 eV as compared to 1.11 eV for silicon and 0.67 eV for germanium, all measured at a temperature of 300 K), diamond is an insulator at room temperature, so it is depleted by itself, thus the detector noise is low and dark currents are negligible.

As a result of phononic excitation, the electron–hole pair-production energy is larger than the band gap, namely 13 eV, three to four times higher than for silicon. The average ionization density for a minimum-ionizing particle is 36 electron–hole pairs per μm only, so sensitive frontend electronics needs to be used for readout.

Instead, both an electron mobility of 2,200 $cm^2/(V\,s)$ and a hole mobility of 1,700 $cm^2/(V\,s)$ (as compared to 1,450 cm^2/Vs and 450 $cm^2/(V\,s)$ for silicon) make diamond detectors good candidates for fast detectors with response times on the order of a few nanoseconds.

When it comes to the question of radiation hardness, the high lattice displacement energy of 37–47 eV (Koike et al. 1992) as compared to 11 to 22 eV for silicon is expected to be an advantage of diamond over silicon and germanium.

2.3 Charge Collection in Polycrystalline CVD Diamonds

The lifetime of charge carriers in silicon or monocrystalline CVD detectors is theoretically unlimited. But any defect in the lattice of diamond helps electrons and holes to recombine or simply traps charge, limiting the lifetime of charge carriers and thus the charge collection efficiency of the sensor. The corresponding parameter is called charge collection distance, often referred to as CCD. This is somehow unfortunate, since the acronym CCD also stands for *charge-coupled device* and may thus be a source of confusion. Here, the symbol δ_q is being used, since the charge collection distance can be understood as the thickness of an ideal diamond, in which no charge is being trapped or recombination of electrons and holes takes place. δ_q is thus smaller or equal to the physical thickness d of the diamond. Both parameters are measured parallel to the lines of the electric field the sensor is biased with. For contemporary polycrystalline

CVD diamond sensors, δ_q used to have a value between 200 and 300 µm. The minimum energy necessary to produce one electron–hole pair is on the order of 13 eV (Meier 1999).

Oh describes the effect of charge-carrier lifetime by the following ansatz (Oh 1999) (also citing from (Schleich 2008)): An ionization process creates a free electron. After Ramo's theorem, its movement in the electrical field induces a current, which is

$$I = ev/d \tag{1}$$

with the electron charge e, the drift velocity v, and the distance between the metalization electrodes d. Accounting for the carrier lifetime τ, the total charge Q_m seen outside the detector is:

$$Q_m = \frac{e}{d}\int_0^\tau v\,\mathrm{d}t = \frac{e}{d}v\tau, \tag{2}$$

where the schubweg $\delta := v\tau = \mu E\tau$ can be expressed in terms of the charge-carrier mobility μ and the electric field strength E. In the following calculations, the statistical nature of electron capture and local variations of the schubweg have to be taken into account, and finally the following approximation can be made:

$$Q_m \approx Q_i\frac{\overline{\delta}}{d}, \tag{3}$$

where Q_i is the total charge created by ionization and $\overline{\delta} = \frac{1}{d}\int_0^d \delta(z)\,\mathrm{d}z$. z is counted along the axis parallel to the electric field. Accounting for both, electrons and holes, one obtains the expression (Oh 1999):

$$Q_m = Q_i\frac{\delta_q}{d} \tag{4}$$

with a parameter called the charge collection distance:

$$\delta_q = \delta_e + \delta_h = E(\mu_e\tau_e + \mu_h\tau_h). \tag{5}$$

It can be interpreted as an average separation distance of electron–hole pairs before they recombine. It is important to notice that the charge-carrier mobility depends on the electric field E and that $\mathrm{d}\delta_q/\mathrm{d}E \propto \frac{1}{E}$ with the electric field strength approaching saturation. Measurements (Fernandez-Hernando et al. 2005) with diamond samples similar to the ones used in LHCb have shown that $\delta_q = 140$ µm at 250 V. Applying a voltage four times as high increases δ_q only by 43%. In addition, the signal-to-noise ratio decreases considerably above 300 V bias voltage (Müller 2011). Due to the saturation of the charge collection distance above 250 V and erratic dark currents occurring especially at higher field strengths, a bias voltage value below that value was chosen for the Beam-Conditions Monitor of the LHCb experiment.

2.4 Radiation Effects

Radiation causes damage to the diamond bulk, resulting in a decrease of the charge-carrier lifetime, which reduces the charge collection distance. Exposure to protons of 24 GeV has shown that the charge collection distance remains constant up to a fluence of $3 \cdot 10^{15}/\mathrm{cm}^2$ (Adam et al. 2000). However, protons of 25 MeV only, instead, degraded the charge collection properties of polycrystalline CVD diamond significantly after exposure to fluences on the order of $10^{16}/\mathrm{cm}^2$.

The effect of several types of radiation on artificial diamond is summarized in *Table 1*.

Table 1
Performance of diamond sensors after irradiation

Radiation	Energy	Fluence or dose	Effect on charge collection distance	Reference
Protons	24 GeV	$3 \cdot 10^{15}$ cm^{-2}	No degradation	Adam et al. (2000)
Protons	25 MeV	$5.7 \cdot 10^{16}$ cm^{-2}	50% of initial value	Domke et al. (2008)
Neutrons	1 MeV	$2 \cdot 10^{15}$ cm^{-2}	70% of initial value	Adam et al. (2000)
Pions	300 MeV	$1.7 \cdot 10^{15}$ cm^{-2}	70% of initial value	Adam et al. (2000)
Alpha particles	5 MeV	$2 \cdot 10^{15}$ cm^{-2}	70% no degradation	Dulinski et al. (1994)
Photons	100 keV	6.8 MGy	No degradation	Dulinski et al. (1994)
Photons	1 MeV	10 MGy	No degradation	Dulinski et al. (1994)

Schleich (2008) explains the reversible effect of ionizing particles as being linked to the ionization density: This is nonuniform for hadrons, which leads to polarization effects diminishing the electric field and the charge collection distance of the detector. In contrast, electron and gamma radiation excite a constant ionization density leading to filling of traps. This effect increases the charge collection distance.

This is called *pumping*. Most polycrystalline CVD diamond detectors can only be used after having received a certain radiation dose to fill up the charge traps in their electronic band structure.

An adverse effect of radiation is the occurrence of erratic dark currents. These are current spikes lasting several hundred milliseconds, well exceeding nominal signals by several orders of magnitudes. As they only occur in polycrystalline CVD diamonds, grain boundaries seem to play a role in their generation. These erratic dark currents can be suppressed by a magnetic field of the order of 0.1–0.6 T, which is oriented perpendicularly to the grain boundaries or by lowering the bias voltage to 0.2 V per µm of sensor thickness (Edwards et al. 2005).

The radiation hardness of chemical vapor deposition (CVD) diamond has exhaustively been investigated by the RD42 collaboration at CERN (http://rd42.web.cern.ch/rd42) up to fluences of $3 \cdot 10^{15}$ protons per cm^2 (Kagan 2005).

2.4.1 The Beam-Conditions Monitor of the LHCb Experiment: An Application Example

For the Beam-Conditions Monitor of the LHCb experiment (The LHCb Collaboration 2008), polycrystalline CVD diamond sensors are being used. These sensors are arranged in the LHCb experimental cavern around the beam pipe of the Large Hadron Collider in a way that the sensitive area of the innermost sensor starts at a radial distance of 37.0 mm from the beam axis. From the mechanical dimensions and the relative permittivity of 5.7 for diamond, the capacitance of one sensor can be calculated to be 6.5 pF (Schleich 2008).

The charge collection distances of the LHCb BCM diamonds range from 192 to 240 µm with the exception of one diamond showing a charge collection distance of 132 µm (Schleich 2008). In order to estimate the particle flux through the sensors at LHCb, simulation results for a detector in close proximity are used (Lieng 2008): During 10 years of LHCb operation, $8 \cdot 10^{12}$ neutrons and $9.6 \cdot 10^{13}$ charged hadrons per cm^2 are expected, so no significant reduction of

the charge collection distance is expected to take place. The simulated minimum bias signal of the BCM detectors (Lieng 2008) under normal running conditions according to the simulated radiation field will then be between 5 and 20 nA.

Diamond sensors show a linear relationship between the flux of particles and the current signal over up to nine orders of magnitude, as shown in *Fig. 5*. This is essential for applications in beamline instrumentation at particle accelerators, where diamonds are more and more being used to protect other detectors sensitive to excess radiation. Here it is essential that the diamond sensor yields a measurable current signal under normal operating conditions, and a much higher signal when it comes to a malfunction of the accelerator. This sounds trivial at a first glance, but saturation effects upon excessive exposure to radiation that prevent the sensor from yielding sufficiently high a current onto which a beam abort trigger is based must be excluded. The linearity could be demonstrated by shining protons of 24 GeV onto a polycrystalline diamond sensor of the Beam-Conditions Monitor of the CMS experiment at CERN (Chong et al. 2007).

In order to validate the durability of metal contacts under irradiation, two different contact systems (titanium–silver and titanium–gold with a glued-on copper strip) were exposed to 10^{15} protons/mm^2 at an energy of 25 MeV over a period of 18 h. Although the diamond bulk suffered from radiation damage, it could be demonstrated that both contact systems maintained their ohmic properties and will thus depass the lifetime of the diamond sensor itself.

The CVD diamond sample used in the irradiation test was a prototype sensor for the Beam-Conditions Monitor of CERN's LHCb experiment (The LHCb Collaboration 2008), as described above and shown in *Fig. 3*. A metalized area of 8 mm × 8 mm covered the central part on both sides of the sensor. Upstream, with respect to the beam direction, the contact was made of a 50-nm gold layer with 50 nm of titanium as the undermetal. For the downstream contact, the gold layer was replaced by a silver layer of 50 nm, respectively.

In this design, the electrical contact is established with a glued-on copper-cladded HFS1 strip, using Epotecny E205 silver-doped conductive glue (see list of equipment providers at the end of this article). The advantage of this design is the large surface of the contact, in case readout at high frequencies is intended. A disadvantage is the danger of silver atoms diffusing into

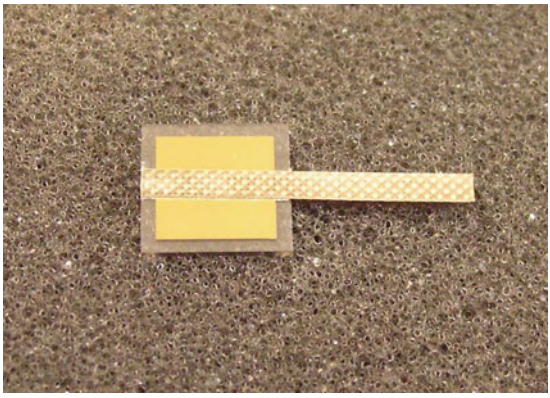


Fig. 3
A single polycrystalline CVD sensor of the Beam-Conditions Monitor of the LHCb experiment

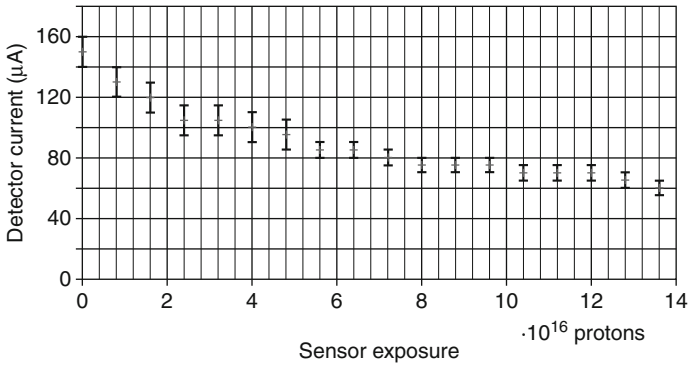


Fig. 4

Decrease of the sensor current during exposure to protons of 25 MeV over a surface of 2 mm^2 over 34 min

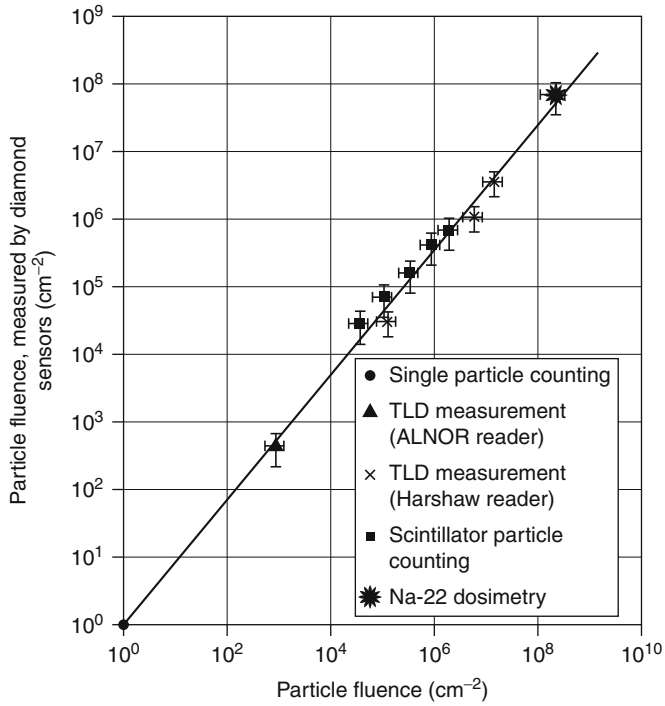


Fig. 5

Diamond sensor response to the a fluence of 24 GeV protons, as measured by a scintillator hodoscope of 5 mm × 5 mm cross section, by Na-22 dosimetry in aluminum, LiF thermoluminescence dosimetry, and by particle counting with scintillators (Chong et al. 2007)

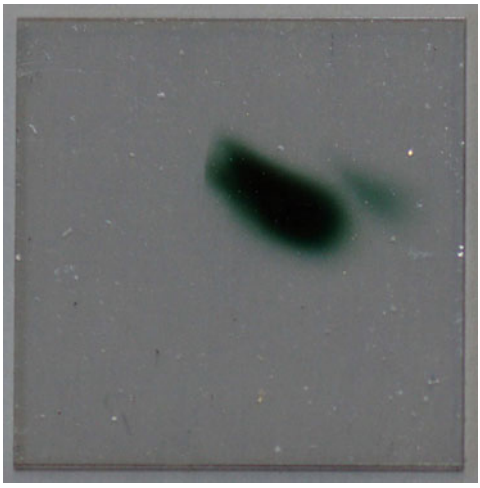


Fig. 6
Irradiation damage in polycrystalline CVD diamond

the diamond bulk. In order to prevent this from happening, the layer of the undermetal titanium was chosen to be significant (50 nm), i.e., equal to the thickness of the gold contact itself. Of course, bonding techniques are also an option. Initially, the resistivity of this glued contact system was measured to be below 1 Ω.

After an exposure to 10^{15} protons/mm^2 at an energy of 25 MeV shined on a surface of 4 mm^2, both contacts have maintained their ohmic properties. Despite the fact that the metallic contact was still operational after exposure, showing no measurable degradation, the current signal from the diamond had completely vanished. Over 34 min, this effect was further studied at a previously unexposed spot of the diamond sensor. The current decrease is shown in *Fig. 4*.

The conclusion that can be drawn is that an electrical contact as described above shows, in terms of radiation damage, a lifetime well depassing that of the artificial diamond material itself.

Sauerbrey (2009) applied a number of analysis methods to the irradiated sensor. After removal of the metalization as shown in *Fig. 6*, he could confirm that no new crystalline structures inside the diamond had formed.

By scanning electron microscopy, it was shown that also no change in grain size or other visible damage had taken place. Also, by energy-dispersive X-ray microanalysis (EDX), it could be shown that no gold or silver had diffused into the diamond sensor.

By absorption spectroscopy, the presence of nitrogen in the bulk could be determined, which had formed green color centers. These damages have spoilt the charge collection capacity of the sensor.

For completeness, in *Fig. 7*, the vanishing of the damage is shown under heating. Curing radiation-damaged sensors in real applications is certainly not an option, as the metalization would suffer from the heat, at least diffusion of atoms from the metalization into the diamond bulk would take place, deteriorating its charge collection properties.

In *Table 1*, the influence of radiation of various types on diamond sensors is given.

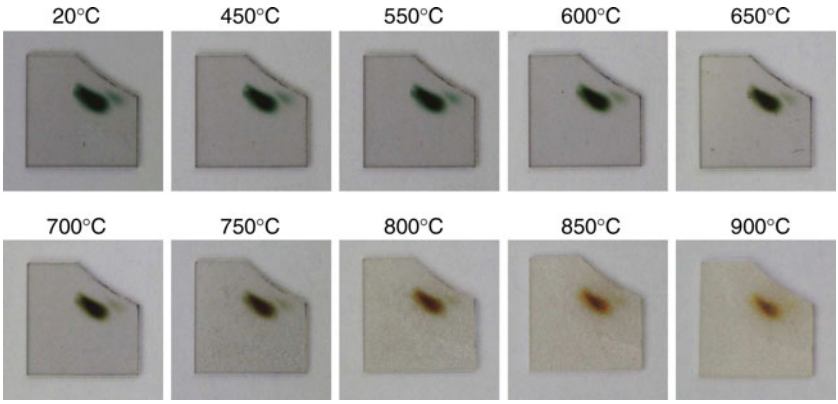


Fig. 7
Healing of a radiation-damaged diamond sensor by heating

Recently, a new way of metalizing diamond sensors using a diamond-like carbon (DLC) layer has been proposed. This way of applying a metallic contact to the diamond sensors promises to transfer higher current densities. First tests with this new metalization are very promising (Galbiati et al. 2009).

3 Cadmium Telluride and Cadmium Zinc Telluride as Sensor Materials

Cadmium telluride (CdTe), a cadmium and tellurium compound material, is widely used as a solar cell material in photovoltaics. For this application, a *pn* junction is created between CdTe and layers of CdS (cadmium sulfide).

With the highest (linear) electro-optic coefficient among all known II-VI compound materials ($r_{41}r_{52}r_{63}$), CdTe is also used as an electro-optical modulator.

For detector applications, CdTe is used in the form of various alloys, such as HgCdTe, which is sensitive to infrared radiation. Doped with chlorine, CdTe is in use as a radiation detector also sensitive to electrons and alpha particles. But of particular importance for radiation measurements is its alloy with zinc.

Cadmium zinc telluride (CdZnTe or CZT) is an alloy of cadmium telluride and zinc telluride. It is a room-temperature direct-bandgap semiconductor directly sensitive to X-rays and gamma rays. CZT is the outcome of intensive research with the goal of finding a sensor material with better detection properties than silicon and germanium: CZT can be operated at room temperature and saturates only beyond a photon flux of $10^6/(mm^2 \cdot s)$. Also, the energy resolution of CZT, important for spectroscopic applications, is better than that of scintillation detectors. Combining good energy resolution and high count-rate capability in a detector operating at room temperature makes CZT an interesting material for applications in fundamental research, medicine, industry, and homeland security.

The properties of most of the discussed sensor materials including diamond are summarized in *Table 2*.

Table 2

Properties of detector materials. Zincblende (ZnS) has a cubic crystal system, its lattice type is face centered. The diamond structure is fairly similar, but obviously consists of carbon atoms only. (Based on data published by Sze (1981), Dulinski et al. (1994), Schieber et al. (1996), Lindner (1997), Nuclear Institute of Standards and Technology (1999), collected by Klaiber-Lodewigs (1999))

Material	Si	Ge	GaAs	CdTe (CdZnTe)	Diamond (C)
Lattice structure	Diamond	Diamond	Zincblende	Zincblende	Diamond
Bandgap type	Indirect	Indirect	Direct	Direct	Indirect
Atomic number – Z	14	32	32	50	6
Atomic mass – A [atomic mass units]	28.09	72.61	72.32	120.0	12.01
Density – ρ [g/cm^3]	2.33	5.32	5.31	6.20	3.50
Ionization potential V_i [eV]	173.0	350.0	384.9	539.3	≈ 78
Bandgap width E_{gap} [eV]	1.12	0.66	1.42	1.56	5.47
Medium electron–hole production energy – E_{eh} [eV]	3.6	2.9	4.2	4.7	≈ 13
Electron mobility μ_e [cm^2/(V s)]	1,500	3,900	8,500	1,050	1,800
Hole mobility μ_h [cm^2/(V s)]	450	1900	400	100	1200
Specific resistance R_s [Ω cm]	$2.3 \cdot 10^5$	47	10^7–10^8	10^9–10^{11}	$> 10^{11}$
Differential energy loss for minimum-ionizing particles [eV/μm]	358.0	667.8	661.7	698.5	585.5
Energy-dependent intrinsic energy resolution $\Delta E/E$ [(eV/E)$^{1/2}$]	1.55	1.39	1.67	1.77	2.94

4 New Passive Thermoluminescence Detectors

Also in the field of passive radiation sensors, new developments have taken place. Here, the term "*passive radiation sensor*" is used to describe a system that changes certain properties under the influence of ionizing radiation. After exposure, these properties are measured. This way, the passive sensor is read out and the integrated value of the energy dose the sensor was exposed to can be determined. Especially for thermoluminescence detectors, advancements could be achieved.

Thermoluminescence is a form of luminescence shown by certain crystalline materials, such as lithium fluoride. Energy previously absorbed from exposure of the crystal to electromagnetic radiation or other ionizing radiation is re-emitted as light when the material is heated up to several 100 °C. The way light is emitted during this heating process is described by the glow curve of the material.

Radiation creates excited electronic states in crystalline materials. In some materials, these states are trapped by lattice defects. Although stable in time, energetically, these states are not stable. Heating the material enables the trapped states to interact with lattice vibrations (phonons), rapidly decaying into lower-energy states. This process leads to the emission of photons.

Thermoluminescent (TL) dosimeters made from lithium fluoride are routinely used to monitor absorbed doses in many kinds of radiation fields which contain photons, electrons, and neutrons.

Lithium-fluoride detectors doped with manganese, copper, and phosphorus (LiF:Mg,Cu,P), usually referred to as MCP detectors, show a very high sensitivity and a simple signal-to-dose relation. They can be considered a de facto standard in modern environmental thermoluminescence dosimetry. According to Obryk et al. (2011), these sensors are capable of measuring doses at μGy levels and even below.

The shape of the glow curve resulting from doses ranging from a μGy to a kGy is practically identical (Obryk et al. 2011). The glow curve represents the light emission of the sensor when it is heated up at a given rate. However, significant changes of their glow-curve shape at high and very high doses have been discovered (Bilski 2002; Olko et al. 2010). High-temperature peaks start to grow at doses above 1 kGy and continue to grow up to doses of about 50 kGy when a completely new peak appears in the MCP's glow curve beyond 400 °C.

Using these properties, one single sensor covers a dynamic range of nine orders of magnitude. A dynamic range covering five orders of magnitude can be seen in *Fig. 8*. This feature opens up applications of thermoluminescence dosimeters also in the field of dosimetry at contemporary particle accelerators in fundamental research, where the expected dose cannot easily be determined beforehand.

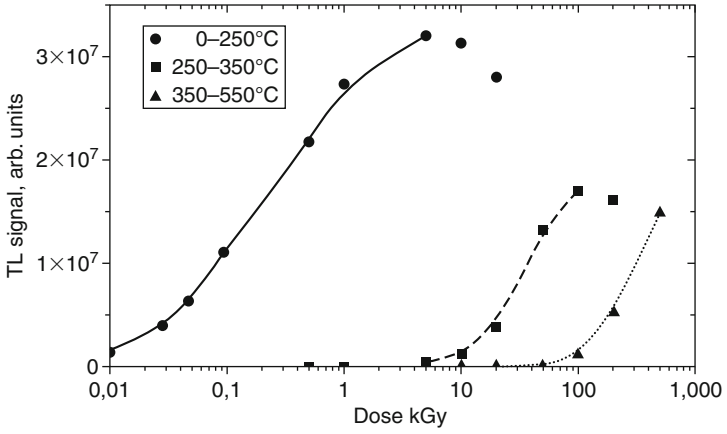


Fig. 8
Dynamic range of MCP TLD (LiF:Mg,Cu,P) (Obryk et al. 2011)

5 Conclusions

Coping with the radiation environment at modern high-luminosity particle accelerators, such as the Large Hadron Collider at CERN or Tevatron at Fermilab, represents a major challenge. Accelerators need to be equipped with radiation-hard sensors to provide input to their control systems. Chemical vapor deposition diamond can already be considered an established sensor material for these applications. Activities are going on to develop this material further, in order to use it also for position-sensitive devices in high-energy physics experiments, the way the less radiation-hard silicon is used today.

When energy resolution at relatively high particle fluxes is needed, cadmium zinc telluride can be used for sensors which can be operated at room temperature.

Also in the field of passive sensors for longer-term radiation monitoring, for instance in experimental caverns, developments have taken place. New thermoluminescence detectors made from lithium fluoride doped with manganese, copper, and phosphorus (LiF:Mg,Cu,P) are very sensitive, offer a simple signal-to-dose relation, but nevertheless cover a dynamic range of nine orders of magnitude.

Acknowledgments

I would like to thank Harris Kagan, Dirk Meier, Alexander Oh, Shaun Roe, and Peter Weilhammer for valuable discussions on the subject of diamond detectors. In the same way, I owe thanks to Pawel Bilski, Maciej Budzanowski, Barbara Obryk, and Pawel Olko, when it comes to the discussion of thermoluminescence detectors. Jan Sauerbrey was so kind to provide material that originates from the work he carried out for his diploma thesis.

The detector design referred to in this article as an example is based on concepts developed when the author was with Technische Universität Dortmund, Experimentelle Physik 5, 44221 Dortmund, Germany.

Suppliers of Equipment

CVD diamond sensors, also metalized ones: Diamond Detectors Ltd., United Kingdom, www.diamonddetectors.com

Conductive glue: Epotecny, France, www.epotecny.com/uk/; www.bicron.com/

Thermoluminescence detecors: Thermo Scientific, Germany, www.thermoscientific.com/; Institut Fizyki Jadrowej, Poland, www.ifj.edu.pl/

Metalized polymers as sensor contacts: ISTechnologie, Germany, www.istechnologie.de/; MiCryon Technik, Germany, www.micryon.de/

References

Adam W et al (2000) Pulse height distribution and radiation tolerance of CVD diamond detectors. Nucl Instrum Methods Phys Res A 447: 244–250

Bates AG, Borel J, Buytaert J, Collins P, Eckstein D, Eklund L (2006) IEEE Trans Nucl Sc 53 (Part 3):3

Bilski P (2002) Lithium fluoride: from LiF:Mg,Ti to LiF:Mg,Cu,P, Radiat Prot Dosim 100(1–4): 199–206

Chong D, Fernandez-Hernando L, Gray R, Ilgner CJ, Macpherson AL, Oh A, Pritchard TW, Stone R, Worm S (2007) Validation of synthetic diamond

for a beam condition monitor for the compact muon solenoid experiment. IEEE Trans Nucl Sci 54(1):182–185

Domke M, Gernhäuser R, Ilgner C, Schwertel S, Warda K (2008) Validation of titanium-gold and titanium-silver and copper contacts on CVD diamond sensors for beam-conditions monitors and tracking detectors for heavy ions under proton irradiation. Maier-Leibnitz-Laboratorium der Universität München und der Technischen Universität München

Dulinski W et al (1994) Diamond detectors for future particle physics experiments. In: 27th international conference on high-energy physics, Glasgow, CERN-PPE-94-222

Edwards AJ, Bruinsma M, Burchat P, Kagan H, Kass R, Kirkby D et al (2005) Radiation monitoring with CVD diamonds in BABAR. Nucl Instrum Methods Phys Res A 552: 176–182

Eversole WG (1962) Synthesis of diamond, US Patent, o. 3,030,187

Fernandez-Hernando L, Chong D, Gray R, Ilgner C, Macpherson A, Oh A, Pritchard T, Stone R, Worm S (2005) Development of a CVD diamond beam condition monitor for CMS at the Large Hadron Collider. Nucl Instrum Methods Phys Res A 552:183–188

Galbiati A, Lynn S, Oliver K, Schirru F, Nowak T, Marczewska B et al (2009) Performance of monocrystalline diamond radiation detectors fabricated using TiW, Cr/Au and a novel Ohmic DLC/Pt/Au electrical contact. IEEE Trans Nucl Sci 56(4):1863–1874

Kagan H (2005) Recent advances in diamond detector development. Nucl Instrum Methods Phys Res A 541:221–227

Klaiber-Lodewigs JM, Eigenschaften und Einsatz von CdTe/CdZnTe-Mikrostreifendetektoren, diploma thesis, Bonn University 1999, BONN-IB- 99-08

Koike J, Parkin DM, Mitchell TE (1992) Displacement threshold energy for type IIa diamond. Appl Phys Lett 60(12):1450–1452

Lieng M (2008) Summary of simulations for the beam conditions monitor at the LHCb. Technical report LHCb-2008-027, Technische Universität Dortmund

Lindner M (1997) Einsatz von GaAs-Mikrostreifendetektoren im Bioscope-System. Diploma thesis, Bonn University, BONN-IB-97-22

Meier D (1999) CVD diamond sensors for particle detection and tracking. PhD thesis, Ruprecht-Karls-Universität Heidelberg

Müller S (2011) The beam condition monitor 2 and the radiation environment of the CMS detector at the LHC. PhD thesis, Karlsruhe Institute of Technology

Nuclear Institute of Standards and Technology, USA, 1999

Obryk B, Bilski P, Olko P (2011) Method of thermoluminescent measurement of radiation doses from micrograys up to a megagray with a single LiF:Mg,Cu,P detector. Radiat Prot Dosimetry 144(1–4):543–547, doi: 10/1093/rpd/ncq339

Oh A (1999) Particle detection with CVD diamond. PhD thesis, University of Hamburg

Olko P, Bilski P, El-Faramawy NA, Gōksu HY, Kim JL, Kopec R, Waligórski MPR (2010) On the relationship between dose-, energy- and LET-response of thermoluminescent detectors. Radiat Prot Dosim 119:15–22

Sauerbrey J (2009) Upgrade evaluation of the LHCb beam conditions monitor and pCVD diamond sensor irradiation analysis. Diploma thesis, Technical University of Dortmund

Schieber M et al (1996) Material properties and room-temperature nuclear detector response of wide bandgap semiconductors, NIM A 377: 492–495

Schleich S (2008) FPGA based data acquisition and beam dump decision system for the LHCb beam conditions monitor. Diploma thesis, Technical University of Dortmund

Shintani Y (1996) J Mater Res 11:29–55

Sze SM (1981) Physics of semiconductor devices, 2nd edn. Wiley New York

The LHCb Collaboration (2008) JINST S08005, 3. doi: 10.1088/1748-0221/3/08/S08005

The web site of the RD42 collaboration at CERN is http://rd42.web.cern.ch/rd42. Accessed 4 March 2011

22 Radiation Damage Effects

R.-Y. Zhu
Physics, Mathematics and Astronomy Division, California Institute of Technology, Pasadena, CA, USA

1 Introduction ***536***

2 Scintillation-Mechanism Damage ***538***

3 Radiation-Induced Phosphorescence and Energy-Equivalent Readout Noise ***539***

4 Radiation-Induced Absorption ***539***
4.1 Recovery of Radiation-Induced Absorption 542
4.2 Radiation-Induced Color Centers 543
4.3 Dose-Rate Dependence and Color-Center Kinetics 546

5 Light-Output Degradation ***546***

6 Light-Response Uniformity ***547***

7 Damage Mechanism in Alkali Halide Crystals and CsI(Tl) Development ***549***

8 Damage Mechanism in Oxide Crystals and PWO Development ***551***

9 Conclusion ***553***

Acknowledgments ***554***

References ***554***

Further Reading ***555***

C. Grupen, I. Buvat (eds.), *Handbook of Particle Detection and Imaging*,
DOI 10.1007/978-3-642-13271-1_22, © Springer-Verlag Berlin Heidelberg 2012

Abstract: Radiation damage is an important issue for the particle detectors operated in a hostile environment where radiations from various sources are expected. This is particularly important for high energy physics detectors designed for the energy and intensity frontiers. This chapter describes the radiation damage effects in scintillating crystals, including the scintillation-mechanism damage, the radiation-induced phosphorescence, and the radiation-induced absorption. The radiation damage mechanism in crystal scintillators is also discussed. While the damage in halides is attributed to the oxygen/hydroxyl contamination, it is the structure defects, such as the oxygen vacancies, which cause the damage in oxides. Various material analysis methods used in investigations of the radiation damage effects as well as the improvement of crystal quality through systematic R&D are also presented.

1 Introduction

Total-absorption shower counters made of inorganic crystal scintillators have been known for decades for their superb energy resolutions and detection efficiencies (Gratta et al. 1994). In high energy and nuclear physics experiments, large arrays of scintillating crystals of up to 10 m^3 have been assembled for precision measurement of photons and electrons. These crystals are working in a radiation environment, where various particles, such as γ rays, neutrons, and charged hadrons, are expected. *Table 1* (Mao et al. 2008) lists the basic properties of the heavy-crystal scintillators commonly used in high energy physics detectors. They are NaI(Tl), CsI(Tl), undoped CsI, BaF_2, bismuth germanate (BGO), lead tungstate (PWO), and Ce-doped lutetium oxyorthosilicate ($Lu_2(SiO_4)O$ or LSO(Ce)) (Melcher and Schweitzer 1992). All have either been used in, or actively pursued for, high energy and nuclear physics experiments. Some of them, such as NaI(Tl), CsI(Tl), BGO, LSO(Ce), and cerium-doped lutetium–yttrium oxyorthosilicate ($Lu_{2(1-x)}Y_{2x}SiO_5$:Ce, LYSO) (Cooke et al. 2000; Kimble et al. 2002) are also widely used in the medical industry.

All known crystal scintillators suffer from radiation damage (Zhu 1998). There are three possible radiation damage effects in crystal scintillators: (1) the scintillation-mechanism damage, (2) the radiation-induced phosphorescence (afterglow), and (3) the radiation-induced absorption (color centers). A damaged scintillation mechanism would reduce the scintillation light yield and cause a degradation of the light output. It may also change the light-response uniformity along the crystal length since the radiation dose profile is usually not uniform. The radiation-induced phosphorescence, commonly called afterglow, causes an increase of the dark current in the photodetectors, and thus an increase of the readout noise. The radiation-induced absorption reduces the light attenuation length (Ma and Zhu 1993), and thus the light output and possibly also the light-response uniformity.

Table 2 summarizes γ-ray-induced radiation damage effect for various crystal scintillators. There is no experimental data supporting a scintillation-mechanism damage. All crystal scintillators studies so far, however, suffer from the radiation-induced phosphorescence and the radiation-induced absorption.

The radiation-induced absorption is caused by a process called color-center formation, which may recover spontaneously under the application temperature through a process called color-center annihilation. If so, the damage would be dose-rate dependent (Ma and Zhu 1993, 1995; Zhu 1997). If the radiation-induced absorption does not recover, or the recovery speed is

Table 1

Properties of some heavy-crystal scintillators

Crystal	NaI(Tl)	CsI(Tl)	CsI	BaF_2	BGO	PWO	LSO(Ce)
Density (g/cm^3)	3.67	4.51	4.51	4.89	7.13	8.3	7.40
Melting point (°C)	651	621	621	1280	1050	1123	2050
Radiation length (cm)	2.59	1.86	1.86	2.03	1.12	0.89	1.14
Molière radius (cm)	4.13	3.57	3.57	3.10	2.23	2.00	2.07
Interaction length (cm)	42.9	39.3	39.3	30.7	22.7	20.7	20.9
Refractive index[a]	1.85	1.79	1.95	1.50	2.15	2.20	1.82
Hygroscopicity	Yes	Slight	Slight	No	No	No	No
Luminescence[b] (nm)	410	560	420	300	480	425	420
(at Peak)			310	220		420	
Decay time[b] (ns)	245	1220	30	650	300	30	40
			6	0.9		10	
Light yield[b,c]	100	165	3.6	36	21	0.30	85
			1.1	4.1		0.077	
$d(LY)/dT^{b,d}$ (%/°C)	−0.2	0.4	−1.4	−1.9	−0.9	−2.5	−0.2
				0.1			
Experiment	Crystal	CLEO	KTeV	TAPS	L3	CMS	KLOE
	Ball	BaBar			BELLE	ALICE	SuperB
		BELLE				PrimEx	
		BES III				Panda	

[a] At the wavelength of the emission maximum

[b] *Top line*: slow component, *bottom line*: fast component

[c] Relative light yield of samples of 1.5 X_0 and with the PMT QE taken out

[d] At room temperature

Table 2

Radiation damage in crystal scintillators

Item	CsI(Tl)	CsI	BaF_2	BGO	PWO	LSO/LYSO
Scintillation mechanism	No	No	No	No	No	No
Phosphorescence (afterglow)	Yes	Yes	Yes	Yes	Yes	Yes
Absorption (color centers)	Yes	Yes	Yes	Yes	Yes	Yes
Recover at room temperature	Slow	Slow	No	Yes	Yes	No
Dose-rate dependence	No	No	No	Yes	Yes	No
Thermally annealing	No	No	Yes	Yes	Yes	Yes
Optical bleaching	No	No	Yes	Yes	Yes	Yes

very low, the color-center density would increase continuously under irradiations until all defect traps are fully filled. In this case, the corresponding radiation damage effect is not dose-rate dependent.

Color centers may also be annihilated thermally by heating the crystal to a high temperature through a process called thermal annealing, or optically by injecting light of various wavelengths to the crystal through a process called optical bleaching (Ma and Zhu 1993, 1995). The recovery process, either spontaneous or manual through thermal annealing or optical bleaching,

reduces the color-center density or the radiation-induced absorption. At the same time, it also introduces an additional instability for the crystal's light output because of the variation of the crystal's transparency. In this case, a precision monitoring system is mandatory to follow the variations of the crystal's transparency.

The radiation damage caused by neutrons and charged hadrons may differ from that caused by γ rays. Studies (Huhtinen et al. 2005, 2006, 2008) on proton-induced radiation damages in PWO crystals, for example, show a very slow (or no) recovery at room temperature, contrary to the radiation damage caused by γ rays. This leads to a cumulative damage in PWO with no dose-rate dependence for hadrons.

The radiation damage level is also different at different temperatures for crystals with dose-rate-dependent damage since the spontaneous recovery speed is temperature dependent. PWO crystals used at low temperature, for example, suffer more damage than that at high temperature (Semenov et al. 2007, 2008, 2009).

Commercially available mass-produced crystals usually do not meet the quality required for high energy physics detectors. The quality of mass-produced crystals, however, may be improved by removing harmful impurities and defects in the crystal.

The rest of this chapter discusses γ-ray-induced radiation damage phenomena in scintillating crystals, the origin of the radiation damage in halides and oxides, as well as the improvement of crystal quality through systematic R&D. All data presented in this chapter are measured for full-size crystals adequate for calorimeter construction, which is typically 18–25 X_0 long. Since both the radiation-induced phosphorescence and absorption are a bulk effect, it is important that only the full-size crystals are used in such studies.

2 Scintillation-Mechanism Damage

Experimental facts show that the crystal's scintillation mechanism is not damaged. This is observed for irradiations of γ rays, neutrons, as well as charged hadrons (Batarin et al. 2003; Huhtinen et al. 2005, 2006, 2008; Batarin et al. 2004, 2005). A common approach is to compare the shape of the emission spectra measured before and after irradiations. Direct comparison of the overall intensity of the emission spectra suffers from a large systematic uncertainty caused by the sample position and orientation, the surface quality, and the internal absorption which may be induced by the radiation.

The top plots of ❯ *Fig. 1* show the photoluminescence spectra measured before (blue) and after (red) γ-ray irradiations for a PWO sample (left) and an LYSO sample (right). These spectra are normalized to the integration around the emission peaks as shown in the figure. The relative difference between these normalized spectra (green) is shown in the bottom plots. Also shown in the bottom plots are the averages of the absolute values of the relative difference. The numerical values are 0.7% and 0.6%, respectively, for PWO and LYSO, which are much less than the systematic uncertainty of these measurements, indicating that no statistically significant difference is observed between the photoluminescence spectra taken before and after irradiations for both PWO and LYSO. This observation consists with no damage to the scintillation mechanism. Similar studies show that there is no scintillation-mechanism damage observed for BGO (Wei et al. 1990; Zhu et al. 1991), BaF_2 (Zhu 1994), and CsI(Tl) (Zhu et al. 1996) as well. This conclusion is also supported by more complicated measurements of the light-response uniformity before and after irradiations with a nonuniform dose profile (Batarin et al. 2003, 2004, 2005).

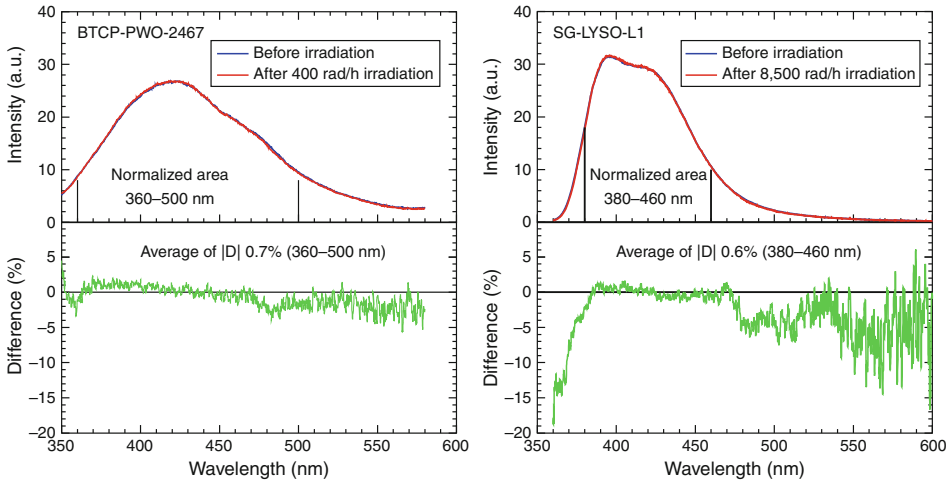


Fig. 1

Normalized photoluminescence spectra measured before (*blue*) and after (*red*) γ-ray irradiation and corresponding difference (*green*) are shown as a function of wavelength for a PWO sample (*left*) and an LYSO sample (*right*)

3 Radiation-Induced Phosphorescence and Energy-Equivalent Readout Noise

The radiation-induced phosphorescence can be measured as the residual photocurrent after the radiation is turned off. The left plot of *Fig. 2* shows the γ-ray-induced photocurrent, normalized to that during the irradiations, as a function of time during and after the γ-ray irradiations for several crystal samples: PWO, BGO, and LSO/LYSO. All samples are of full size adequate for calorimeter applications. The amplitude of the normalized phosphorescence is at a level of 10^{-5} for BGO and PWO, 3×10^{-4} for LYSO, and 2×10^{-3} for LSO. The LYSO samples are also observed as having a smaller phosphorescence than the LSO sample.

The right plot of *Fig. 2* shows γ-ray-induced anode photocurrents as a function of the γ-ray dose rate applied to several LSO and LYSO samples. Consistent slopes are observed for all samples indicating similar light yield for these samples. The slope may be used to calculate the readout noise in the number of electrons for a certain integration gate and be converted to the energy-equivalent readout noise by normalizing to the crystal's light output (Mao et al. 2009a, b). Because of its high light yield (200 times PWO and 5 times BGO) and short decay time (40 ns), the energy-equivalent readout noise in LSO and LYSO is an order of magnitude lower than that in PWO for both γ-ray and neutron irradiations.

4 Radiation-Induced Absorption

The main consequence of radiation damage in scintillation crystals is the radiation-induced absorption or color-center formation. Depending on the type of the defects in the crystal, the color centers may be electrons located in the anion vacancies (F center) and holes located in the

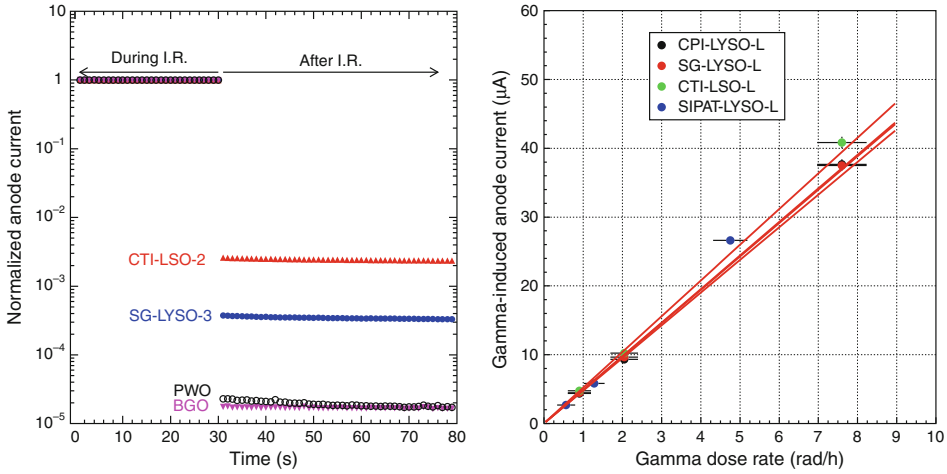


Fig. 2

***Left*: Normalized anode current is shown as a function of time during and after γ-ray irradiations for the BGO, PWO, LSO, and LYSO samples. *Right*: γ-ray-induced anode photocurrent is shown as a function of the dose rate applied to several LSO and LYSO samples**

cation vacancies (V center), as well as interstitial anion atoms (H center) or ions (I center), etc. Radiation-induced absorption is observed by comparing the longitudinal optical transmittance spectra measured after and before the irradiations.

Figure 3 shows the longitudinal transmittance spectra as a function of wavelength measured before and after several steps of irradiations for four halide crystals: CsI(Tl) (top left) and BaF_2 (top right) and two oxide crystals: PWO (bottom left) and LYSO (bottom right). While the color-center width is narrow in CsI(Tl), it is relatively wide in other crystals. It is interesting to note that the CsI(Tl) sample SIC-5 suffers much less radiation damage than other two CsI(Tl) samples since it was grown with a scavenger in the melt to remove the oxygen contamination, which is an effective approach to improve radiation hardness for the halide crystals as discussed in Sect. 7. For the BaF_2 sample, we also notice that the fast dose rate (top) is up to a factor of 30 of the slow rate (bottom) while the damage levels of the longitudinal transmittance are identical for the same integrated dose. This is expected since no recovery at the room temperature was observed for BaF_2 as discussed in Sect. 4.3.

It is also interesting to note that the radiation-induced absorption is much smaller in LSO and LYSO than that in other crystals. *Figure 4* shows an expanded view of the longitudinal transmittance spectra measured before and after several steps of γ-ray irradiations for a PWO (left) and an LYSO (right) sample. Also shown in the figure is the corresponding photoluminescence spectra (blue) and the numerical values of the photoluminescence-weighted longitudinal transmittance (EWLT), which is defined as:

$$\mathrm{EWLT} = \frac{\int LT(\lambda)\, Em(\lambda)\, \mathrm{d}\lambda}{\int Em(\lambda)\, \mathrm{d}\lambda}. \tag{1}$$

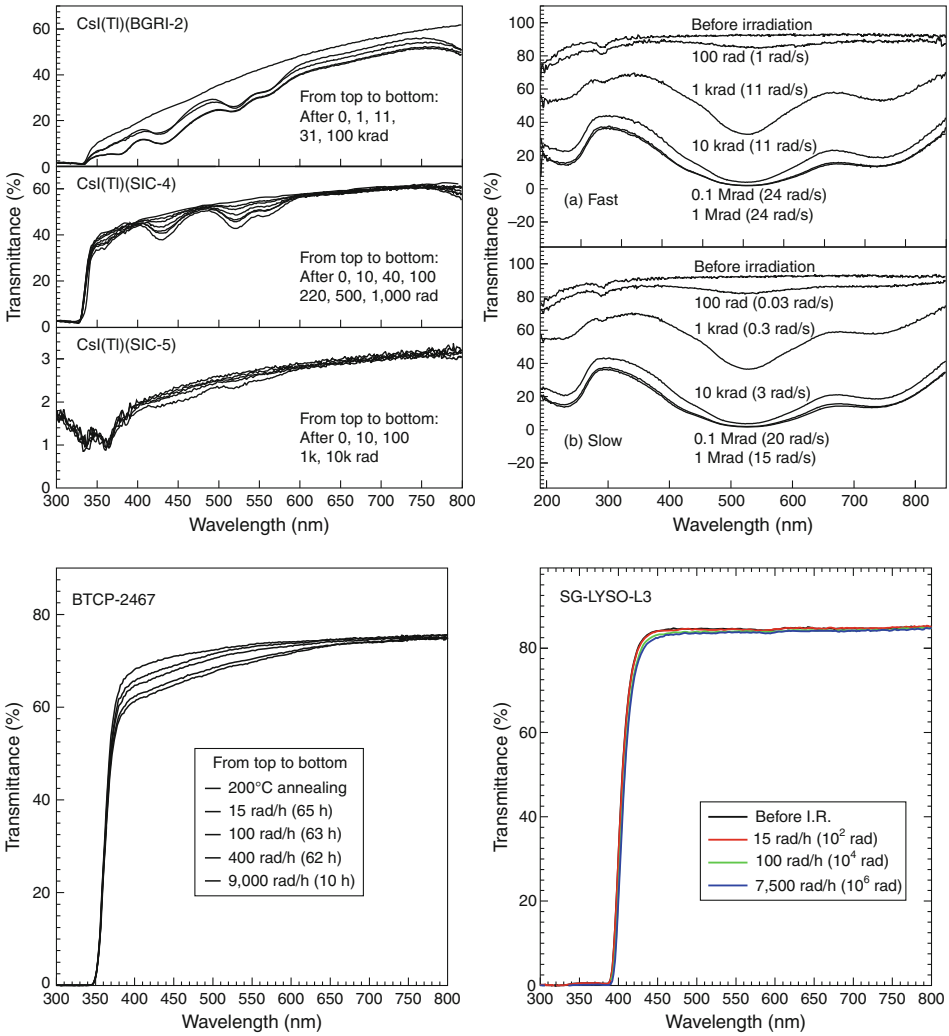


Fig. 3

The longitudinal transmittance spectra measured before and after several steps of the irradiation are shown as a function of wavelength for several CsI(Tl) (*top left*), BaF_2 (*top right*), PWO (*bottom left*), and LYSO (*bottom right*) samples

The EWLT values represent the crystal's transparency more accurately than the transmittance at the emission peak, which is commonly used in various radiation damage studies. This is particularly true for LSO and LYSO, which have a non-negligible self-absorption since their emission spectra are not entirely within the transparent region of the crystal (Chen et al. 2005, 2007).

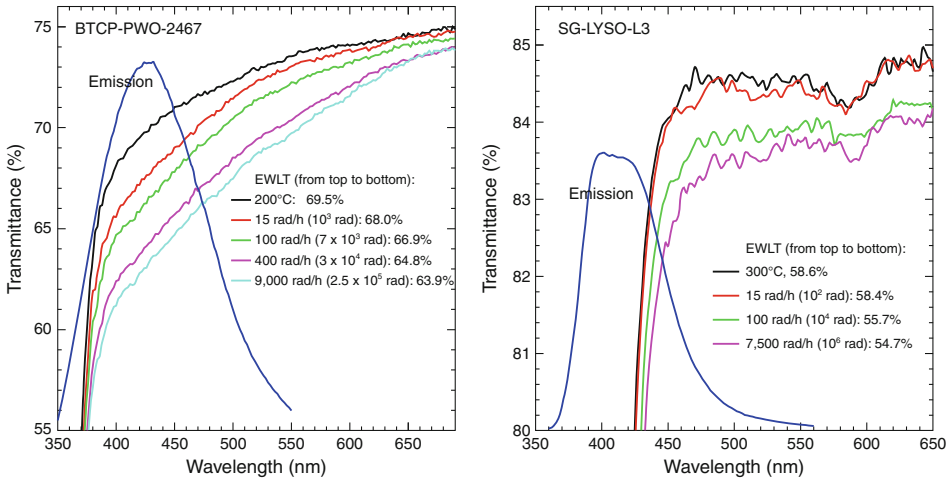


Fig. 4
Degradation on EWLT for a PWO sample (*left*) and an LYSO sample (*right*)

4.1 Recovery of Radiation-Induced Absorption

Depending on the color-center depth, the radiation-induced absorption may recover spontaneously at the room or application temperature. *Figure 5* shows the recovery behavior of the longitudinal optical transmittance measured after the γ-ray irradiations up to 4,000 and 500 h, respectively, for two PWO samples at 440 nm (left) and four LSO and LYSO samples at 420 nm (right). Three recovery time constants were determined by using exponential fits for these PWO samples. While the short time constant is at a few tens of hours, the medium time constant is at a few thousand hours, and the third time constant is much longer, which may be considered no recovery for the time scale of these measurements. It is also interesting to note that the LSO and LYSO samples show very slow recovery speed, which is consistent with no recovery. Similarly, the radiation-induced absorption does not recover at the room temperature for BaF_2 (Zhu 1994) and CsI(Tl) (Zhu et al. 1996) as well.

In addition to the spontaneous recovery at the room or application temperature, the radiation damage level may also be reduced by heating crystals to a high temperature (thermal annealing) or injecting light of various wavelengths to the crystal (optical bleaching). The γ-ray-induced absorption can be thermally annealed entirely at 200°C for BaF_2 (Zhu 1994), BGO (Wei et al. 1990; Zhu et al. 1991), and PWO (Zhu et al. 1996, 1998, 1999, 2002, 2004), or 300°C for LSO and LYSO (Chen et al. 2005, 2007). Optical bleaching was also found effective for BaF_2 (Zhu 1994), BGO (Wei et al. 1990; Zhu et al. 1991), and PWO (Zhu et al. 1996, 1998, 1999, 2002, 2004). On the other hand, the γ-ray-induced absorption in CsI(Tl) can neither be annealed thermally or bleached optically (Zhu et al. 1996). Optical bleaching may be used to reduce the color-center density for crystals of poor radiation hardness. It has been extensively studied for BaF_2 (Ma and Zhu 1995) and PWO (Zhu et al. 1996, 1998, 1999, 2002, 2004) in the past and is actively pursued for PWO (Semenov et al. 2007, 2008, 2009). In this case, a precision monitoring is mandatory to follow the variations of the crystal's light output caused by the variations of the crystal's transparency.

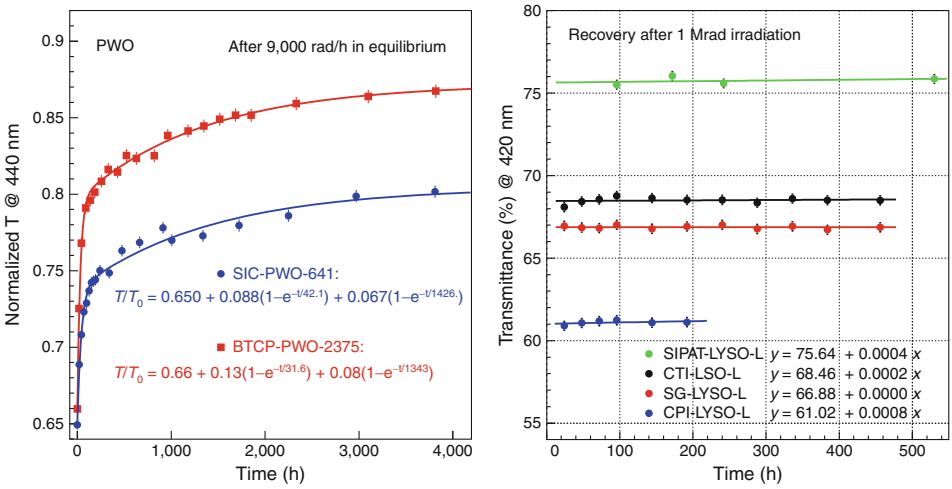


Fig. 5

The recovery of γ-ray-induced transmittance damage is shown as a function of time after the irradiation for a PWO sample (*left*) and an LSO/LYSO sample (*right*)

4.2 Radiation-Induced Color Centers

The longitudinal transmittance data can be used to calculate the light attenuation length of the crystal according to (Ma and Zhu 1993)

$$LAL = \frac{\ell}{\ln\left\{[T(1-T_s)^2]/\left[\sqrt{4T_s^4 + T^2\left(1-T_s^2\right)^2} - 2T_s^2\right]\right\}}, \tag{2}$$

where T is the longitudinal transmittance measured along crystal length ℓ, and T_s is the theoretical transmittance assuming multiple bouncing between two crystal ends and without internal absorption:

$$T_s = (1-R)^2 + R^2(1-R)^2 + \cdots = (1-R)/(1+R), \tag{3}$$

and

$$R = \frac{(n_{\text{crystal}} - n_{\text{air}})^2}{(n_{\text{crystal}} + n_{\text{air}})^2}, \tag{4}$$

where n_{crystal} and n_{air} are the refractive indices for crystal and air, respectively.

The radiation-induced absorption coefficient, or the color-center density D, can be calculated according to (Ma and Zhu 1993, 1995)

$$D = 1/LAL_{\text{after}} - 1/LAL_{\text{before}}, \tag{5}$$

where LAL_{after} and LAL_{before} are the light attenuation lengths after and before the irradiation.

The radiation-induced absorption-coefficient spectrum can also be presented as a function of the photon energy and be further decomposed to a sum of several color centers with Gaussian energy distributions (Wei et al. 1990; Zhu et al. 1991).

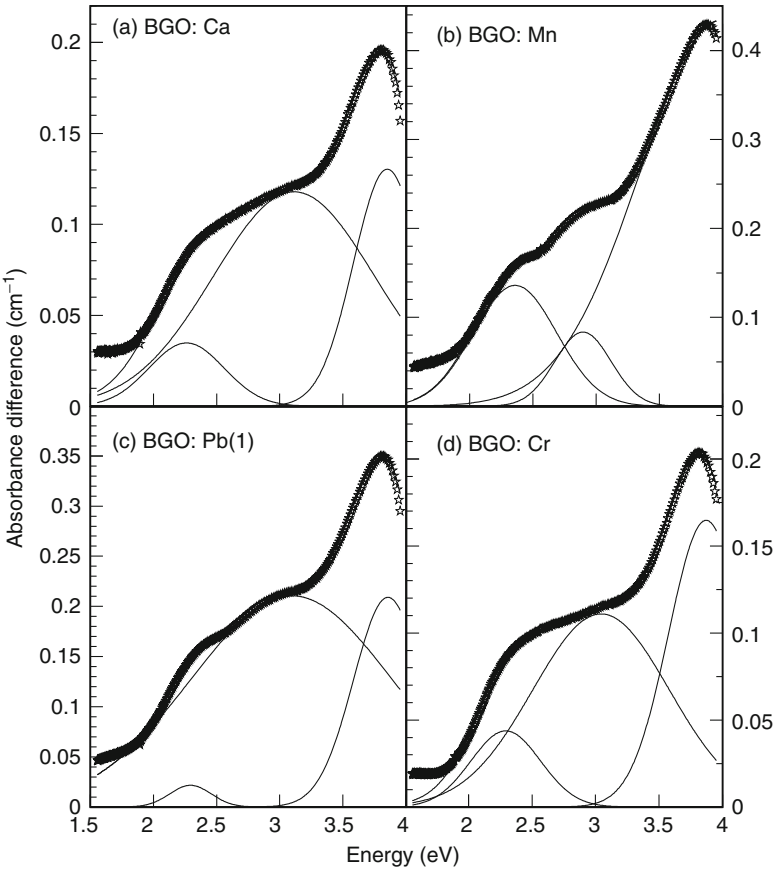


Fig. 6

The radiation-induced absorption-coefficient spectra (*lines with stars on top*) are shown as a function of the photon energy for four doped BGO samples. The spectra are decomposed to sums of three color centers (*plain solid lines*)

$$D = \sum_{i=1}^{n} A_i e^{-\frac{(E-E_i)^2}{2\sigma_i^2}}, \tag{6}$$

where E_i, σ_i, and D_i denote the energy, width, and amplitude of the color center i, and E is the photon energy.

Figures 6 and *7* show the radiation-induced color-center densities plotted as a function of the photon energy, respectively, for four BGO samples doped with Ca, Mn, Pb, and Cr (left) and two PWO samples in the equilibrium under the γ-rays irradiations with dose rate of 100 rad/h and 9,000 rad/h (right). It is interesting to note that although the overall shapes of these radiation-induced absorption coefficients are rather different, the Gaussian decompositions show the color centers located at the same energy and with the same width. While there

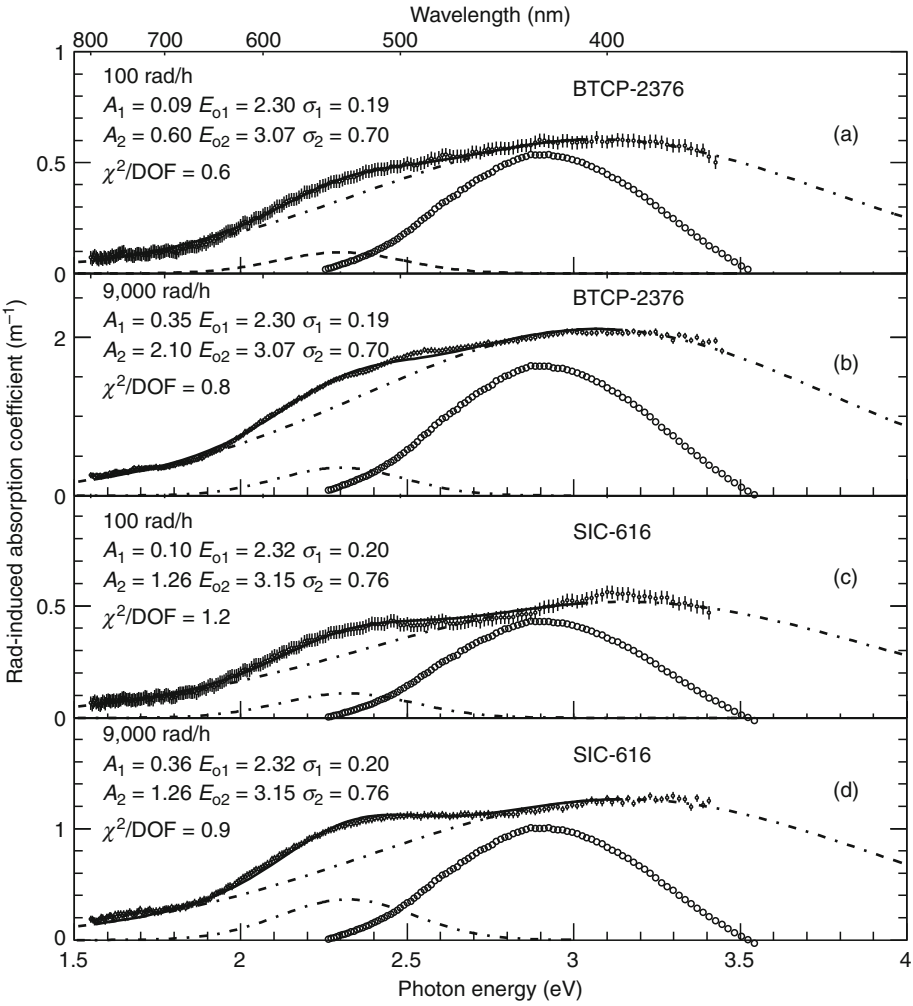


Fig. 7

The radiation-induced absorption-coefficient spectra (*data points with error bars*) are shown as a function of the photon energy for two PWO samples in the equilibrium at two different dose rates. The spectra are decomposed to sums of two color centers (*dot-dashed lines*). Also shown in the figure is the emission spectrum of PWO

are three color centers peaked at 2.3, 3.0, and 3.8 eV for all BGO samples, the PWO samples show two color centers peaked at 2.3 and 3.1 eV.

These observations hint that the color centers in these oxide crystals are caused by crystal-structure-related defects, such as oxygen vacancies, not particular impurities. The readers are referred to the corresponding references (Wei et al. 1990; Zhu et al. 1991, 1996, 1998, 1999, 2002, 2004) for more discussions about these color centers.

4.3 Dose-Rate Dependence and Color-Center Kinetics

Because of the balance between two processes – the color-center creation (irradiation) and the color-center annihilation (room-temperature recovery) – the radiation damage may be dose-rate dependent. Assuming that the annihilation speed of the color center i is proportional to a constant a_i and its creation speed is proportional to a constant b_i and the dose rate (R), the differential change of color-center density when both processes coexist can be expressed as (Ma and Zhu 1993, 1995):

$$\mathrm{d}D = \sum_{i=1}^{n} \left\{ -a_i D_i \,\mathrm{d}t + \left(D_i^{\mathrm{all}} - D_i \right) b_i R \,\mathrm{d}t \right\}, \tag{7}$$

where D_i is the density of the color center i in the crystal and the summation goes through all the centers. The solution of Eq. 7 is

$$D = \sum_{i=1}^{n} \left\{ \frac{b_i R D_i^{\mathrm{all}}}{a_i + b_i R} \left[1 - \mathrm{e}^{-(a_i + b_i R)t} \right] + D_i^0 \mathrm{e}^{-(a_i + b_i R)t} \right\}, \tag{8}$$

where D_i^{all} is the total density of the trap related to the color center i and the D_i^0 is its initial value. The color-center density in the equilibrium (D_{eq}) depends on the dose rate (R):

$$D_{\mathrm{eq}} = \sum_{i=1}^{n} \frac{b_i R D_i^{\mathrm{all}}}{a_i + b_i R}. \tag{9}$$

Following this equation, the optical transmittance, and thus the light output, would decrease when crystals are exposed to a radiation with a certain dose rate until they reach an equilibrium. At the equilibrium the speed of the color-center formation (damage) equals to the speed of the color-center annihilation (recovery), so that the color-center density (radiation-induced absorption) does not change unless the dose rate applied changes. More detailed discussions on the behavior of the color centers with bleaching light can be found in Ma and Zhu (1993, 1995).

Equation 9 also indicates that the damage level is not dose-rate dependent if the recovery speed (a_i) is small, which is the characteristics of the radiation damage caused by deep color centers. For crystals with no dose-rate dependence, an accelerated irradiation with a high dose rate would reach the same result as a slow irradiation with a low dose rate provided that the total integrated dose is the same. This is clearly shown in the transmittance data of a BaF_2 crystal in the top right plot of *Fig. 3*.

5 Light-Output Degradation

The light output of a crystal scintillator is a convolution of the crystal's emission spectrum, the light propagation inside the crystal, and the quantum efficiency (QE) of the photodetector. All these are wavelength-dependent. Although the crystal emission and photodetector QE are not affected by the radiation, the efficiency of the light propagation is affected by the variations of the light attenuation length and thus the radiation damage.

The left plot of *Fig. 8* shows the normalized light output as a function of time when the γ rays are applied at a defined dose-rate step by step from 15 rad/h up to 400 rad/h for a PWO sample. The dose-rate dependence of the γ-ray-induced radiation damage in PWO is clearly

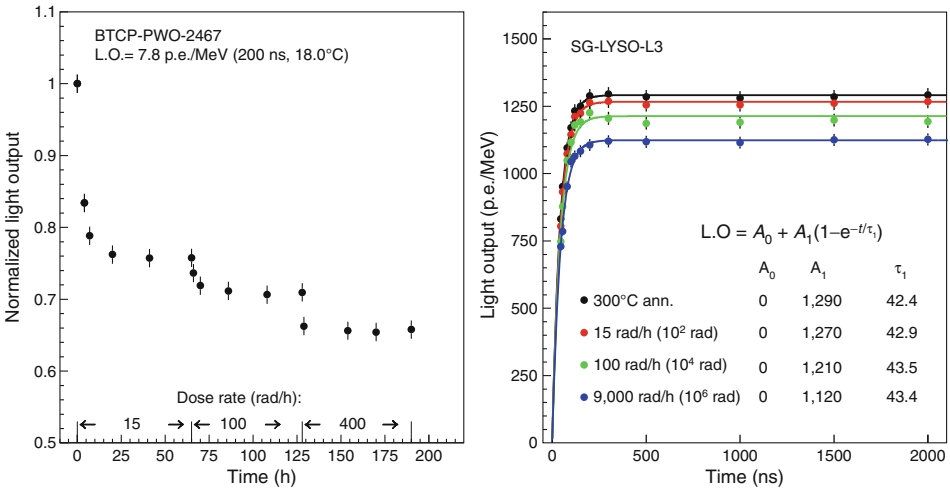


Fig. 8

***Left*: The normalized light output is shown as a function of time during several steps of the γ-ray irradiations with the dose rate up to 400 rad/h for a PWO sample. *Right*: The light output is shown as a function of the integration time after several steps of γ-ray irradiations with integrated dose up to 1 Mrad for an LYSO sample**

shown. The right plot of *Fig. 8* shows the light output as a function of the integration time together with the exponential fits to the scintillation decay time for a LYSO sample after several steps of irradiations with the cumulated dose up to 1 Mrad. It is clear that while the light output degrades the scintillation decay time remains the same, which is consistent with no damage to the scintillation mechanism.

The radiation hardness of LSO and LYSO crystals against γ rays (Mao et al. 2009a), neutrons (Mao et al. 2009b), and charged hadrons (Nessi-Tedaldi et al. 2009) has been found to be excellent. *Figure 9* shows the expanded longitudinal transmittance spectra (left) and the normalized average light output (right) for four LSO and LYSO samples. For the light-output measurement the seed/ID end of these samples is coupled to the readout device: XP2254 PMT (top) and two S8664-55 APDs (bottom). All samples tested have a consistent radiation resistance, with the degradations of the EWLT and the light output of approximately 12% for a γ-ray dose of 1 MRad. Because of these advantages, LYSO crystals are being considered for several future HEP experiments, such as SuperB and the CMS endcap calorimeter upgrade.

6 Light-Response Uniformity

An adequate light-response uniformity along the crystal length is a key for maintaining the crystal precision at high energies. The light-response uniformity of a long crystal may be parameterized as a linear function

$$\frac{LY}{LY_{\text{mid}}} = 1 + \delta(x/x_{\text{mid}} - 1), \tag{10}$$

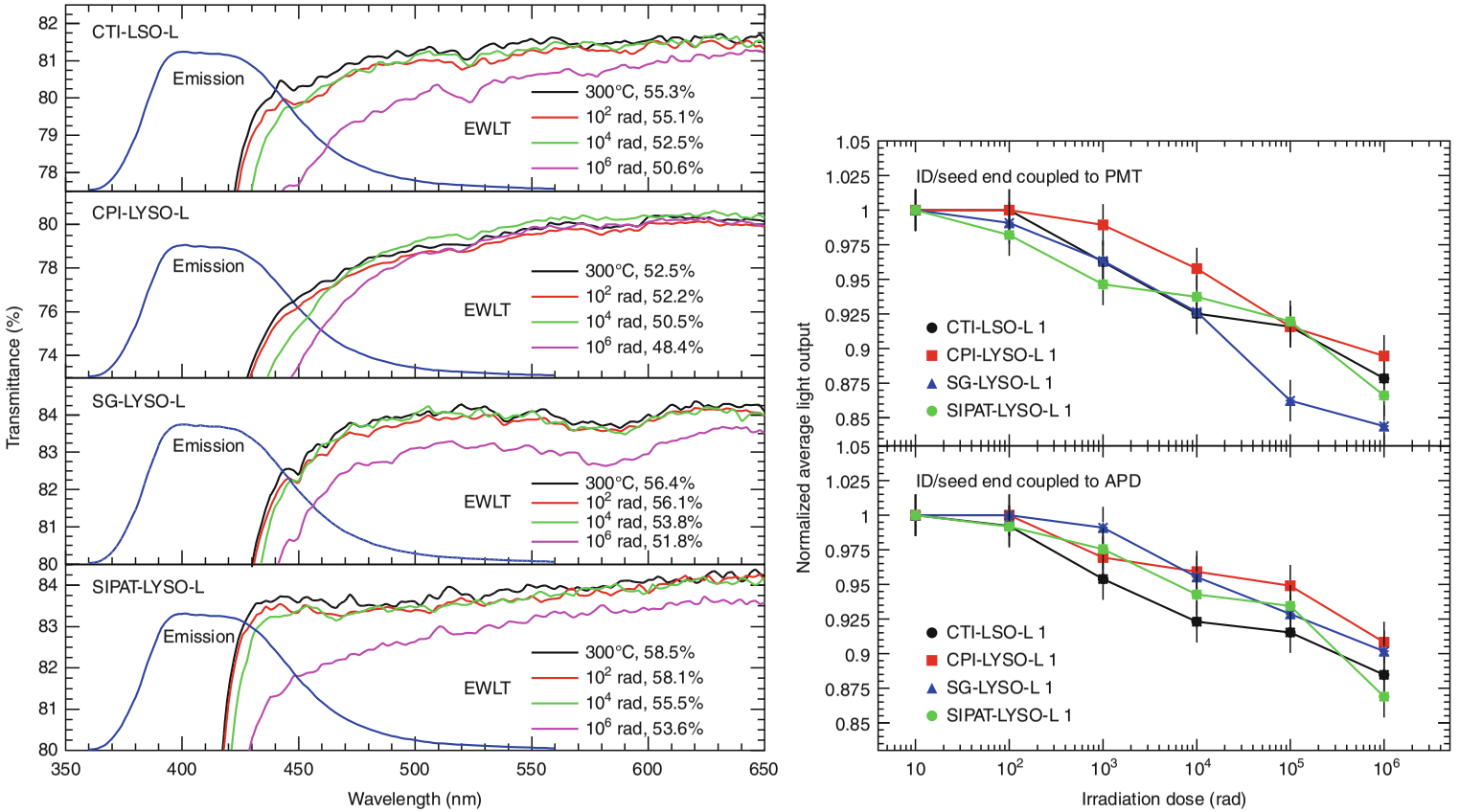


Fig. 9

The longitudinal transmittance spectra (*left*) and the normalized light output (*right*) are shown as a function of the integrated dose for four long LSO and LYSO samples

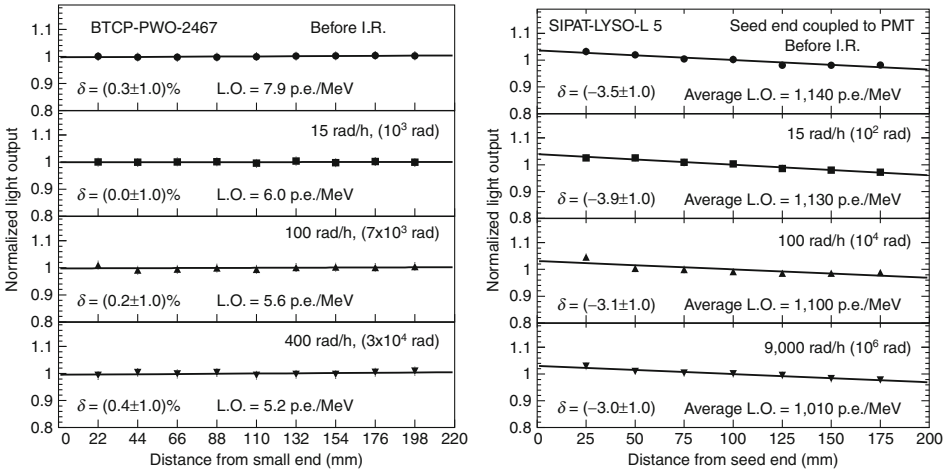


Fig. 10

The light-response uniformity as a function of the distance to the small end and the end coupled to the PMT for a PWO sample (*left*) and an LYSO sample (*right*), respectively, after several steps of γ-ray irradiations

where (LY_{mid}) represents the light output measured at the middle point of the crystal, δ represents the deviation from the flat response, and x is the distance from one end of the crystal.

Figure 10 shows the light-response uniformity after several steps of the γ-ray irradiations for a PWO sample (left) and an LYSO sample (right). The γ-ray irradiations were carried out step by step under a fixed dose rate in each step for the PWO sample. It is clear that the shape of the light-response uniformity was not changed for both crystals, indicating that the energy resolution is not compromised by the γ-ray irradiations. This is due to the fact that the degraded light attenuation length is long enough to maintain the light-response uniformity as predicted by a ray-tracing simulation for the light propagation inside the crystal (Zhu 1998).

7 Damage Mechanism in Alkali Halide Crystals and CsI(Tl) Development

Material analysis is crucial for identifying the radiation damage mechanism. The Glow Discharge Mass Spectroscopy (GDMS) analysis was used to search for correlations between the trace impurities in the CsI(Tl) crystals and their radiation hardness. Samples were taken 3–5 mm below the surface of the crystal to avoid surface contamination. A survey of 76 elements, including all of the lanthanides, indicates that there are no obvious correlations between the detected trace impurities and the crystal's susceptibility to the radiation damage. This indicates an important role of the oxygen contamination which cannot be determined by the GDMS analysis.

Oxygen contamination is known to cause radiation damage in the alkali halide scintillators. In BaF_2 (Zhu 1994), for example, hydroxyl (OH^-) may be introduced into crystal through a

hydrolysis process and later decomposed to interstitial and substitutional centers by radiation through a radiolysis process. Equation 11 shows a scenario of this process:

$$OH^- \rightarrow H_i^0 + O_s^- \text{ or } H_s^- + O_i^0, \tag{11}$$

where the subscript *i* and *s* refer to the interstitial and substitutional centers, respectively. Both the O_s^- center and the U center $\left(H_s^-\right)$ were identified (Zhu 1994).

Following the BaF_2 experience, significant improvement of the radiation hardness was achieved for CsI(Tl) crystals by using a scavenger to remove oxygen contamination (Zhu et al. 1996). *Figure 11* (left) shows the normalized light output as a function of the integrated dose for several CsI(Tl) samples, and compared to the *BaBar* radiation-hardness specification (solid line) (Zhu 1998). While the late samples SIC-5, 6, 7, and 8 satisfy the *BaBar* specification, early samples SIC-2 and 4 do not.

The improvement of the CsI(Tl) quality was achieved following an understanding that the radiation damage in the halide crystals is caused by the oxygen or hydroxyl contamination. Various material analyses were carried out to quantitatively identify the oxygen contamination in the CsI(Tl) samples. Gas Fusion (LECO) was found not sensitive enough to identify the oxygen contamination in CsI(Tl) samples. The identification of oxygen contamination was achieved by using the Secondary Ionization Mass Spectroscopy (SIMS) analysis. A Cs ion beam of 6 keV and 50 nA was used to bombard the CsI(Tl) sample. All samples were freshly cleaved prior to being loaded into the UHV chamber. An area of 0.15 × 0.15 mm^2 on the cleaved surface was analyzed. To further avoid the surface contamination, the starting point of the analysis is at about 10 μm deep inside the freshly cleaved surface. The right plot of *Fig. 11* shows the depth profile of the oxygen contamination for two radiation-soft samples (SIC-T1 and SIC-2)

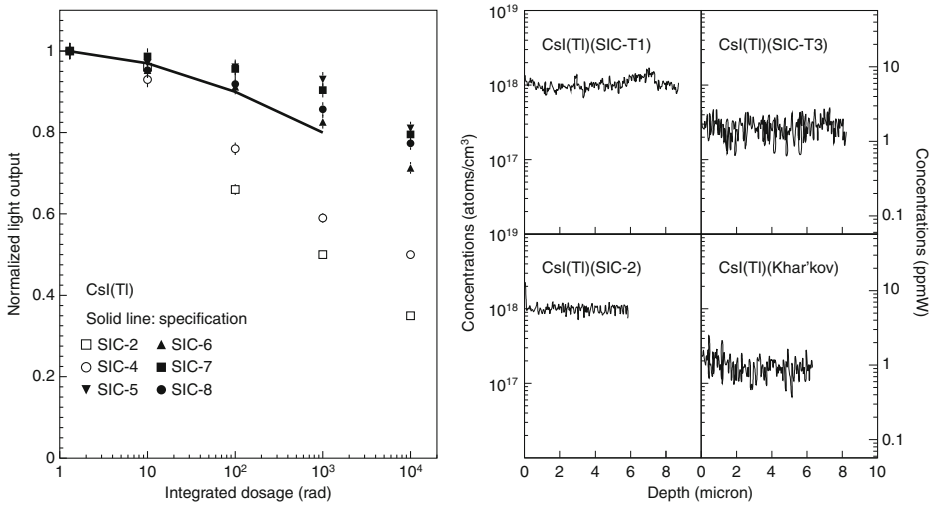


Fig. 11

***Left*: The progress of the CsI(Tl) radiation hardness is shown for CsI(Tl) samples together with the *BaBar* radiation-hardness specification. *Right*: The depth profile of the oxygen contamination is shown for two rad-soft CsI(Tl) samples (SIC-T1 and SIC-2) and two rad-hard samples (SIC-T3 and Khar'kov)**

and two radiation-hard samples (SIC-T3 and Khar'kov). Crystals with poor radiation resistance have oxygen contamination of 10^{18} atoms/cm^3 or 5.7 PPMW, which is five times higher than the background count (2×10^{17} atoms/cm^3, or 1.4 PPMW).

8 Damage Mechanism in Oxide Crystals and PWO Development

Similarly, GDMS analysis was carried out for BGO and PWO crystals, and was found to have no particular correlation with the crystal's radiation hardness. This hints an important role of the structure-related defects in the crystal which cannot be determined by the GDMS analysis. Crystal structure defects, such as oxygen vacancies, are known to cause radiation damage in oxide scintillators. In BGO, for example, three common radiation-induced absorption bands at 2.3, 3.0, and 3.8 eV were found in a series of 24 doped samples (Wei et al. 1990; Zhu et al. 1991) as shown in *Fig. 6*. Observations in the PWO crystals are similar with two color centers peaked at 2.3 and 3.1 eV as shown in *Fig. 7*. Following these observations, effort was made to reduce the oxygen vacancies in PWO crystals. An oxygen compensation approach, which was carried out by a post-growth thermal annealing in an oxygen-rich atmosphere, was found effective in improving PWO's radiation hardness for samples up to 10 cm long (Zhu et al. 1996, 1998, 1999, 2002, 2004). This approach, however, is less effective for longer (25 cm) crystals which show a variation of the oxygen vacancies along the crystal. In practice, yttrium doping, which provides a local charge balance for oxygen vacancies and so prevents the color-center formation, was found effective for PWO (Zhu et al. 1996, 1998, 1999, 2002, 2004). *Figure 12* shows the normalized light output as a function of time for three PWO samples under the γ-ray irradiations with a dose rate of 15 rad/h. PWO samples, produced in late 2002 with yttrium doping, are much more radiation hard than the early samples.

This improvement of PWO quality was achieved following an understanding that the radiation damage in the oxide crystals is caused by the oxygen vacancies. Various material analyses were carried out to quantitatively identify the stoichiometry deviation and the oxygen vacancies in the PWO samples. Particle-Induced X-ray Emission (PIXE) and quantitative wavelength-dispersive Electron Micro-Probe Analysis (EMPA) were tried. PWO crystals with poor radiation hardness were found as having a non-stoichiometric W/Pb ratio. Both PIXE and EMPA, however, does not provide the oxygen analysis. X-ray Photoelectron Spectroscopy (XPS) was found to be very difficult because of the systematic uncertainty in oxygen analysis. Electron Paramagnetic (or Spin) Resonance (EPR or ESR) and Electron-Nuclear Double Resonance (ENDOR) were tried to find unpaired electrons, but were also found to be difficult to reach a quantitative conclusion.

The identification of the oxygen vacancies is achieved by using the Transmission Electron Microscopy (TEM) coupled to Energy Dispersion Spectrometry (EDS) with a localized stoichiometry analysis. A TOPCON-002B Scope was first used at 200 kV and 10 μA. The PWO samples were made to powders of an average grain size of a few μm and then placed on a sustaining membrane. With a spatial resolution of 2 Å, the lattice structure of the PWO samples was clearly visible. *Figure 13* (left) shows a TEM picture taken for a sample with poor radiation hardness. Black spots of a diameter of 5–10 nm were clearly seen in the picture. On the other hand, the samples with good radiation hardness show a stable TEM picture with no black spots, as shown in *Fig. 13* (right). By employing TEM/EDS, a localized stoichiometry analysis

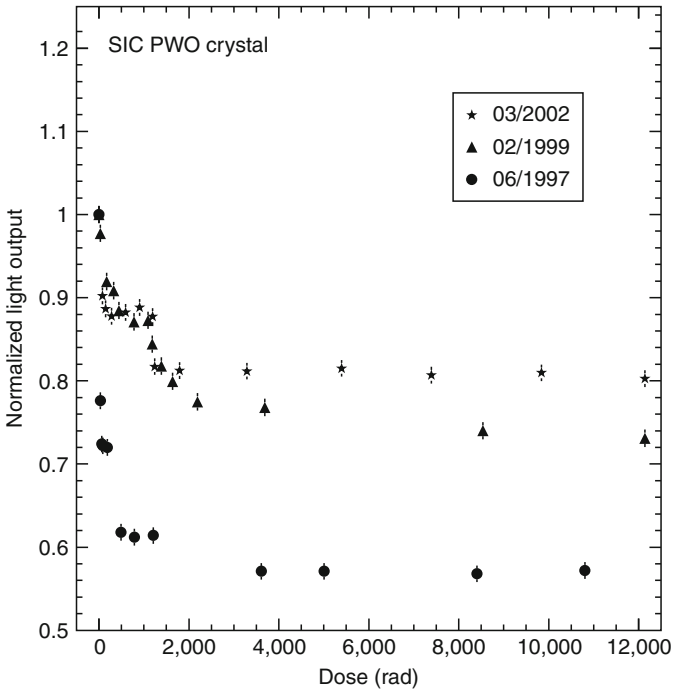


Fig. 12

The progress of PWO radiation hardness is shown for PWO samples from SIC

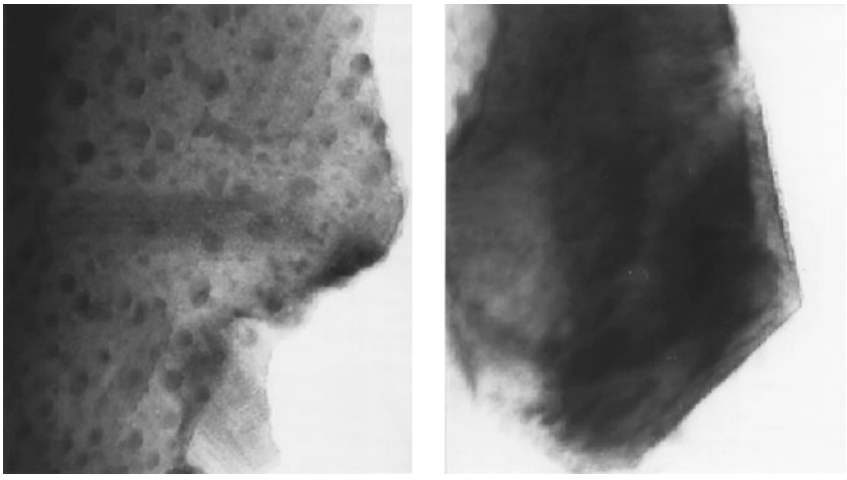


Fig. 13

TEM pictures of a PWO crystal of poor radiation hardness (*left*), showing clearly the black spots of ø5–10 nm related to oxygen vacancies, as compared to that of a good one (*right*)

Table 3

Atomic fraction (%) of O, W, and Pb in PWO samples measured by TEM/EDS (Yin et al. 1997)

As-grown sample				
Element	Black Spot	Peripheral	$Matrix_1$	$Matrix_2$
O	1.5	15.8	60.8	63.2
W	50.8	44.3	19.6	18.4
Pb	47.7	39.9	19.6	18.4
The same sample after oxygen compensation				
Element	$Point_1$	$Point_2$	$Point_3$	$Point_4$
O	59.0	66.4	57.4	66.7
W	21.0	16.5	21.3	16.8
Pb	20.0	17.1	21.3	16.5

was carried out. The system is a JEOL JEM-2010 scope and a Link ISIS EDS. The spatial resolution of this system allows a localized stoichiometry analysis in a region of a diameter of 0.5 nm. Black lack spots were observed in an as-grown sample. Localized stoichiometry analysis was carried out for points inside and surrounding the black spots, as well as points far away from the black spots. The uncertainty of the analysis is about 15%. The resultant atomic fractions (%) at these areas are listed in *Table 3* (Yin et al. 1997).

A clear deviation from the atomic stoichiometry of O:W:Pb = 66:17:17 was observed for samples taken inside these black spots, pointing to a severe deficit of the oxygen component. In the peripheral area, the oxygen deficit was less, but still significant. There was no oxygen deficit observed in the area far away from the black spots. As a comparison, the same sample after a thermal annealing in an oxygen-rich atmosphere was reanalyzed. No black spot was found. The result of the analysis is also listed in *Table 3*. In all randomly selected points no stoichiometric deviation was observed. This analysis thus clearly identified oxygen vacancies in PWO samples of poor radiation hardness.

9 Conclusion

Crystal scintillators suffer from radiation damage with possible effects including: (1) the scintillation-mechanism damage, (2) the radiation-induced phosphorescence, and (3) the radiation-induced absorption. No experimental evidence has been observed for the scintillation-mechanism damage in any crystals studied so far. All crystals show the radiation-induced phosphorescence and absorption. The radiation-induced phosphorescence increases the dark current of the photodetector, and thus the readout noise. The energy-equivalent noise is low for crystals with high light yield. The predominant radiation damage effect in the crystal scintillators is the radiation-induced absorption, or color-center formation. The radiation-induced absorption may recover spontaneously at the application temperature and leads to dose-rate dependence. Thermal annealing and optical bleaching are also found to be effective for shallow color centers, but not for all crystal scintillators.

The radiation damage in the alkali halide crystals is understood to be caused by the oxygen and/or hydroxyl contamination as shown by the SIMS analysis. By using a scavenger to remove the oxygen contamination, the radiation hardness of the mass-produced CsI(Tl)

crystals is improved. The radiation damage in the oxide crystals is understood to be caused by the stoichiometry-related defects, for example oxygen vacancies, as shown by the localized stoichiometry analysis with TEM/EDS. By using yttrium doping, the radiation hardness of the mass-produced PWO crystals is improved.

Acknowledgments

Measurements at Caltech were carried out by Drs. J.M. Chen, Q. Deng, H Wu, D.A. Ma, R.H. Mao, X.D. Qu and L.Y. Zhang. This work was supported in part by the US Department of Energy under grant DE-FG03-92-ER-40701 and the US National Science Foundation Award PHY-0612805.

References

Batarin VA, Butkler J, Chen TY, Davidenko AM, Derevschikov AA, Goncharenko YM et al (2003) Study of radiation damage in lead tungstate crystals using intense high-energy beams. Nucl Instr Meth A512:488–505, A530:286–292 (2004) and A540:131–139 (2005)

Chen JM, Mao RH, Zhang LY, Zhu R-Y (2005) Large size LYSO crystals for future high energy physics experiments. IEEE Trans Nucl Sci 52:2133–2140 (2005) and IEEE Trans Nucl Sci 54:718–724 (2007)

Cooke DW, McClellan KJ, Bennett BL, Roper JM, Whittaker MT, Muenchausen RE (2000) Crystal growth and optical characterization of cerium-doped $Lu1.8Y0.2SiO5$. J Appl Phys 88:7360–7362

Gratta G, Newman H, Zhu R-Y (1994) Crystal calorimeters in particle physics. Annu Rev Nucl Part Sci 44:453–500

Huhtinen M, Lecomte P, Luckey D, Nessi-Tedaldi F, Pauss F (2005) High-energy proton induced damage in PbWO4 calorimeter crystals. Nucl Instr Meth A545:63, A564:164 (2006) and A587:266 (2008)

Kimble T, Chou M, Chai BHT (2002) Scintillation properties of LYSO crystals. In: IEEE NSS Conference Record, Norfolk, pp 1434–1437

Ma DA, Zhu R-Y (1993a) Light attenuation length of barium fluoride crystals. Nucl Instr Meth A333:422–424

Ma DA, Zhu R-Y (1993) On optical bleaching of barium fluoride crystals. Nucl Instr Meth A332: 113–120 and Nucl Instr Meth A356:309–318 (1995)

Mao RH, Zhang LY, Zhu R-Y (2008) Optical and scintillation properties of inorganic scintillators in high energy physics. IEEE Trans Nucl Sci NS-55:2425–2431

Mao RH, Zhang LY, Zhu R-Y (2009a) Gamma ray induced radiation damage in PWO and LSO/LYSO crystals. Paper N32-5 in IEEE NSS 2009 Conference Record

Mao RH, Zhang LY, Zhu R-Y (2009b) Effect of neutron irradiations in various crystal samples of large size for future crystal calorimeter. Paper N32-4 in IEEE NSS 2009 Conference Record

Melcher C, Schweitzer J (1992) Cerium-doped Lutetium oxyorthosillicate: a fast efficient new scintillator. IEEE Trans Nucl Sci NS-39:502–505

Nessi-Tedaldi F, Dissertori G, Lecomte P, Luckey D, Pauss F (2009) Studies of Ceriun fluoride, LYSO and lead tungstate crystals exposed to high hadron fluences. Paper N32-3 in IEEE NSS 2009 Conference Record

Semenov PA, Uzunia AV, Davidenko AM, Derevschikov AA, Goncharenko YM, Kachanov VA et al (2007) First study of radiation hardness of lead tungstate crystals at low temperature. Nucl Instr Meth A562:575–580, IEEE Trans Nucl Sci NS-55:1283–1288 (2008) and Paper N32-2 in IEEE NSS 2009 Conference Record (2009)

Wei ZY, Zhu RY, Newman H, Yin ZW (1990) Radiation resistance and fluorescence of Europium doped BGO crystals. Nucl Instr Meth A297:163–168

Yin ZW, Li PJ, Feng JW (1997) TEM study on lead tungstate crystals. In: Zhiwen Yin et al (eds) Proceedings of SCINT97 International Conference. CAS, Shanghai Branch, pp 191–194

Zhu RY (1994) On quality requirements to the barium fluoride crystals. Nucl Instr Meth A340:442–457

Zhu RY (1997) Precision crystal calorimetry in future high energy colliders. In: IEEE NSS1996 Conference Record, published in IEEE Trans Nucl Sci NS-44:468–476

Zhu R-Y (1998) Radiation damage in scintillating crystals. Nucl Instr Meth A413:297–311 and references therein

Zhu RY, Stone H, Newman H, Zhou TQ, Tan HR, He CF (1991) A study on radiation damage in doped BGO crystals. Nucl Instr Meth A302:69–75

Zhu RY, Ma DA, Wu H (1996) CsI(Tl) radiation damage and quality improvement. In: Antonelli A et al (eds) Proceedings of the 6th International Conference on Calorimetry in High Energy Physics. Frascati Physics Series, Bologna, Italy, 589–598

Zhu RY, Ma DA, Newman HB, Woody CL, Kierstead JA, Stoll SP, Levy PW (1996) A study on the properties of lead tungstate crystals. Nucl Instr Meth A376:319–334 (1996), IEEE Trans Nucl Sci 45:688–691 (1998), Nucl Instr Meth A438:415–420 (1999), Nucl Instr Meth A480:470–487 (2002) and IEEE Trans Nucl Sci 51:1777–1783 (2004)

Further Reading

Claeys C, Simoen E (2002) Radiation effects in advanced semiconductor materials and devices. Springer, Berlin

Grupen C, Shwartz B (2008) Particle detectors. Cambridge University Press, Cambridge

Holmes-Siedle A, Adams L (2002) Handbook of radiation effects. Oxford University Press, Oxford

Iniewski K (2010) Radiation effects in semiconductors. CRC Press, Boca Raton

Knoll G (2000) Radiation detection and measurement, 3rd edn. Wiley, New York

Lecoq P, Annekov A, Gektin A, Korzhik M, Pedrini C (2005) Inorganic scintillators for detector systems. Springer-Verlag, Berlin, Heidelberg

Part 3

Applications of Detectors in Particle and Astroparticle Physics, Security, Environment and Art

23 Astrophysics and Space Instrumentation

John W. Mitchell[1] · *Thomas Hams*[1,2]
[1]NASA/GSFC, Greenbelt, MD, USA
[2]CRESST/University of MD Baltimore County, Baltimore, MD, USA

1 Introduction ***560***

2 Photon Instruments ***561***
2.1 X-Ray Calorimeters 561
2.2 Grazing-Incidence Optics 564
2.3 Coded Aperture Masks 566
2.4 Pair Conversion 568

3 Cosmic-Ray Instruments ***571***
3.1 Time-of-Flight Versus Energy Measurements 573
3.2 dE/dx Versus Total Energy 575
3.3 Magnetic Rigidity Spectrometers 577
3.4 Calorimeters 582
3.5 Large-Area Composition Experiments 585
3.6 Indirect Measurements 587

4 Conclusion ***591***

References ***591***

C. Grupen, I. Buvat (eds.), *Handbook of Particle Detection and Imaging*,
DOI 10.1007/978-3-642-13271-1_23, © Springer-Verlag Berlin Heidelberg 2012

Abstract: Instrumentation for particle and high-energy photon measurements in space must provide high levels of performance while meeting the severe constraints imposed by flight. Direct measurements are required spanning over 13 decades in energy and covering species ranging from photons to the heaviest nuclei in the periodic table. Indirect measurements increase the energy range by another five decades. Many of the detection techniques used are shared with accelerator instruments and other ground-based applications, but the implementation is often unique to space. This chapter sets the context for the required measurements and reviews representative instruments for direct measurements of photons and particles from 100 eV to 10^{15} eV and indirect measurements to over 10^{20} eV.

1 Introduction

Instruments to measure high-energy photons, X-rays and γ-rays, and energetic particles are key tools in modern astronomy and astrophysics. High-energy photons are produced by a wide variety of processes in which particles, particularly electrons, are accelerated to relativistic velocities, or in which material is elevated to extreme temperatures. The particle acceleration processes that produce high-energy photons are also likely sources of highly energetic particles. Detected at velocities approaching the speed of light, these particles, known as cosmic rays, include atomic nuclei and electrons, as well as positrons and antiprotons. Direct high-energy photon and particle detection spans 13 orders of magnitude in energy from X-rays of ~100 eV to particles near the "knee" of the cosmic-ray spectrum at about 10^{15} eV. The instrumentation required varies greatly depending on the energy and species to be observed (Chap. 1, "Interactions of Particles and Radiation with Matter"). Indirect measurements extend more than another five orders of magnitude to above 10^{20} eV, see Chap. 24, "Indirect Detection of Cosmic Rays." In this chapter, the techniques used for high-energy astrophysics measurements from flight platforms are reviewed and representative instruments are discussed.

Designers of space instrumentation face a number of special challenges. Whether for balloons, sounding rockets, or satellites, instruments must conform to strict weight, dimension, and power limits. Size is a particular issue. Larger, heavier instruments cost more to build and test. For space-based instruments, higher weight also demands more powerful launch vehicles with accompanying, often dramatic, increases in cost. Balloon payloads are limited by the capacity of the balloon vehicles and payload weight determines the altitude, and therefore the level of residual atmosphere, that can be reached. This is a compound problem because heavier payloads require stronger balloons, which themselves are heavier. Power is usually supplied by photovoltaic arrays that have limited area, supplemented by batteries whose weight has to be considered. Heat generated by the electronics of flight instruments must be dissipated by radiation to space because even at balloon altitudes (~36 km) the atmosphere is too thin to support convective cooling. At the same time, instruments must contend with the heat load of exposure to unobstructed sunlight. Instruments must be reliable and largely autonomous while incorporating the versatility to allow reconfiguration on command to change operational modes or compensate for degradation. Space-based instruments must contend with the rigors of the launch environment including shock, vibration, and acoustic loads. Balloon instruments

have to survive transportation to remote launch locations and shocks at parachute deployment and landing. For all instruments, cost is a major factor.

2 Photon Instruments

In this section, we review space and suborbital instrumentation for the direct measurement of highly energetic photons (X-ray, γ-ray), covering an energy range from ~0.1 keV to ~300 GeV. Given the large number of current instruments in this category, we can only select representative missions and discuss their instrumentation as exemplary of other missions using similar techniques. The techniques discussed include collimation (RXTE), grazing-incidence focusing optics (Chandra), coded aperture mask (Swift-BAT), and pair-production tracking (Fermi-LAT). The need to carry out X-ray observations above the Earth's atmosphere is apparent when considering that a 20 keV photon has an interaction length of 10 m in air. Starting in the hard X-ray range, suborbital observations become possible with balloon-borne or sounding rocket payloads. The first X-ray observations were conducted as early as 1962 (Giacconi et al. 1962) employing Geiger counters on board a sounding rocket to demonstrate detection of an X-ray source outside the solar system.

2.1 X-Ray Calorimeters

The Rossi X-Ray Timing Explorer (RXTE) is the longest operating of the currently active NASA X-ray missions. RXTE was launched in December 1995, into a 580 km circular orbit with 23° inclination, and the instrument is still providing data. The main science objective of RXTE is a time variability study of X-ray emissions in the energy range from 2 to 250 keV with microsecond time resolution and moderate energy resolution for bright sources. The RXTE spacecraft has the ability to quickly repoint the observatory to highly variable sources, such as Gamma-Ray Bursts (GRBs).

The nonimaging and non-focusing X-ray detector system of RXTE is a good starting point for this discussion, since similar techniques (proportional and scintillation counters) had been used in earlier missions (Chap. 11, "Gaseous Detectors," Chap. 15, "Scintillation Counters"). RXTE employs large-area collimated X-ray detectors giving a narrow field of view of the target region and reducing unwanted background detection.

RXTE is composed of three main instruments: a large-area Proportional Counter Array (PCA, in Chap. 11, "Gaseous Detectors") covering the range 2–60 keV (Jahoda et al. 2006), the High Energy X-ray Timing Experiment (HEXTE) (Gruber et al. 1996) with an energy range of 15–250 keV, and an All Sky Monitor (ASM) operating in the 2–10 keV energy range (Levine et al. 1996). The spacecraft can point the fixed PCA in any desired target direction except for a 30° exclusion zone in the direction of the Sun. Two star trackers provide pointing information to better than 130 arcsec and the spacecraft maintains a pointing accuracy for the PCA of better than 6 arcmin. The spacecraft can be brought on target with a slewing rate of 180° in 30 min. *Figure 1* shows the RXTE spacecraft with location of the detector systems.

The PCA is made up of five identical modules with a total collection area of 6,500 cm^2. Each PCA module has two sealed gas volumes. The main detection volume is filled with Xe/CH_4 gas and has four layers of proportional counters. Each layer is made up of a frame with 20

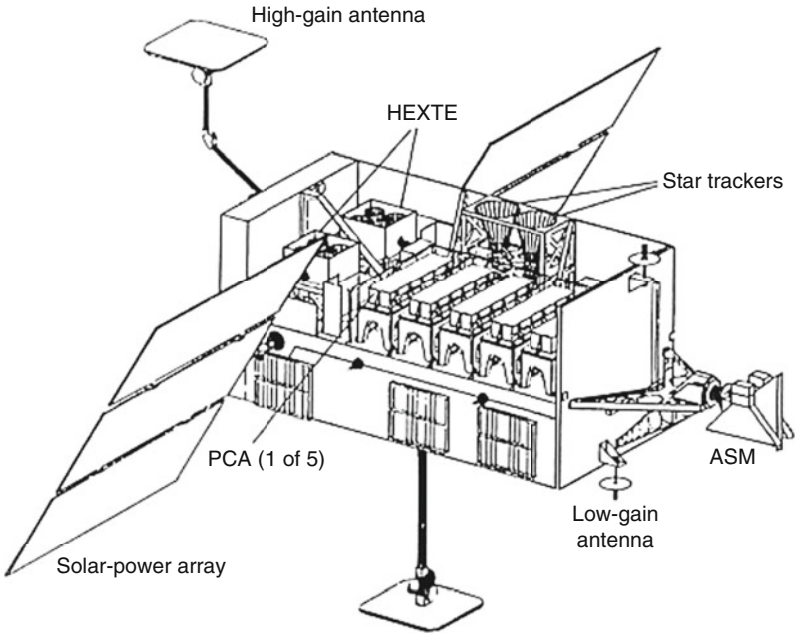


Fig. 1
RXTE spacecraft

cells, each nominally 1.3 cm × 1.3 cm × 100 cm, with an anode wire in the center. The second gas volume is filled with propane and uses a single layer of proportional cells, placed on the Xe volume, that serves as an entrance veto and is similar in construction to the other layers. In addition, the anode signals of the lower (fourth) layer and the outermost cells in the remaining layers of the Xe volume are used to form a three-sided anticoincidence. Both gas volumes are maintained at ~1 atm of pressure. The sides and bottom of the PCA employ a passive graded-Z shielding of tantalum, tin, and the aluminum housing to reject cosmic rays or X-rays that do not enter through the front. The remaining 18 anode wires in each of the three Xe layers define the active detection volume of the calorimeter. Collimators in front of the PCA provide a 1° (FWHM) field of view (FOV). The entrance window to the propane volume and the divider between the propane and Xe volumes are thin aluminized Mylar.

Telemetry bandwidth limits do not permit transmission of the digitized pulse heights of all anode signals. To maintain high internal background rejection and good detection efficiency, a signal encoding scheme was incorporated. The anode signals of the side cells and of the bottom layer are combined into a single veto signal that is sensitive both to penetrating radiation from the sides and to partially contained events. The anode wires of the propane layer are combined to build an anticoincidence to reject incident cosmic rays. In each layer, the remaining anode signals of alternate cells are combined to form a total of six signals in the active volume. Each of the six anode signals and two veto signals is individually discriminated and digitized. A nominal good event requires there to be no signal in any of the veto layers and a signal in only a single proportional cell.

Incident X-rays below 60 keV predominantly lose energy through photoelectric absorption in the Xe gas. The X-ray produces a photoelectron which in turn creates electron–ion pairs in

the Xe. The number of pairs is directly proportional to the energy of the incident X-ray. The proportional cell amplifies the signal and generates a measurable charge pulse on the anode wire. A pressure transducer and temperature sensor on each PCA unit record the state of the gas and thus the gas amplification. An ^{241}Am source in the PCA provides gain calibration. Time tagging of events is fast enough to allow for detection of microsecond-time-scale X-ray flux changes. The instrument has a sensitivity of 0.1 mCrab and an energy resolution of 18% at 6 keV. Higher-energy X-rays lose energy through Compton scattering and spread their energy deposit over a larger volume in the PCA. An increased number of proportional cells with a signal from a high-energy X-ray reduces the background rejection of the PCA and thus sets the upper energy limit of ~60 keV.

The High-Energy X-ray Timing Experiment (HEXTE), which is co-aligned with the PCA, provides high-energy measurements from 15–250 keV. HEXTE is divided into two clusters each employing four phoswich detectors with an effective collection area of ~800 cm^2 and a FOV of 1°. To veto charged particles entering the phoswich detectors, the sides of the cluster are surrounded by a particle anticoincidence shield of four plastic-scintillator tiles, each viewed by two photomultipliers (PMT) via wavelength-shifting light guides positioned along two sides. This anticoincidence provides prompt vetoing of spurious background from effects such as Cherenkov radiation in the PMT glass and secondary particles generated in the collimators. Each cluster can tilt off target by ±1.5°. The tilts for the two clusters are in orthogonal planes and the tilt motion always maintains one cluster on target while the other is sampling off-target background data. The on-target dwell time is programmable between 16–128 s. Much like the PCA, each phoswich scintillator individually records the arrival time and energy of incident X-rays.

Each HEXTE phoswich detector is contained in an opaque, sealed housing, which prevents stray light and moisture from entering and provides magnetic shielding, see Chap. 15, "Scintillation Counters." The detector material is an inorganic scintillator, NaI(Tl) (18.29 cm diameter, 0.32 cm thick), followed by a 5.71 cm thick CsI(Na) crystal, which serves as an anticoincidence. The scintillation light is viewed by a 12.7 cm PMT. X-rays enter the detector through a Pb honeycomb collimator, restricting the FOV to 1°, and a 0.5 mm thick Be entrance window. The two scintillator materials have different characteristic rise times. NaI(Tl) emits light after roughly 0.25 μs whereas CsI(Na) has a rise time of 0.63 μs. A pulse-shape analysis of the PMT signal can discriminate a purely fast NaI(Tl) signal (i.e., a good event) from an event with a slower CsI(Na) component. The latter class of events are rejected since they could result from an incident cosmic ray, partial energy containment in the detector by an X-ray with a Compton-scatter electron escaping the NaI(Tl) crystal, or an X-ray entering from the side. The energy resolution of the detector, 15% at 60 keV, relies on a calibration of measured PMT signal amplitude as a function of incident X-ray energy. The scintillation light yield is proportional to the X-ray energy deposit in the NaI(Tl) and the signal amplitude of the PMT is proportional to the collected scintillation light. Each phoswich module has a gain calibrator mounted in one cell of the collimator. The calibrator uses a plastic scintillator doped with a small amount of ^{241}Am and viewed by a 1.27 cm PMT. The plastic scintillator is placed directly in front of the entrance window. The primary decay scheme of ^{241}Am yields a 60 keV X-ray in coincidence with a 4 MeV alpha particle. The X-ray, leaving the plastic scintillator, provides a spectral line for gain calibration of the PMT, and the alpha, which stops in the plastic, provides a coincidence signal. To maintain uniform response across the phoswich aperture, both crystals are highly polished, optically coupled to the PMT, and wrapped in a diffuse Teflon reflector.

2.2 Grazing-Incidence Optics

The instruments discussed in the previous subsection do not provide imaging of X-ray sources, but rather obtain temporal or spectroscopic observations from a given region in the sky defined by the viewing angle of the collimator. Here we will discuss current X-ray instruments for astrophysical observations, which employ focusing optics. Given the strong absorption of X-rays traversing matter, refractive optics used in telescopes for visible light are not applicable. However, X-rays can be reflected off mirror surfaces provided the incident angle is shallow and the glancing angle is less than the critical angle, as Compton pointed out in 1923. For this grazing-incidence technique, the critical angle depends on mirror material and X-ray energy. The critical angle is inversely proportional to the photon energy. Mirror surface materials with increasing Z have larger critical angles at the same photon energy, and most grazing-incident-angle X-ray telescopes employ high-Z gold- or iridium-coated mirrors. The grazing-incidence technique had been explored for X-ray microscopy when in the early 1950s Hans Wolter suggested three geometries for focusing X-rays. In the mid 1960s a prototype of a Wolter Type-I X-ray telescope was developed and later flown on a sounding rocket to image the soft X-rays from the Sun. The first space-based imaging X-ray telescope (Wolter Type-I) was onboard NASA's second High Energy Astrophysical Observatory, HEAO-2. The satellite, later renamed Einstein Observatory, was launched in late 1978, collecting data for 2 years and providing, among other important studies, the first high-resolution X-ray study of supernova remnants. Focusing systems have a much higher signal-to-noise (background) performance than non-focusing systems because the X-rays collected over a large area are imaged on a much smaller sensor. As a consequence, focusing systems are less likely affected by spurious background events. Presently, a number of space-based grazing-incident telescopes are in operation, most notably the European Space Agency's X-ray Multi-Mirror Mission (XMM-Newton) and NASA's Chandra X-ray Observatory (Chandra or CXO). These latter two missions have similarities in their basic design and have complementary science goals. Here we will illustrate the technique with Chandra.

Chandra, formerly known as the Advanced X-ray Astrophysical Facility (AXAF), was conceived as a large, high-sensitivity telescope serving the astrophysical community by accessing the entire sky with an availability of greater than 85% at any time, providing spectroscopy and imaging, including achieving modest energy resolution ($E/\Delta E \approx 10$–50) for spatially resolved spectroscopy. The optical system achieves imaging better than 0.5 arcsec FWHM. The instrument suite covers the energy range from 0.1 to 10 keV and has an effective collection area of 800 and 400 cm^2 at 0.25 and 5 keV, respectively. An artist's view of the Chandra spacecraft is shown in ❯ *Fig. 2* and instrumental details can be found on the Chandra X-ray Center web site (cxc.harvard.edu).

The observatory was launched into a low Earth orbit with the Space Shuttle Columbia (STS-93) in July 1999; the Inertial Upper Stage booster rocket carried the spacecraft into a highly elliptic orbit (apogee height ~142,400 km, perigee height ~6,400 km) with a period of approximate 63.5 h (2010). The first-light image was taken of Cassiopeia A and was released on August 6, 1999.

The main science components of Chandra are the High-Resolution Mirror Assembly (HRMA), which focuses incoming X-rays onto the Science Instrument Module (SIM) at the far end of the Optical Bench Assembly (OBA). The SIM contains the High-Resolution Camera (HRC) and the Advanced CCD Imaging Spectrometer (ACIS). Each instrument type is divided into an imager (-I) and spectrometer (-S) subset. The ACIS imagers have moderate

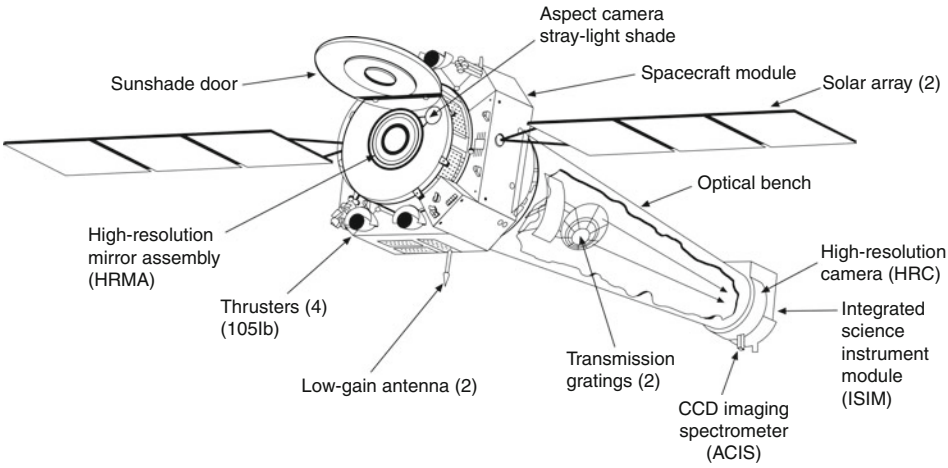


Fig. 2
Chandra spacecraft (NASA/CXC/NGST)

intrinsic spectral resolution, but in conjunction with the Low- (LETG) or High-Energy Transmission Gratings (HETG), which can be inserted into the beam, both photosensor instruments (HRC, ACIS) provide high-resolution spectra.

The science topics that can be addressed by an observatory with high availability, large collection area, and high-resolution imaging and spectroscopy are wide ranging, and these capabilities enable observation of faint sources, distant sources, or the investigation of the structure and physical processes in astrophysical objects.

The HRMA utilizes a Wolter Type-I design comprised of four concentric, nested, grazing-incident X-ray mirror segments fabricated from Zerodur glass with a 330 Å iridium coating. The outer mirror segment has a diameter of 123 cm and the inner segment 65 cm. The paraboloid followed by hyperboloid mirror sections are each 85 cm long and together with the thermal pre- and post-collimator and aperture plate, the HRMA has a total length of 276 cm. The HRMA has a weight of 1,484 kg and a focal length of 10 m. The precision achieved in the HRMA drives the spatial resolution of 0.5 arcsec for Chandra and provides an order of magnitude improvement over previous instruments.

The ACIS is made up of ten planar, charged-couple devices (CCDs) each with 1,024 × 1,024 pixels, see Chap. 16, "Semiconductor Counters." An incident X-ray deposits its energy via photoelectric interaction in the CCDs. The hit pixel and the amount of charge stored characterize the incident location and energy of the X-ray. The CCDs are exposed with a frame time of ~3.2 s and read out with transfer time of ~41 ms. Telemetry limitations of the spacecraft permit only the transmission of data from up to six preselected CCDs. Four CCDs are arranged in a 2 × 2 array (ACIS-I) and used as an imager, the remaining six CCDs are arranged in a linear 1 × 6 array (ACIS-S) and are either used for imaging or as a grating readout. The ACIS has the capability to simultaneously acquire high-resolution images and moderate-resolution spectra. Most of the CCDs in the ACIS are front-illuminated except for two back-illuminated CCDs in the ACIS-S, which have a lower threshold. Since the CCDs are sensitive to the detection of visible light and X-ray photons, they are covered by a visible/UV-blocking filter, which is a thin

composite of polyimide sandwiched between two thin layers of aluminum. The quantum efficiency of the CCDs matches the energy range of the HRMA. To achieve high-quality spectra of point and slightly extended (few arc-seconds) sources, the linear ACIS-S array is arranged in the expected diffraction direction of the transmission grating system. The ACIS instrument can be used in conjunction with the HETG or LETG, but in normal operation is mostly used with the HETG.

The HETG is mounted on a support frame that can be swung into the beam behind the HRMA and contains two grating patterns: the Medium-Energy Grating (MEG) and the High-Energy Grating (HEG). The MEG, covering the range 0.4–5.0 keV, places the grating behind the outer two mirror segments, whereas the HEG, covering the range from 0.8–10 keV, places the grating behind the inner two mirror segments. The nominal grating parameters for the HEG (MEG) are as follows: The grating lines are 5,100 Å (3,600 Å) thick gold deposited on a 9,800 Å (5,500 Å) thick polyimide film. The periodicity is 2,000.81 Å (4,001.95 Å) and the bars have a width of 1,200 Å (2,080 Å). The ruling patterns on the HEG and MEG form an X on the ACIS-S, allowing it to separate the different diffraction patterns.

The second instrument in the focal plane of the HRMA is the High-Resolution Camera (HRC), which employs a chevron-type microchannel plate (MCP) coated with CsI to enhance photoelectric conversion and read out by a crossed-grid charge detector. The HRC has a spatial resolution of ~20 μm. To eliminate cosmic-ray background, the HRC incorporates a plastic-scintillator anticoincidence detector. Passive tantalum shielding on the inside of the titanium housing rejects X-rays entering from the side. The HRC has two subsets, one optimized for imaging (HRC-I) and the other (HRC-S) serving as the readout for the LETG, which is similar in its function to the HETG discussed above but optimized for low-energy observations starting at 70 eV. The HRC-I provides the largest FOV (~$30' \times 30'$) on the observatory, has an energy threshold below that of the ACIS, and has a good time resolution of ~ 16 μs, but lacks the spectral resolution of the ACIS.

In order to smooth out the pixel-to-pixel variation, the instantaneous image is spread over different pixels by dithering the spacecraft over the target in a Lissajous pattern. The amount and period of the dithering depends on the instrument in use and is 20 arcsec for the HRC and 8 arcsec for the ACIS with a nominal period of 700 s (pitch) and 1,000 s (yaw). The controlled motion of the spacecraft has to be taken into account to obtain final images.

Other mirrors used to focus X-rays using grazing incidence employ foil substrates instead of glass, allowing segments to be packed closer and weigh less. Foil mirrors were used in the Japanese ASCA (formerly Astro-D) satellite, which was launched on February 20, 1993, and operated for over 7 years. This mission is also noteworthy since it was the first X-ray satellite mission to use a CCD imager. New high-resolution mirrors developed for the NuStar mission (Koglin et al. 2005) use thin thermally formed glass shells with graded-depth multilayer coatings to extend focusing into the hard X-ray band from 8 to 80 keV.

2.3 Coded Aperture Masks

X-ray focusing using conventional grazing-incidence mirrors with single-layer coatings is technically feasible up to energies of ~10 keV. This method provides high angular resolution (0.5 arcsec, previous section) but has a narrow field of view of ~1° and a small collection area.

In the energy range above ~10 keV, where conventional grazing-incidence focusing is less effective, and below ~50 MeV, where pair production can be used to reconstruct the energy and direction of the incident photon, other imaging techniques are needed. One way to accomplish

imaging above ~10 keV is similar to a pinhole camera, where a position-sensitive photosensor is placed under a mask with a pinhole. A point source will illuminate a particular region of the photosensor and this location determines the arrival direction of the photons. By replacing the pinhole with a coded aperture mask, which for a given source direction casts a unique pattern on the photosensor, very large collection areas and good angular resolution can be achieved. To obtain the image of the source, the detected photons over the entire photosensor need to be deconvolved. Encoding incident photons can be done in one of two ways, temporally or spatially. Most present missions employ a spatial-coded aperture mask, which we discuss below. An example of current temporal encoding can be found on the All Sky Monitor (ASM) of RXTE. In the ASM three narrow FOV detectors sweep in different planes over the region of interest in the sky. Each detector records the detected count rate as a function of time (direction). The direction of a point source is marked by the time (direction) for which the photon flux is the greatest. The three scans in different planes constrain the source location.

While the image in a focusing X-ray telescope is only affected by the noise of pixels in the immediate vicinity of the focused image, in the coded-aperture technique noise from the entire photosensor affects the quality of the image.

Current space missions using coded aperture masks are the International Gamma-Ray Astrophysics Laboratory (INTEGRAL) and Swift. We will use the latter to illustrate this technique. INTEGRAL was launched by the European Space Agency (ESA) in late 2002 and is an active mission. INTEGRAL is dedicated to the fine spectroscopy ($E/\Delta E = 500$) and fine imaging (angular resolution: 12 arcmin FWHM) of celestial γ-ray sources in the energy range 15 keV–10 MeV with concurrent source monitoring in the X-ray (3–35 keV) and optical (V-band, 550 nm) energy ranges and employs a number of coded-aperture-mask instruments.

Swift is the latest active mission that utilizes a coded aperture mask and has the largest mask ever deployed. This mission is conducted as an international collaboration among Italy, the UK, and the USA, funded in the USA by NASA. The main science focus of the Swift mission is to study GRBs (Gehrels et al. 2009). Swift has a complement of three co-aligned instruments to view GRBs and their afterglows at γ-ray, X-ray, ultraviolet (UV), and optical wavelengths: the Burst Alert Telescope (BAT), the X-Ray Telescope (XRT), and the Ultraviolet/Optical Telescope (UVOT) (Gehrels et al. 2004). Swift was launched November 20, 2004, and detects ~300 GRBs/year, localizing ~100 GRBs/year.

The largest instrument on board Swift is the BAT (Barthelmy et al. 2005), which can view approximately one sixth of the sky at any time or 1.4 sr (half-coded). The BAT can detect and acquire high-precision locations for GRBs and then relay a position estimate accurate to within 1–4 arcmin to the ground in approximately 15 s. ❯ *Figure 3* shows the detector plane of the BAT instrument. The main components of the BAT are the coded aperture mask, the detector array, and the graded-Z shielding. The detector array covers an area of 1.2 m × 0.6 m and has 32,768 pixels. Each pixel is a 4 mm × 4 mm × 2 mm CdZnTe crystal operated in photon counting mode. To maintain signal uniformity across the detector plane, the detector-array temperature is kept at 294 ± 1 K. The energy range of the detector is from 15 to 150 keV with an energy resolution of ~5 at 60 keV. The coded aperture mask has a D-shape with an area of 2.7 m^2 and is placed 1 m above the detector array. The approximately 54,000 Pb coding titles are 5 mm × 5 mm × 1 mm and are mounted to a composite substrate. The mask uses a completely randomized pattern (50% open and 50% closed) rather than the commonly used Unified Redundant Array (URA) pattern. To reduce the effect of cosmic rays on the detector array by 95%, a graded-Z shield surrounds the sides and the underside of the array. The average BAT background event rate is 10,000 events s^{-1} (or about 0.3 count s^{-1} per pixel), with orbital variations of a factor of two around this value. This yields a GRB fluence sensitivity of 2×10^{-8} erg cm^{-2} s^{-1} (15–150 keV).

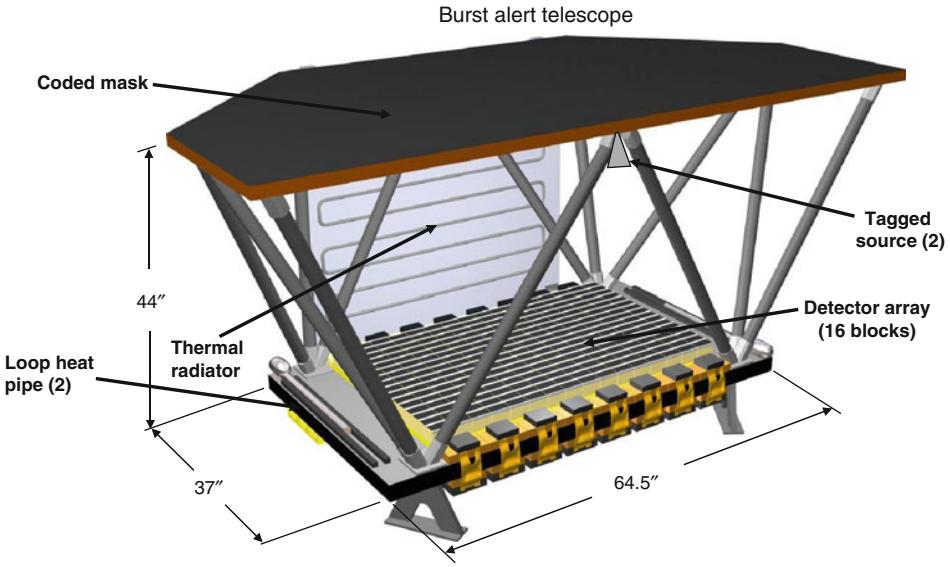


Fig. 3
BAT detector on Swift spacecraft

2.4 Pair Conversion

Above a few MeV, the dominant interaction mechanism of γ-rays with matter is their conversion to e^+e^- pairs in the electric field of a nucleus or an electron. Pair production has a threshold of 1.022 MeV for interactions in the field of the nucleus and the rate of production goes as Z^2. The threshold for production by interaction in the field of an electron is ~2 MeV and the production rate for atomic electrons is proportional to Z. Thus, in heavy materials, the interaction takes place primarily near the nucleus. Pair production can be used as the basis for a high-energy γ-ray observatory by tracking the momentum vectors of the e^+ and e^- and measuring their combined energy, see ❯ Chap. 12, "Tracking Detectors." The directions of the particles are rapidly altered by multiple Coulomb scattering, and it is important to measure the tracks before scattering has a significant effect. Provided the e^+ and e^- have not deviated too much from their original track before measurement, the photon momentum vector is approximately the vector sum of the momenta of the particles. Similarly, the energy of the incident photon is the small recoil energy plus the sum of the energies of the e^+ and e^-. The latter can be measured using a calorimetric technique or, at low energies, by measuring the rate of multiple Coulomb scattering.

The pair-conversion telescope technique was pioneered in the NASA SAS-2 (Small Astronomy Satellite 2) instrument that flew from November 1972 to June 1973 and measured γ-rays in the 35–200 MeV range using a spark-chamber tracking system with thin tungsten plates for pair production. Energy was measured by determining the rate of multiple scattering. This was followed by the ESA Cos-B instrument that was launched in 1975 and operated for 6.5 years, ending in 1982, producing a catalog of 25 γ-ray sources, a γ-ray map of the Milky Way disc, and detection of the first extragalactic γ-ray source 3C273. Cos-B was able to add a crystal calorimeter to improve energy resolution, see ❯ Chap. 20, "Calorimeters." The promise of high-energy

γ-ray astronomy was realized in the EGRET (Energetic Gamma-Ray Experiment Telescope) instrument on the NASA Compton Gamma-Ray Observatory (CGRO) that launched in 1991 on the Space Shuttle Atlantis and operated until it was deorbited in 2000. EGRET used a spark chamber for pair production and tracking and had more than 20 times the geometry factor (area × solid angle) of its predecessors. The energy of the e^+e^- pair was measured using a NaI(Tl) crystal calorimeter. A plastic-scintillator anticoincidence dome vetoed charged-particle background. EGRET measured γ-rays from 20 MeV to 30 GeV and revolutionized γ-ray astronomy with the first all-sky survey above 50 MeV, detections of γ-ray pulsars and blazars (a class of active galactic nuclei) as well as observations of diffuse γ-ray emission, delayed emission from GRBs, and γ radiation from high-energy solar flares.

The Large Area Telescope (LAT) on the Fermi Gamma-ray Space Telescope (Fermi) (Atwood et al. 2009) was designed to clarify and extend EGRET observations. Fermi (formerly the Gamma-ray Large Area Space Telescope) was launched in June 2008. Fermi-LAT was designed to provide the observations needed to understand the nature of the high-energy photon sky including identification of sources, determining the origins of the diffuse emission, understanding the mechanism of particle acceleration, using γ-rays to probe the nature of dark matter, and using γ-rays to study the early universe and the evolution of γ-ray sources. LAT released a catalog of 1,451 sources in 2010, the largest γ-ray source catalog to date. About half the sources are either blazars (~600) or pulsars (~60). Other source classes, some newly discovered, include pulsar wind nebulae, supernova remnants, globular clusters, starburst galaxies, Seyfert galaxies, and X-ray binaries. In addition, the LAT has proven to be a highly effective detector for high-energy electrons with measurements of the electron spectrum to energies approaching 1 TeV.

The central design goal of Fermi-LAT was to measure the directions, energies, and arrival times of incident γ-rays over a wide FOV so that much of the sky is viewed on each orbit. All of the detector technologies used were based on current practice at particle accelerators. The LAT is designed to span the energy range from below 20 MeV to above 300 GeV using a pair-conversion telescope based on a silicon-strip-detector (SSD) tracking system interleaved with thin tungsten converter layers. The energies of the e^+e^- pair are measured in a fully active CsI(Tl) crystal calorimeter. The use of a monolithic anticoincidence limited the upper energy range of EGRET due to self-vetoing by backsplash albedo from interactions in the calorimeter. To eliminate this problem, LAT uses a highly segmented anticoincidence detector (ACD) to detect charged particles entering the FOV of the telescope. There is only a small probability of a backsplash event coincident in the same ACD tile as an incident γ-ray, and the ACD is not automatically included as a veto in the trigger. The instrument is arranged in 16 modules or towers, each incorporating both a tracking section and a calorimeter section. The tracker array is completely covered on the outside by the ACD. LAT dimensions are 1.8 m × 1.8 m × 72 cm. The instrument weighs 2,789 kg and consumes 650 W. The effective area is 9,500 cm^2 at normal incidence and the FOV is 2.4 sr. A schematic view of the Fermi-LAT is shown in Fig. 4.

The LAT converter–tracker has 18 tracking planes, each with two layers (x and y) of single-sided SSDs with 228 μm pitch and 400 μm thickness. The top 16 layers are preceded by thin tungsten converter foils with 0.03 radiation length (X_0) thickness (0.01 cm) over the first 12 layers and 0.18 X_0 (0.072 cm) foils over the last four layers. Each tracker layer has about 0.014 X_0 of support and detector material. The use of thin converters in the upper layers improves the point spread function (PSF) for the lowest energies. The thicker foils at the back of the tracker enhance the effective area and FOV at high energies while reducing the angular resolution at 1 GeV by less than a factor of 2. The lowest two tracker planes do not have converter material overlaying. The detectors and foils are supported in low-mass carbon-composite "trays" with

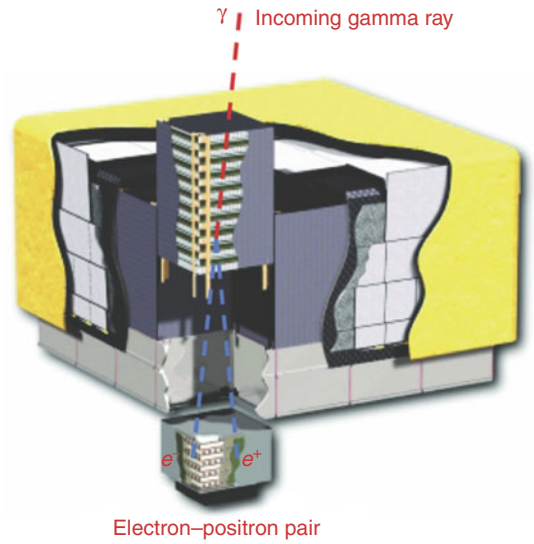


Fig. 4
Schematic view of the Fermi-LAT (Atwood et al . 2009)

aluminum honeycomb cores. Each tray is about 3 cm thick and carries two single layers of detectors, a converter foil, and front-end electronics. A detector layer is at the top and bottom of each tray and the foil is located above the lower detector layer. The SSD strips at the top and bottom of a tray are parallel, but successive trays are rotated by 90° to place x and y layers together. The tracker defines the FOV and provides the principal instrument trigger. The tracker design is compact, giving an aspect ratio for the full telescope of 0.4 to maximize the FOV.

The LAT calorimeter, below the tracker, has 96 CsI(Tl) crystals in each module, each 2.7 cm × 2 cm × 32.6 cm. This gives a lateral segmentation of 1 Molière radius and a longitudinal segmentation of 1.08 X_0. The crystals are optically isolated using reflective wrapping material and are arranged in 8 layers of 12 crystals each. Each layer is rotated 90°, forming an x, y hodoscope. The total depth of the calorimeter is 8.6 X_0. Combined with the converter–tracker, the full depth of the LAT is 10.1 X_0. Each crystal is read out by two photodiodes at each end, one with an area of 147 mm^2 and the other with an area of 25 mm^2, to span the needed dynamic range. The relative light levels measured at the ends of each crystal give a measure of the shower position. By fitting the profile of particle cascades, the longitudinal segmentation (segmentation in calorimeter depth) allows the energy of incident particles to be measured even if the shower produced by the particle is not fully contained in the calorimeter. This technique extends the energy measurements to ~1 TeV, although for uncontained showers the energy resolution is limited by fluctuations in shower leakage.

As noted, the LAT uses a highly segmented ACD to avoid self-vetoing by backsplash from high-energy showers in the calorimeter. The ACD is required to have a very high detection efficiency for charged particles, 0.9997% averaged over the full FOV, and to impose as little interaction mass in the photon beam as possible. This was achieved using an array of plastic-scintillator tiles read out by wavelength-shifting optical fibers coupled to PMTs at the periphery

of the ACD. Each tile is read out by two PMTs for redundancy. Tiles are overlapped in one dimension to eliminate gaps. The remaining gaps between tiles are closed by scintillating optical fiber ribbons with >90% detection efficiency for singly charged particles.

The LAT measures γ-rays and electrons with an energy resolution ranging from 9% at 100 MeV to 18% at 300 GeV. The single-photon angular resolution ranges from 3° at 100 MeV to less than 0.15° for energies above 10 GeV. The instrument normally operates in a scanning mode in which the axis of the instrument alternates on successive orbits between +35° from zenith and toward the pole of the orbit and −35° from zenith. For the 25.5° inclination 565 km Fermi orbit this gives virtually uniform sky coverage every two orbits (3 h). The telescope can also be pointed to address targets of opportunity.

3 Cosmic-Ray Instruments

Cosmic rays, high-energy charged particles traveling at speeds that can approach that of light, are a rich source of information on the chemical evolution of the Galaxy as well as on some of the most extreme environments and exotic processes in the universe. Cosmic rays have been detected with energies exceeding 10^{20} eV. Except at the highest energies, cosmic rays are isotropized by intergalactic, galactic, and heliospheric magnetic fields and their arrival directions do not point back to their sources.

The majority of cosmic rays are atomic nuclei from hydrogen to the heaviest elements in the periodic table, with energies spanning more than 13 orders of magnitude and steeply falling spectra decreasing by a factor of ~50 per decade in energy. Protons make up about 85% of cosmic-ray nuclei and helium about 12%. Galactic cosmic-ray (GCR) nuclei are most likely accelerated in supernova remnants and their elemental and isotopic composition probe nucleosynthesis, nuclear interactions with the interstellar medium (ISM), the distribution of freshly synthesized elements, and the mechanism of supernova explosions.

Primary cosmic-ray nuclei are produced directly in nucleosynthesis processes. Secondary cosmic-ray nuclei, including Li, Be, and B, and those elements directly below Fe in the periodic table are produced mainly by fragmentation of more abundant nuclei in interactions with the ISM. Both elemental and isotopic measurements are important. Measurements of stable secondary-to-primary ratios such as B/C and sub-Fe/Fe provide important information on the path length of ISM traversed by the GCRs. Measurements of radioactive secondary GCRs such as ^{10}Be, ^{26}Al, ^{36}Cl, ^{54}Mn, and ^{14}C can be used as "clocks" to determine the age of the cosmic rays and the fraction of time spent in the galactic halo. The abundances of other radioactive isotopes such as ^{59}Ni that decay by electron capture probe the time between nucleosynthesis and acceleration. Measurements of isotopes such as ^{22}Ne and the abundances of elements heavier than iron probe the origins of GCRs and the mechanisms by which nuclei are selected for acceleration.

The spectra of cosmic-ray nuclei reflect both source and transport effects (Strong et al. 2007). At energies above the influence of solar modulation, the GCR nuclear spectrum falls rapidly with an all-particle differential spectral index of about $E^{-2.7}$ to an energy of about 10^{15} eV. At this point, known as the "knee," the spectrum steepens. The reason for this is uncertain, but is commonly attributed to progressive failure of the supernova acceleration mechanism. Knee energies are at about the limit of direct measurements and most of the data comes from ground-based measurements (Chap. 24, "Indirect Detection of Cosmic Rays").

Between 10^{17} and 10^{18} eV, an extragalactic ultrahigh-energy cosmic-ray (UHECR) component progressively becomes dominant and above the "ankle" at about 5×10^{18} eV the spectrum flattens again. Both the nature of the cosmic acceleration engines responsible for such extreme energies and the composition of the UHECRs are uncertain and are the subjects of intense study. Candidate UHECR sources include active galactic nuclei, neutron stars, galaxy clusters, and the progenitors of GRBs. UHECRs are almost certainly atomic nuclei, but models of UHECR chemical composition range from pure protons, through mixtures of light, intermediate, and heavy species, to pure Fe. Current data from the Pierre Auger Observatory favor a heavier overall composition at the highest energies, but data from HiRes do not.

At energies above a few times 10^{19} eV, UHECRs interact with cosmic microwave and infrared background photons. The details of this interaction depend on the composition of the UHECRs. Protons interact by photoproduction, yielding pions, protons, and neutrons. Heavier nuclei undergo photodisintegration in which nucleons are scattered from the nucleus. Either process results in the loss of energy and causes a dramatic steepening of the cosmic-ray spectrum known as the GZK effect, named after Greisen, Zatsepin, and Kuzmin who predicted this suppression in 1966. In 2008, HiRes published the first significant observation of the GZK suppression, confirmed by the Pierre Auger Observatory. UHECRs measured with energies above the GZK cutoff must come from sources within a radius of about 100 megaparsecs of Earth. Particles at these energies have large gyroradii in extragalactic and galactic magnetic fields and their arrival directions should point back to their sources. This opens the possibility of charged particle astronomy by identifying and characterizing individual sources.

The interactions that produce the GZK cutoff also produce ultrahigh-energy neutrinos, known as GZK neutrinos or cosmogenic neutrinos. The expected flux of these neutrinos depends strongly on the UHECR composition. UHECR proton interactions produce e.g., π^+ that decay to a ν_μ and μ which then decay to a ν_μ, a ν_e, and an e. This process produces a spectrum of neutrinos extending to ultrahigh energies. Neutrons are also produced and can decay to produce neutrinos, but this is a minor component. Similarly, if nucleons resulting from photodisintegration of UHECR nuclei are above the pion photoproduction threshold then neutrinos are produced. The neutrino flux from this process depends on the opacity of the photon backgrounds to UHECR nuclei and the spectrum of the nuclei before photodisintegration. Oscillations over astrophysical distances result in a 1:1:1 ratio between the three ν flavors. Because the ν are not absorbed during propagation through the universe, the spectrum of cosmogenic ν arriving at the Earth should reflect the accumulated contribution of sources extending to high redshift.

Other cosmic-ray components include electrons, positrons, and antiprotons. While these are largely the result of interactions of nuclear cosmic rays with the ISM, they may have other origins. Positrons and electrons can be produced directly in astrophysical objects such as pulsars, and features in their spectra can provide important insights into nearby sources. Cosmic-ray particles may also be produced directly by the annihilation of dark-matter candidates such as neutralinos and Kaluza–Klein particles. Details of the spectra of resulting particles, especially positrons, electrons, and antiprotons, provide important constraints on the nature of dark matter.

The energy spectra of cosmic-ray species other than nuclei also reflect both their origins and transport to Earth. Electrons are largely secondary and the spectrum from the superposition of distant sources falls approximately as E^{-3} and softens rapidly above 1 TeV. Electrons lose energy quickly by synchrotron and inverse Compton processes and any detected with TeV

energy must have been accelerated within about 10^5 years and can have traveled at most a few hundred parsecs. At energies above 1 TeV, features from discrete sources might become evident in the high-energy cosmic-ray electron spectrum. A significant feature in the electron spectrum below 1 TeV might also indicate a nearby source of electrons, a pulsar, or dark-matter annihilation. Recent measurements of the high-energy electron spectrum differ considerably. ATIC and PPB-BETS report a feature in the 300–800 GeV range. Measurements by Fermi-LAT and H.E.S.S. show some excess flux near 1 TeV compared to model predictions, but not a distinct feature. H.E.S.S. measurements also indicate that the spectrum steepens above 1 TeV.

The positron spectrum also exhibits interesting features. PAMELA measurements show a significant excess of positrons over expectations of secondary production for energies above 10 GeV. Two general classes of explanation have been offered for the e^+ excess observed by PAMELA: the signature of a nearby pulsar, or group of pulsars, and annihilation radiation from a dark-matter clump. There is ample evidence that e^- and e^+ pairs are produced by primary e^- accelerated within pulsars and that the e^- and e^+ are subsequently accelerated to ultra-relativistic velocities. However, the mechanism by which the particles might escape the pulsar is unclear. Models for the dark-matter source are constrained by measurements of antiprotons from BESS and PAMELA. Most cosmic-ray antiprotons are secondaries produced by interactions of GCRs with the ISM. Production kinematics and the energy spectra of the primary cosmic rays give a characteristic secondary antiproton spectrum with a peak around 2 GeV and sharp decreases below and above the peak. The presence of an additional source such as dark-matter annihilation might be seen as a deviation from the secondary spectrum above or below the peak. Thus far, precision measurements from BESS-Polar and PAMELA have shown no significant excess.

The Sun acts both as a source of energetic particles and as a modifier of the GCR flux. Particles are emitted by the Sun in the solar wind and in coronal mass ejections (CMEs). Solar energetic particles (SEPs) are accelerated by impulsive solar flares and by interplanetary shocks from CMEs. In addition, matter from the local ISM can enter the solar system as neutrals and then be ionized by the solar wind. These ions can then be picked up by the solar wind and subsequently accelerated at the solar-wind termination shock, becoming anomalous cosmic rays (ACRs). The ACRs sample matter from the local ISM.

The magnetic fields entrained in the outflowing solar wind reduce the energies of GCRs entering the heliosphere. This effect, known as solar modulation, acts to redistribute the GCRs to lower energies and is often simplified as a spherically symmetric force field opposing the incoming GCRs (Fisk 1971). Solar modulation has a significant effect on the measured GCR spectrum below ~10 GeV/nucleon. This effect is not constant, but tracks the 11-year cycle of solar activity, having its greatest influence at solar maximum. Solar modulation depends on the magnetic polarity of the Sun as well, and particles of different charge sign, e.g., electrons and positrons or antiprotons and protons, are affected differently depending on the magnetic polarity of the Sun (Bieber et al. 1999). Understanding this charge-sign-dependent solar modulation is critical to developing detailed models of low-energy antiparticle fluxes.

3.1 Time-of-Flight Versus Energy Measurements

Ions with energies below a few MeV/nucleon, often reached by SEPs, CMEs, corotating-interaction-region (CIR) ions, and ACRs, can be measured with isotopic resolution using a time-of-flight (TOF) mass spectrometer (Chap. 6, "Particle Identification.") A representative

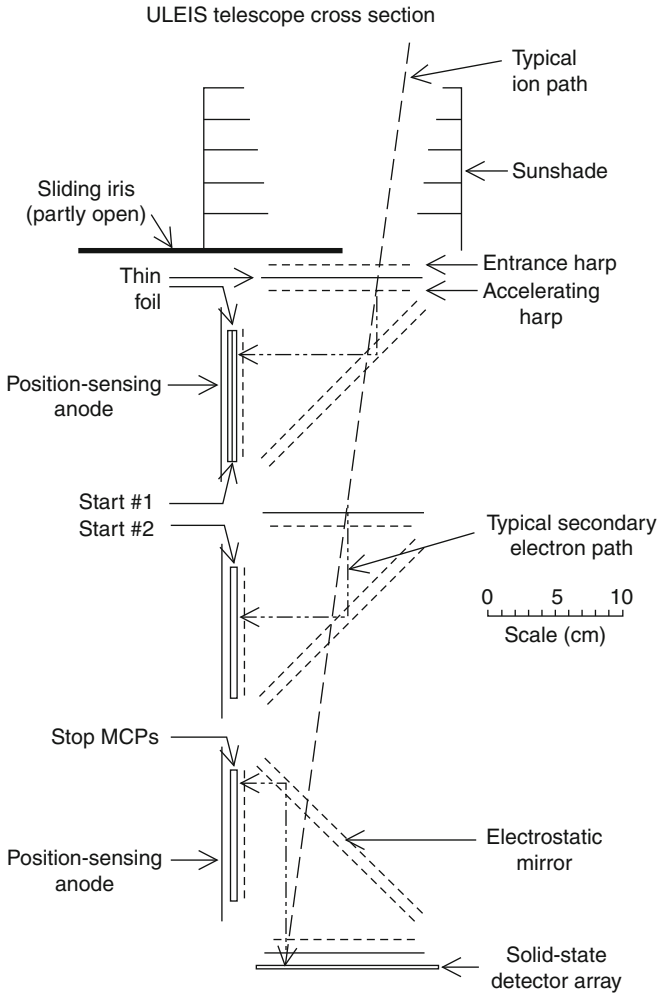


Fig. 5

Cross-sectional view of the ultralow-energy isotope spectrometer onboard ACE (Mason et al. 1998)

detector using this technique is the Ultralow-Energy Isotope Spectrometer (ULEIS) (Mason et al. 1998) on the Advanced Composition Explorer (ACE) mission, launched in 1997. A schematic cross-sectional view of ULEIS is given in *Fig. 5*. An ion entering the ULEIS acceptance cone passes through a series of thin metal foils, causing the emission of secondary electrons. The ion is then stopped in a solid-state detector, which measures its residual energy. The secondary electrons emitted from the foils are accelerated to ~1 keV and deflected onto microchannel plates (MCPs) using electrostatic mirrors. The resulting MCP pulses are discriminated and the time between pulses is measured. The secondary electron optics are isochronous so that the point of impact on the foil does not influence the measured time. The first two foils give redundant start times while the last foil gives a stop.

TOF is measured with ~300 ps resolution and with a 50 cm flight path ULEIS can measure ions with energies up to a few MeV/nucleon. The lower energy of ~45 keV/nucleon is set by the thicknesses of the foils and the front contact of the solid-state detector. The total kinetic energy of the particle, $E = mv^2/2$, corrected for the energy loss in the foils and detector contact, and the measured velocity $v = L/t$, where L is the flight path in the detector, give the mass of the ion: $m = 2E\,(t/L)^2$. ULEIS cannot measure charge. Particle species are determined by comparison to solar system abundances to determine the dominant isotope among isomers (same mass but different charge). This does not contribute significant ambiguity. ULEIS uses MCPs with areas of 8 cm× 10 cm and an array of solid-state detectors with an area of 73 cm^2 to give a geometric factor of 1.3 cm^2 sr. Simplified implementations of this technique can reach lower energies using thinner foils. A wide range of energies can be measured in a single instrument by replacing the single solid-state detector with a dE/dx-E telescope.

3.2 d*E*/d*x* Versus Total Energy

For energies below a few hundred MeV/nucleon, one of the most common techniques used for the identification of GCR or SEP nuclei is the simultaneous measurement of specific ionization energy loss, dE/dx, and total kinetic energy, E. This dE/dx-E technique, often implemented as dE/dx-total E or dE/dx-residual E methods, is capable of determining the charge, mass, and energy of incident particles.

In its simplest form, a dE/dx-E telescope utilizes two detectors, usually silicon diodes, to separately measure dE/dx and E. The thickness of the entrance detector, which measures dE/dx, sets the effective lower energy limit of the device and this detector is often as thin as possible. The second detector must be thick enough to stop particles within the energy range of interest. dE/dx is approximately $dE/(dL\sec(\theta))$, where dL is the thickness of the entrance detector and θ is the angle of the particle with respect to the telescope axis. Similarly, the particle's kinetic energy is approximately the energy deposited in the stopping detector if the dE/dx detector is thin. Ionization energy loss is proportional to Z^2/v^2 and kinetic energy equals $mv^2/2$. Thus, the product $dE/dx \times E$ is proportional to $Z^2m/2$ and is independent of velocity. As illustrated in ❯ *Fig. 6*, when dE/dx is plotted against total E, the result is a series of hyperbolas in constant $Z^2m/2$. Elements are separated more than isotopes, so the telescope functions to determine both the charge and mass of the nucleus. In order to achieve maximum resolution, θ must be either limited by a combination of collimation and detector geometry or the trajectory of the particle must be measured. This is most commonly accomplished with position-sensitive solid-state-detector hodoscopes, although other methods including drift chambers and scintillating-optical-fiber hodoscopes have also been used.

The modern dE/dx-E telescope was introduced on the IMP-1 (Explorer 18) satellite, launched in 1963. This telescope used thin solid-state entrance detectors with a CsI(Tl) crystal as a stopping detector. ISEE-3, launched in 1978, incorporated two instruments that provided measurements of the incoming particle trajectory, one using a drift chamber and one with position-sensitive solid-state detectors. Since then, dE/dx-E telescopes of great sophistication have been used on many missions including Voyager (I and II), CRRES, SAMPEX, Wind, NINA, Ulysses, ACE, and STEREO. Some representative examples are discussed below.

In most high-resolution solid-state telescopes, the charge and mass of detected particles are identified based on the energies they deposit coming to rest in a stack of detectors rather than in a single detector. This allows the instrument to stop particles of much higher energy than

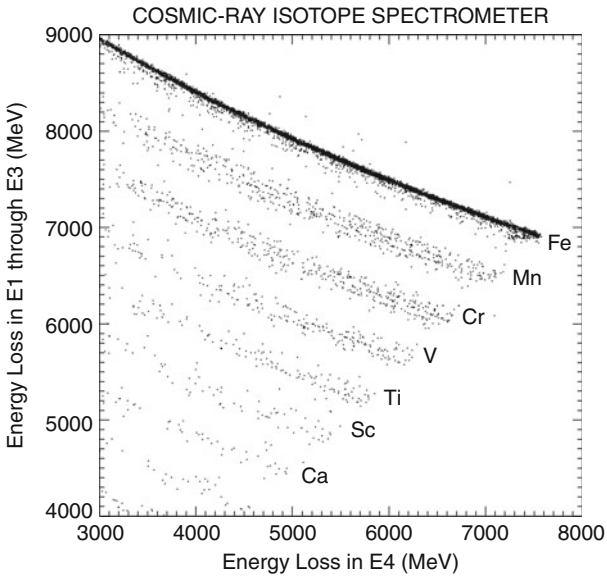


Fig. 6
dE/dx vs E technique used with ACE/CRIS (Stone et al. 1998c)

could be achieved with a single stopping detector. In addition, the resolution improves when dE/dx is measured in a detector or set of detectors whose thickness is a significant fraction of the particles range. Use of a stack of detectors allows optimization of the thicknesses of the detectors for dE/dx measurements over an extended energy range. An iterative technique for solving for mass is discussed in Stone et al. (1998a).

The ULYSSES Cosmic and Solar Particle INvestigation (COSPIN) incorporated five instruments (Dual-Anisotropy Telescopes (ATs), Low-Energy Telescope (LET), High-Energy Telescope (HET), High-Flux Telescope (HFT), and Kiel Electron Telescope (KET)) measuring nuclei and electrons from 0.5 MeV to several hundred MeV, depending on species. The HET was designed to measure elemental and isotopic composition of nuclei from hydrogen to nickel. The lowest energy measured was about 5 MeV for protons and extended to 400 MeV/nucleon for Fe. The telescope utilized a hodoscope formed of six 1,000 μm thick Li-drifted silicon (Si(Li)) detectors incorporating a multi-strip readout with position determined by voltage division. These were arranged in two groups of three detectors with successive detectors rotated by 60°, and measured both trajectory and dE/dx. The resultant resolution was about 150 μm, giving a trajectory resolution on the order of 1°. These were followed by a stack of six 5,000 μm Si(Li) detectors. For particles penetrating these detectors, dE/dx was accurately measured. The stack also stopped incident nuclei and provided the total E measurement. Finally the telescope was terminated in a 1,000 μm Si(Li) diode acting as a penetration detector. The full telescope was surrounded in a scintillator anticoincidence shield. The geometric factor of the HET depended on particle species and energy and ranged from about 3.6 cm^2 sr at the highest energies to ~87 cm^2 sr at the lowest.

ACE includes 4 dE/dx-E telescopes: the Solar Energetic Particle Ionic Charge Analyzer (SEPICA); the Electron, Proton, and Alpha Monitor (EPAM); the Solar Isotope Spectrometer (SIS) (Stone et al. 1998b); and the Cosmic-Ray Isotope Spectrometer (CRIS) (Stone et al. 1998c). SIS and CRIS both measure nuclei from He to Zn with isotopic resolution. The instruments are

complementary. SIS spans the energy range from about 10 MeV/nucleon to 100 MeV/nucleon to measure nuclei accelerated in SEP events, ACRs, and low-energy GCRs. CRIS measures from approximately 100 MeV/nucleon to 600 MeV/nucleon, focusing on measurements of GCR composition.

SIS uses a pair of identical telescopes composed of a stack of 17 solid-state detectors. The top two elements are 75 μm thick ion-implanted x, y (matrix) position-sensitive detectors with 34 cm^2 active areas that measure energy loss and particle trajectories. Each matrix detector has 64 readout strips in x and y each 960 μm wide and separated by 40 μm. Following these are 15 ion-implanted silicon stack detectors with 65 cm^2 active areas and thicknesses ranging from 100 μm for the top two stack detectors, increasing in thickness progressively to 3.75 mm. Finally, the stack has a 1 mm detector acting as a penetration counter. A collimator limits the opening angle of the instrument to 95° full angle. The SIS geometric factor ranges from 19.4 to 38.4 cm^2 sr. Special care was taken in mapping both the thicknesses of the SIS detectors and the thicknesses of the dead layers.

The CRIS instrument consists of a scintillating-optical-fiber-trajectory (SOFT) hodoscope for measuring the trajectory of nuclei, and four silicon solid-state detector telescopes for measuring dE/dx and E. The SOFT system consists of a hodoscope composed of three x, y scintillating-fiber planes (six fiber layers) and a trigger detector composed of a single fiber plane (two fiber layers). The hodoscope and trigger fibers are coupled to an image intensifier that is then coupled to a CCD camera for readout, and to photodiodes to obtain trigger pulses. Fully redundant image-intensified CCD camera systems view opposite ends of the fibers. Only one of these camera systems is operated at any given time because of power and bit-rate limitations. CRIS contains four detector telescopes to achieve a large collecting area and to provide redundancy. Each consists of stacks of 15 silicon detectors. The individual detectors are 10 cm in diameter and 3 mm thick Si(Li). CRIS has a geometric factor of 250 cm^2 sr for isotope measurements.

3.3 Magnetic Rigidity Spectrometers

Magnetic rigidity spectrometers use measurements of the curved trajectory of charged particles in a strong magnetic field to identify incident particles by directly measuring their charge (Z), charge sign, magnetic rigidity ($R = cp/Ze$, where p is momentum and Ze is electric charge), and velocity (β). This information is subsequently used to derive their p, mass (m), and kinetic energy (E_k). Magnetic rigidity spectrometers are unique in their ability to measure the charge sign and so are the principal instruments used for antiparticle (positron and antiproton) measurements and in searches for heavier ($|Z| \geq 2$) antinuclei. They can also provide isotope resolution to much higher energies than dE/dx-E telescopes, limited mainly by velocity resolution, the bending power of the magnet, and multiple scattering. Magnetic rigidity spectrometers require a strong magnet to deflect incident particles and a precise tracking system to measure their trajectories. Most of the instruments built to date have used superconducting magnets and gas-based tracking systems, multi-wire proportional counters or drift chambers, and have flown on balloons. However, recent advances in NdFeB permanent-magnet alloys with saturation magnetization exceeding 1.3 T coupled with silicon position-sensing detectors with resolutions of a few μm have opened the way for space-based magnetic spectrometers. This section will consider both a balloon-borne superconducting magnet instrument, BESS-Polar (Balloon-borne Experiment with a Superconducting Spectrometer – Polar) (Yamamoto et al. 2011) and space-based permanent-magnet instruments, PAMELA (Payload

for Antimatter and Matter Exploration and Light-nuclei Astrophysics) (Picozza et al. 2007) and AMS-02 (Alpha Magnetic Spectrometer – 02) (Kounine et al. 2010).

In the 1960s, Luis Alvarez of the University of California, Berkeley, recognized that a persistent-mode superconducting magnet could be successfully operated on a balloon or space platform. Single-coil balloon-borne magnetic spectrometers were developed at Berkeley and Johnson Space Flight Center, and a spectrometer with a near-Helmholtz magnet was developed at Goddard Space Flight Center. The Berkeley group also formulated plans for a spectrometer to be flown onboard the HEAO-B mission and a prototype was developed and tested. With advances in space cryogenics for IRAS, Spacelab-2, and COBE, this led to a NASA program to develop a large superconducting magnet facility on the International Space Station (ISS), known as Astromag, which would have had provisions to refill the cryostat on orbit and to change out experiments. Astromag was terminated for budget reasons in 1990. A free-flyer version was studied, but not funded.

The Astromag effort led to a new generation of balloon-borne magnetic spectrometers for studies of antimatter, elemental spectra, and light isotopes (LEAP, PBAR, SMILI, MASS, IMAX, BESS, TS93, CAPRICE, HEAT, and ISOMAX). PAMELA had its origin in the WiZard antimatter instrument selected for the first round of Astromag experiments. For historical context, see references in Mitchell et al. (2004). The AMS collaboration that flew the AMS-01 permanent-magnet spectrometer for 10 days on the Space Shuttle in 1998 initially included several members of the Astromag team. The AMS collaboration subsequently prepared a superconducting magnet version, AMS-02, to fly on the ISS. In light of the extended mission planned for the ISS, this was reconfigured to use the AMS-01 permanent magnet, which has no consumables. AMS-02 was successfully installed on the ISS in May 2011.

During the Astromag study, a number of magnet configurations were proposed. BESS originated from a proposal to use a solenoidal superconducting magnet whose coil was thin enough for particles to pass through with minimal interaction probability, tracked by detectors within the warm bore of the magnet. This configuration maximizes the opening angle of the instrument, and hence the geometric factor, making it ideal for rare-particle measurements. Versions of the original instrument were used for nine northern-latitude flights between 1993 and 2002. In order to take advantage of the long flight durations and low geomagnetic cutoff in long-duration balloon (LDB) flights over Antarctica, a completely new version of the instrument, BESS-Polar, was developed. The BESS-Polar magnet has half the areal density of its predecessor, achieved by use of improved Al stabilized NbTi superconducting wire strengthened by cold-working and alloying the Al with Ni filaments. Cryogen lifetime was increased to >25 days by reducing heat transmission to the low-temperature components. In addition, the outer pressure vessel was eliminated, the aerogel Cherenkov counter (ACC) (Chap. 18, "Cherenkov Counters") was moved below the magnet, and a middle TOF layer was added inside the magnet bore. As a result only ~4.5 g/cm^2 is encountered by triggering particles, compared to ~18 g/cm^2 in the previous BESS instrument, lowering the effective energy threshold. BESS-Polar has a mass of 2,000 kg and a geometry factor of 0.3 m^2 sr. BESS-Polar I flew for 8.5 days in 2004, recording 9×10^8 cosmic-ray events. BESS-Polar II flew in 2007–2008, operating at float altitude for 24.5 days with the magnet energized and recording over 4.7×10^9 events.

All versions of BESS use similar instrument configurations with detail changes reflecting the evolution of the instruments and flight-specific requirements. *Figure 7* shows a schematic cross-sectional view of the BESS-Polar II instrument as an example. A central JET-type drift-chamber tracking system and inner drift chambers, giving 52 trajectory points in the bending direction with a resolution of about 130 μm, are located inside

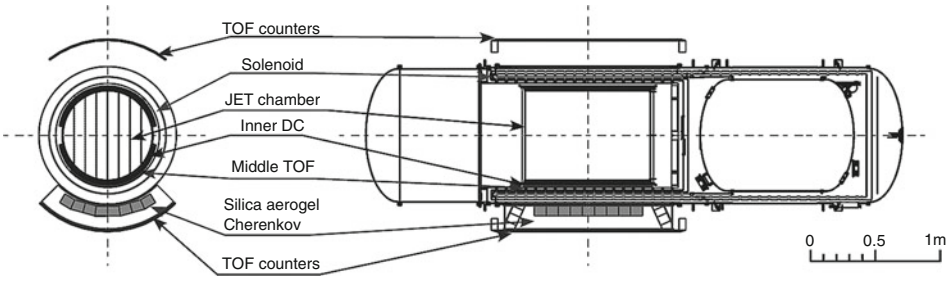


Fig. 7
Cross-sectional view of the BESS-Polar II instrument

the warm bore of the solenoid. R is determined by fitting the curvature of the particle track. Charge sign is determined by the direction of curvature. The horizontal cylindrical configuration of the BESS instrument allows a full opening angle of ~90° with a resulting acceptance of 0.3 m^2 sr. The thin solenoid magnet allows the incoming cosmic rays to penetrate the spectrometer with minimum interactions. Since the magnetic field is uniform inside the solenoid, the deflection distribution for particles of a given R is very narrow and the R resolution is nearly constant for all trajectories within the instrument's geometric acceptance. The superconducting magnet can operate at 1 T and for both BESS-Polar flights a conservative 0.8 T was used. A maximum detectable rigidity (MDR) of 200 GV was achieved in the original BESS instrument and ~280 GV in BESS-Polar. For the BESS-TeV flights in 2001 and 2002, outer drift chambers were added to raise the MDR to 1.4 TV.

Arrays of TOF scintillators are located at the top (UTOF) and bottom (LTOF) of the instrument, with a TOF resolution in BESS-Polar II of ~120 ps over a 1.48 m flight path. In BESS-Polar, developed to extend measurements to as low an energy as possible in long-duration Antarctic flights, a middle TOF scintillator array (MTOF) with resolution ~280–380 ps is located inside the magnet bore below the lower IDC. The TOF scintillators trigger readout of particle events and measure Z and β. The p is determined from R and Z and, in turn, m is determined from p and β as illustrated in *Fig. 8*. BESS-Polar separates antiprotons by mass from negatively charged background particles, mainly muons and electrons, up to E_k of about 1.5 GeV. At higher β, an aerogel Cherenkov ACC identifies low-m high-β background particles with a rejection power >6,000. Additional background rejection is supplied by multiple measurements of dE/dx from the JET. Antiprotons can be identified by mass and charge sign from 0.1 to 4 GeV. Elemental spectra can be measured to >100 GeV.

Where BESS is designed to maximize the geometric acceptance of the spectrometer to measure rare species in balloon flights, PAMELA was conceived to accomplish the same goals as BESS by taking advantage of a long exposure in space flying on a Russian Earth-observing satellite. PAMELA was launched from Baikonur cosmodrome in June 2006. The instrument has been in stable operation since shortly after launch and has made ground-breaking measurements of high-energy positrons and antiprotons, as well as measuring element spectra and SEPs, and carrying out a search for cosmic antimatter.

PAMELA is built around a permanent-magnet-based magnetic rigidity spectrometer using silicon-strip detectors. A plastic-scintillator TOF system measures the charge of incident particles, determines the direction of flight, and provides the instrument trigger. A silicon–tungsten

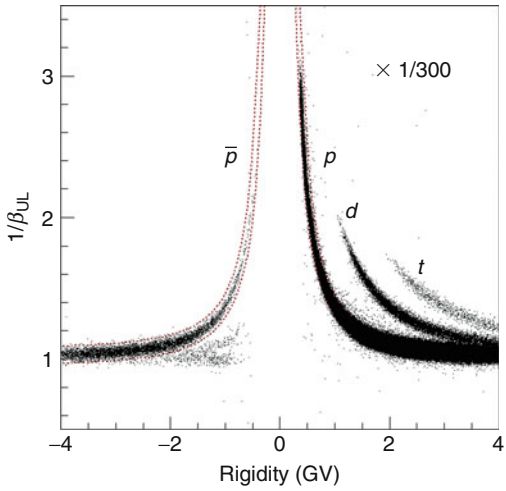


Fig. 8
BESS-Polar II Particle ID plot for singly charged particles

imaging calorimeter measures the energy of incident particles, particularly electrons, and discriminates between electrons and hadrons by examining the shower topology. A plastic-scintillator anticoincidence system protects against particles arising from interactions in the material of the instrument. A plastic-scintillator penetration detector and a neutron detector below the calorimeter aid the selection of high-energy electrons. The instrument has an overall mass of 470 kg and a geometric acceptance of 21.5 cm^2 sr. PAMELA identifies antiprotons 60 MeV–180 GeV, positrons 50 MeV–270 GeV, electrons 50 MeV–400 GeV, protons 80 MeV–700 GeV, and nuclei to O up to 100 GeV. *Figure 9* shows the PAMELA instrument.

The PAMELA spectrometer is based on a magnet composed of 5 layers of NdFeB blocks, 12 in each layer, giving a mean field in the tracking region of 0.43 T. Tracking is provided by six layers of 300 μm double-sided silicon-strip detectors (SSDs) with a readout pitch in the bending direction of 50 μm and 67 μm in the non-bending direction. The tracking resolution is about 3 μm in the bending plane and 11.5 μm in the non-bending plane. The tracking layers are located above and below the magnet and between each of the magnet layers. The MDR is ~1 TV.

The PAMELA TOF system is made up of three layers of plastic scintillator: one (S1) at the top of the instrument and layers just above (S2) and below (S3) the magnet. The full flight distance from S1 to S3 is 77.3 cm and the timing resolution is ~250 ps. This is adequate to separate electrons (positrons) from antiprotons (protons) to ~1 GeV/c and to reject upward-moving particles with about 60 standard deviations. Timing resolution improves for higher charges.

The SiW imaging calorimeter on PAMELA is made up of 44 layers of 380 μm thick single-sided SSDs and 22 W plates each 0.26 cm thick (0.74 X_0). The total depth of the calorimeter is 16.3 X_0 (0.6 nuclear interaction lengths). SSD layers are paired with strips on adjacent layers oriented orthogonal to one another, giving x, y positions. The W plates are interleaved between the pairs of SSD layers. The calorimeter has an energy resolution of 5.5% for electromagnetic showers. The primary goal of the calorimeter is to separate positrons and antiprotons from more abundant background particles of the same charge. Positrons must be separated from protons and antiprotons from electrons. The longitudinal and transverse segmentation of the calorimeter and the measurements of dE/dx by the SSDs identify electromagnetic showers.

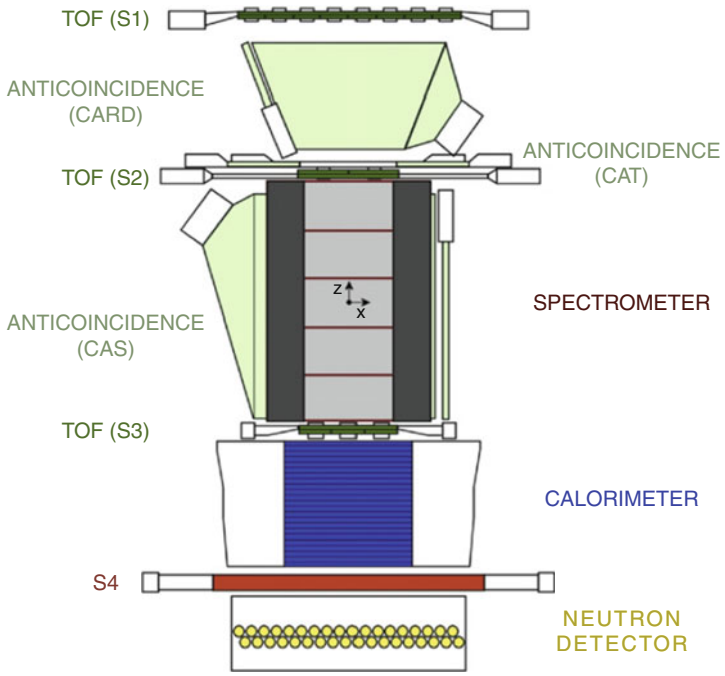


Fig. 9
Schematic view of the PAMELA spectrometer (Picozza et al. 2007)

This allows proton background to the positron measurements and electron background to the antiproton measurements to be rejected by a factor of about 10^5. In both cases the efficiency for selecting the species of interest is about 90%.

AMS-02 is a large, 6,900 kg, instrument designed to exploit the full capabilities of a space-based magnetic rigidity spectrometer. AMS-02 was installed on the ISS by the Space Shuttle Endeavour. As finally configured, the instrument uses the AMS-01 permanent magnet to allow for long-term operation on the ISS without the limitations of liquid-helium consumption. To preserve the performance of the instrument with the weaker permanent magnet, the tracking system was reconfigured as discussed below. AMS-02 is designed to measure cosmic-ray particles and nuclei, examine dark-matter signatures with positron and antiproton measurements, and search for primordial antimatter. It incorporates a TOF, a ring-imaging Cherenkov detector (RICH) using both NaF and silica-aerogel radiators, an electromagnetic calorimeter (ECAL), and a transition radiation detector (TRD). An anticoincidence along the sides of the magnet bore rejects particles outside the instrument acceptance.

The AMS-02 spectrometer employs a NdFeB permanent magnet with a 1 m diameter vertical bore, 1 m height, and 0.15 T field. The magnet is made up of ~6,000 NdFeB blocks glued together in the form of a ring dipole. The ring dipole has very little net dipole moment and so will exert negligible torque on the ISS. The tracking system has nine layers of double-sided SSDs. Position resolution is ~10 μm in the bending direction and ~30 μm in the non-bending direction. In the original configuration for the superconducting magnet, there were eight tracking layers, one at the top of the magnet, one at the bottom, and six distributed in three pairs inside the magnet bore. For the flight configuration, parts of the tracking plane just above the magnet

were moved to just above the ECAL and the plane that had been at the bottom of the magnet was moved to the top of the instrument, above the TRD, giving nine tracking layers. For particles traversing all nine tracking layers, the MDR is 2.14 TV but with a reduced geometric acceptance.

The AMS-02 ECAL uses Pb absorbers interspersed uniformly with layers of 1 mm scintillating optical fibers read out at one end by multi-anode PMTs. The fibers are laid in groups or superlayers of 10 fiber layers and 11 Pb foils. Each successive superlayer orientation is rotated 90° to give topological information on electromagnetic showers, improving rejection of background protons. There are five superlayers in the bending direction and four in the non-bending. The ECAL is ~17 X_0 in total and has an energy resolution of ~2.5% and angular resolution of ~1°. The active area is 648 mm × 648 mm and the detector is 166 mm thick. Expected proton rejection above 200 GeV is on the order of 10^4.

The TOF uses four layers of plastic scintillator, two located above the magnet and two below with a timing resolution of 160 ps. The RICH uses a proximity-focused design with a center 34 cm × 34 cm region of 5 mm thick NaF (n = 1.33) surrounded by an outer ring of silica aerogel (n = 1.05). Cherenkov light is imaged by an array of 10,880 photomultipliers. A 64 cm × 64 cm hole in the middle of the readout plane clears the acceptance of the ECAL. The central NaF radiator provides better resolution at lower energies and its larger refraction angle allows the Cherenkov cones produced to strike the photodetectors outside the hole. The RICH is designed to provide charge resolution up to Fe and a velocity resolution of 0.1%.

The AMS-02 TRD is designed to help distinguish between positrons and protons. The TRD contains 5,248 straw tubes, 6 mm in diameter, arranged in 20 layers alternating with 20 mm layers of polyethylene/polypropylene fleece radiator. The tubes are filled with a 80% : 20% mixture of Xe and CO_2 at 1.0 atm. The gas is cleaned by a recirculating system. The measured leak rate would give an operational life of over 24 years. Measured using 400 GeV protons, the TRD has a rejection power of 10^2 with an electron selection efficiency of 90%. Because this is independent of the ECAL rejection power, the result is net a hadron rejection factor of 10^6.

3.4 Calorimeters

In order to measure GCR protons, helium, and electrons above energies measurable by magnetic rigidity spectrometers (about 1 TeV) ionization calorimetery is required. This is a standard technique for measuring electron and hadron energies at accelerators, but the constraints of balloon or space flight dictate designs that differ considerably from accelerator calorimeters. High-energy GCR calorimetry was pioneered in the Proton satellites (Akimov et al. 1970) and has been used by PAMELA (Picozza et al. 2007), ATIC (Advanced Thin Ionization Calorimeter) (Guzik et al. 2004), CREAM (Cosmic-Ray Energetics And Mass) (Ahn et al. 2007), Fermi-LAT (Atwood et al. 2009), and AMS (Kounine et al. 2010). In this section, we will discuss the ATIC and CREAM balloon instruments. Both instruments were designed to measure the spectra of hadronic GCRs to energies approaching the knee. ATIC has also contributed to the measurement of high-energy CR electron spectra. There are considerable similarities between the two approaches, including the inclusion of an interaction target to boost sensitivity to hadrons. However, they differ greatly in the calorimetry approach. We also briefly discuss the new CALET (CALorimetric Electron Telescope) instrument under construction for flight on the ISS (Toji et al. 2011).

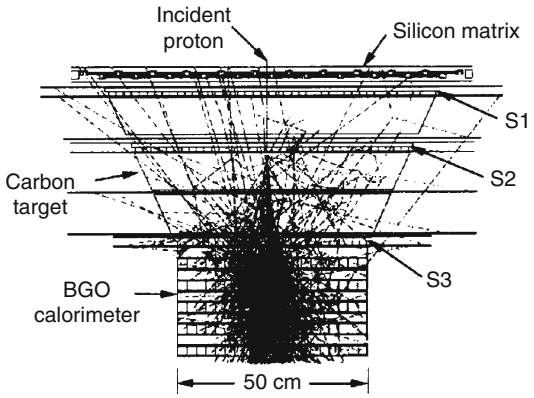


Fig. 10
Simulated proton event in the ATIC flight configuration (Guzik et al. 2004)

High-energy particles interact in material to produce particle cascades (showers) that deposit energy by ionization. The initial interaction probability and subsequent cascade development are characterized by an interaction length (λ_I) for hadrons and radiation length (X_0) for electrons and photons (and the electromagnetic component of hadronic cascades). In dense materials, λ_I is much larger than X_0. An ideal calorimeter fully contains the shower development of the incident particles. For space flight, however, the mass needed for such a full-containment calorimeter for hadrons is prohibitive and optimizing geometric acceptance requires a "thin" calorimeter that contains most of the development of electromagnetic showers but not hadronic showers. Thin calorimeters take advantage of the characteristic development of electromagnetic showers which peak quickly and drop to ~1% of peak after ~30 X_0. Energy resolutions of <10% at 1 TeV are possible for electrons and photons, improving to <3% at 10 TeV. For hadrons resolution is limited by fluctuations in shower energy "leakage" that depend on absorber depth.

ATIC, flown in Antarctic LDB flights in 2000–2001, 2002–2003, and 2007–2008 for a total of 51 days, uses a fully active, bismuth-germanate (BGO) calorimeter to measure the energy deposited by cascades formed by particles interacting in a thick carbon target. A highly segmented silicon matrix, located above the target, resolves the charge of incident particles with little contamination by backsplash albedo from the cascade. Trajectory reconstruction is based on the cascade profile in the BGO calorimeter, plus information from three pairs of scintillator hodoscope layers in the target section. ATIC weighs about 1,500 kg and has a geometric factor of 0.24 m^2 sr. *Figure 10* illustrates a simulated proton event with the ATIC flight configuration.

The target section of ATIC is composed of scintillator hodoscopes with 2 cm segmentation interleaved with carbon target elements. A two-layer scintillator-strip hodoscope is located above a 10 cm thick carbon target layer. This is followed by a second hodoscope, a 20 cm thick target layer, and a third hodoscope. The target section contains about 0.75 λ_I and 1.5 X_0. Together with shower axis measurements from the calorimeter, the hodoscopes can resolve the impact point of incident high-energy particles to better than 1 cm.

The ATIC calorimeter is made of ten layers of BGO bars, each 2.5 cm by 2.5 cm in cross section and 25 cm in length read out by PMTs. Sets of 40 such bars are arranged in 50 cm × 50 cm layers. The full calorimeter depth is about 22 X_0 and 1.1 λ_I. The bars in each successive layer are rotated 90° to resolve the shower in three dimensions and allow its axis to be reconstructed.

CREAM has flown six times over Antarctica since its first flight in 2004–2005, accumulating about 161 days of exposure. The instrumentation suite of CREAM has changed for different flights, although a calorimeter is used in all versions. The charges of incident nuclei are measured using a two-layer plastic-scintillator timing charge detector (TCD), a two-layer silicon pixel detector (SCD), and a silica-aerogel Cherenkov camera (CHERCAM – CREAM-III, IV, V, VI). Depending on the energy and species of the incident particle, its energy is measured by a TRD (CREAM-I) and a tungsten-scintillating-optical-fiber calorimeter (all versions). The geometric acceptance of the TRD is ~1.3 m^2 sr and the effective geometric acceptance (including interactions) for the calorimeter is about ~0.3 m^2 sr for protons and greater for higher-Z nuclei. A new TRD has been developed for CREAM-VII to enable improved measurements of secondary-to-primary ratios. Two 9.5 in. thick trapezoidal densified graphite targets are located just above the calorimeter giving a total of 0.45 λ_I. A Cherenkov detector using an acrylic radiator is located in the upper section of the instrument to veto non-relativistic particles and to provide rapid discrimination of higher-charge nuclei. A scintillating-optical-fiber detector between the target and calorimeter provides a timing reference for the TCD and flags interacting events for triggering.

The CREAM SCD uses an array of 2.12 cm^2 380 μm thick silicon-diode pixels. Each layer has 156 sensors with 16 pixels each and the double-layer SCD has a total of 4,992 pixels. The sensors are slightly tilted and overlap their neighbors so that each layer achieves complete areal coverage. The CHERCAM is a proximity-focused detector using a 1 cm thick silica-aerogel radiator with an index of refraction of 1.07. The detector is designed specifically for charge measurements. The detector is read out by an array of 1,600 PMTs.

CREAM utilizes a sampling calorimeter made of 20 layers of tungsten plates, each 50 cm × 50 cm × 0.35 cm, interleaved with layers of 0.5 mm thick, 1 cm wide scintillating-optical-fiber ribbons. These are coupled to 73 pixel hybrid photodiodes (HPDs), each reading out 64 signals. The fiber ribbons are composed of 19 fibers. Each ribbon is coupled to a light guide that, in turn, is coupled to 48 clear fibers. These are subdivided into groups of 42, 5, and 1 fiber and each group is coupled to an HPD pixel to provide the wide-dynamic-range readout needed by the calorimeter. The total depth of the calorimeter is 20 X_0. Successive fiber layers are rotated by 90°, giving 10 x, y measurements of particle tracks. The longitudinal segmentation of the calorimeter is approximately 1 X_0 and the lateral segmentation is about 1 Molière radius. This allows both the axial and transverse development of the shower to be accurately determined. The resolution of the projected track at the SCD is smaller than an SCD pixel.

CALET is a new mission to measure the high-energy spectra of electrons, nuclei, and γ-rays selected by JAXA for a launch in 2013 to the Japanese Experiment Module Exposed Facility (JEM-EF) on the ISS. The calorimeter-based CALET main telescope has a FOV of ~45° from the zenith and a geometric acceptance of 0.12 m^2 sr. The calorimeter is divided into an imaging calorimeter (IMC) section that provides tracking and accurately determines the starting point of showers, and a total-absorption-calorimeter (TASC) section that measures total particle energy. The IMC contains ~3 X_0 of tungsten interspersed between eight x, y layers of scintillating optical fibers read out by multi-anode PMTs. Most electrons and photons will initiate showers in the IMC, which measures the starting point of the shower and its development until it enters the TASC. The TASC is a stack of lead-tungstate (PWO) crystals arranged in x, y layers to track the axis of the shower. Each crystal is read out by two photodiodes plus an avalanche photodiode. The TASC has a total thickness of 27 X_0. A charge detector subsystem at the top of the telescope measures the charge of incident particles and functions as an anticoincidence detector for γ-ray measurements.

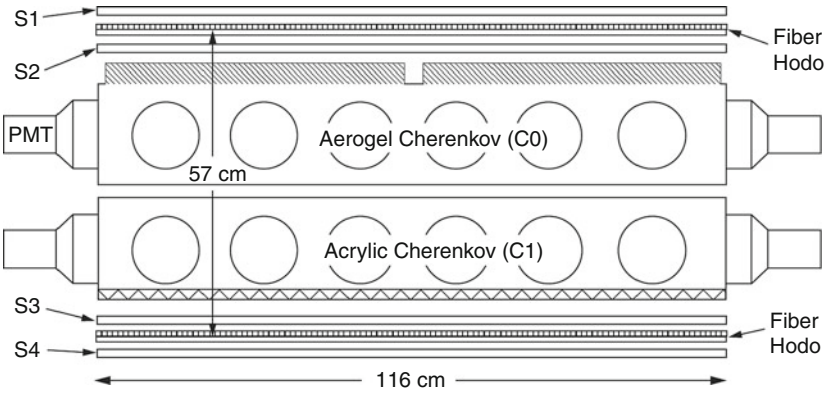


Fig. 11
Schematic view of the TIGER balloon payload (Rauch et al. 2009)

3.5 Large-Area Composition Experiments

One of the main challenges faced by designers of instruments to measure the GCR composition from balloons is to provide sufficient geometric acceptance to measure rare species during a limited exposure. Two approaches to this problem, incorporated into the large TIGER (Trans-Iron Galactic Element Recorder) and TRACER (Transition-Radiation Array for Cosmic Energetic Radiation) instruments, are discussed in this section. Both instruments were designed to provide exposures approaching 100 m^2 sr × days.

TIGER (Rauch et al. 2009) was designed to measure the composition of GCRs heavier than Fe with single-element resolution. TIGER flew twice over Antarctica, 32 days in 2001–2002 and 18 days in 2003–2004. From these flights, TIGER produced the first well-resolved elemental abundances of ^{31}Ga, ^{32}Ge, and ^{34}Se and searched for features in the energy spectrum of Fe from 2.5 to 10 GeV that might indicate the influence of a microquasar. The instrument was designed to limit the material encountered by traversing nuclei to 4.2 g/cm^2 to reduce interaction losses. TIGER is made up of four planes of plastic scintillator to measure charge, Cherenkov detectors with acrylic (n = 1.5) and silica-serogel radiators (n = 1.04) to measure charge and velocity, and a scintillating-optical-fiber hodoscope to measure particle trajectories through the atmosphere and instrument. The active area of TIGER is 114 cm × 114 cm with a full height of 58.3 cm, resulting in a geometric factor of ~1.8 m^2 sr. The instrument weighed 545 kg. ❯ *Figure 11* shows a cross section of the TIGER balloon payload.

Two of the scintillators were located near the top of the instrument and two near the bottom. Particle charge identification used the dE/dx–Cherenkov method up to the threshold of the aerogel Cherenkov. Above that threshold it used the Cherenkov–Cherenkov method. The fiber hodoscope uses a coded readout method to provide good tracking resolution with relatively few PMTs. Each of the two hodoscope planes consisted of two perpendicular layers of 1 mm square scintillating optical fibers. The fibers were formatted into tabs of six or seven fibers, so the effective segmentation of the hodoscope is ~6–7 mm. The fibers at one end of each layer were grouped into 14 "segments" with 14 adjacent tabs forming one segment coupled to one PMT for "coarse" readout. For "fine" readout, the fibers at the other end were grouped with the first tab of each segment going to one photomultiplier, the second tab of each segment going to

another PMT, etc. This coding scheme allowed TIGER to identify uniquely which of the 196 tabs in a layer was hit, using only 28 PMTs per hodoscope. The scintillators use wavelength-shifter readout in which the blue scintillation light is collected by blue-green wavelength-shifter bars on the edges of the scintillator. The light is upshifted by the fluorescent dye in the wavelength shifter, re-emitted, and piped to PMTs at each end. This allowed each large scintillator to be read out using only 8 PMTs. The Cherenkov detectors were configured as light-integrating volumes each viewed by 24 12.7 cm PMTs. The TIGER instrument could not be recovered following its 2003 flight. The collaboration is presently building a new instrument, Super-TIGER, with 6.4 times the effective geometric factor of TIGER, to measure elements to ^{42}Mo, explore the abundances of elements to ^{56}Ba, and measure the energy spectra of the more abundant elements from ^{10}Ne to ^{19}Cu between 0.8 and 10 GeV/nucleon to test the hypothesis that microquasars or other sources could superpose spectral features. The first flight of Super-TIGER is planned for December 2012.

TRACER (Müller et al. 2004; Obermeier 2011) was designed to measure the energy spectra of GCR elements to energies approaching the knee and was based on experience with the CRN (Cosmic Ray Nuclei) experiment that flew on Spacelab 2 in 1985 and used a combination of scintillators, gas Cherenkov detectors, and a TRD. TRACER uses the same transition radiation (TR) radiator material as CRN, taking advantage of the calibrations carried out for that instrument, but uses proportional tubes rather than thin-window multi-wire proportional counters. This allows the pressure vessel to be eliminated. TRACER has had two long-duration balloon flights, 14 days from Antarctica in 2003–2004 and 4.5 days from Kiruna (Sweden) in 2006. The TRACER design goal was to build the largest detector with good energy and charge resolution that could be accommodated on a balloon platform. The technique chosen was to subdivide the energy range into three regions (see ❯ *Fig. 12*), a Cherenkov range from 0.325 GeV/nucleon up to a few GeV/nucleon, relativistic rise in dE/dx from 10–500 GeV/nucleon and a TRD above 700 GeV/nucleon. Overall, the energy measurements span four orders of magnitude. TRACER is made up of two Cherenkov detectors with acrylic radiators and two scintillators. The scintillator/Cherenkov detector pairs are located above and below the TRD. TRACER has an active area of 2 m × 2 m and a height of 1.2 m, giving a geometry factor of 5 m^2 sr. The instrument weighed just less than 2,000 kg. Both Cherenkov detectors and scintillators use wavelength-shifter readout, similar to TIGER. The charges of incident nuclei are identified using the dE/dx–Cherenkov technique. The Cherenkov detectors also identify low-energy particles whose ionization energy deposit in the TRD tubes might mimic the signals of high-energy nuclei. The TRD was designed to measure the Lorentz factor $\gamma = E/(mc^2)$, of nuclei with $Z > 3$. The TRD is made up of 1,584 thin-walled gas proportional tubes, 2 cm in diameter and 2 m long, using Xe/CH_4 gas to detect X-ray photons. The tubes are arranged in double layers with each successive double layer rotated by 90°. Fits to the path length of particles in the double layers of proportional tubes give a position resolution of ~2 mm. The unique feature of the TRACER TRD is that the top four double layers only measure dE/dx. The remaining four double layers are placed below four plastic-fiber blankets and detect the X-ray TR photons produced in the blankets. The dE/dx in the gas falls with energy, reaching a minimum at $\gamma = 3.96$, and then begins a relativistic rise logarithmically with E. TR X-rays are generated beginning at about 7.35×10^{11} eV/nucleon and the signal rises rapidly with γ, reaching saturation in the $\gamma = 10^4$–10^5 range. The TR photons must be detected superimposed on the dE/dx signal and the dedicated dE/dx layers allow the TR signal to be clearly identified. The TR signal increases as Z^2 and the technique is best used for higher-charge nuclei since the TR signals from protons and He have large statistical fluctuations. TRACER has measured the energy spectra of elements from C to Fe to total energies of about 10^{14} eV and the ratio B/C to over 2 TeV/nucleon.

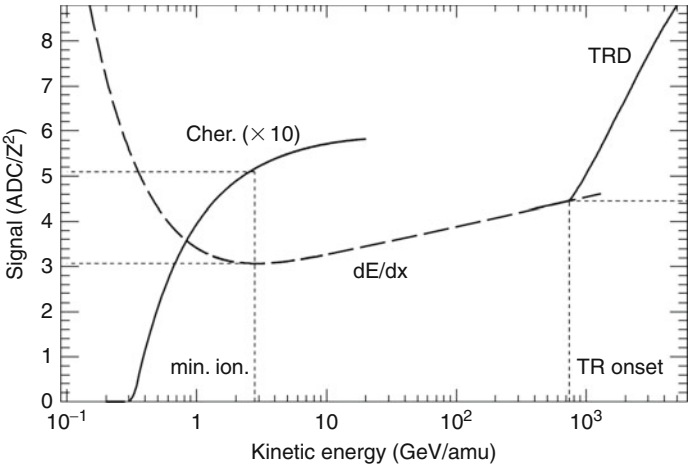


Fig. 12

Response functions, normalized by Z^2, of sub-detectors that are used for energy measurement. The natural normalization of the signals shown here is used throughout the analysis. Minimum ionizing energy (γ = 3.97) is at 3.07 in dE/dx signal and at 0.51 in the bottom Cherenkov signal per Z^2. TR onset (γ = 785) is at 4.45 in dE/dx and TR signal per Z^2. The normalization of the Cherenkov signal is multiplied by 10 for clarity (Obermeier 2011)

3.6 Indirect Measurements

At extremely high energies, the fluxes of cosmic particles are so small that the detectors needed are far too large for direct measurement techniques. These particles have usually been measured by large ground-based instruments as reviewed in Chap. 24, "Indirect Detection of Cosmic Rays" of this handbook. However, in some cases, even larger detectors can be realized using instruments that view the atmosphere or the surface of the Earth from balloon altitudes or space. Some of these indirect measurement instruments are reviewed below.

ANITA (Antarctic Impulsive Transient Antenna) (Nichol et al. 2011) is designed to measure the cosmogenic neutrino flux above about $10^{18.5}$ eV by detecting Askaryan effect radio emission from neutrino interactions in the Antarctic ice. At these energies the neutrino interaction cross sections are vanishingly small, with a mean interaction length in water of hundreds of kilometers. Cross sections are the same for neutrinos and antineutrinos, which are both referred to here as neutrinos. The predicted cosmogenic neutrino flux is also extremely small, on the order of one per square kilometer per week. Thus, a huge interaction target is needed. ANITA views the Antarctic ice from balloon altitude, typically 37 km, and observes an area of about 1.5 million km^2 and an interaction volume of 1–2 million km^3. The solid angle in which a neutrino can arrive and be detected by ANITA is a small fraction of a steradian. Taking this into consideration, the detection volumetric acceptance is hundreds to thousands of km^3 sr. Cosmogenic neutrinos are expected to contain approximately equal numbers of the three ν flavors. However, ANITA is mostly sensitive to ν_e because all of the energy of a ν_e charged-current interaction is converted into an observable shower with, on average, 80% of the incident ν_e energy

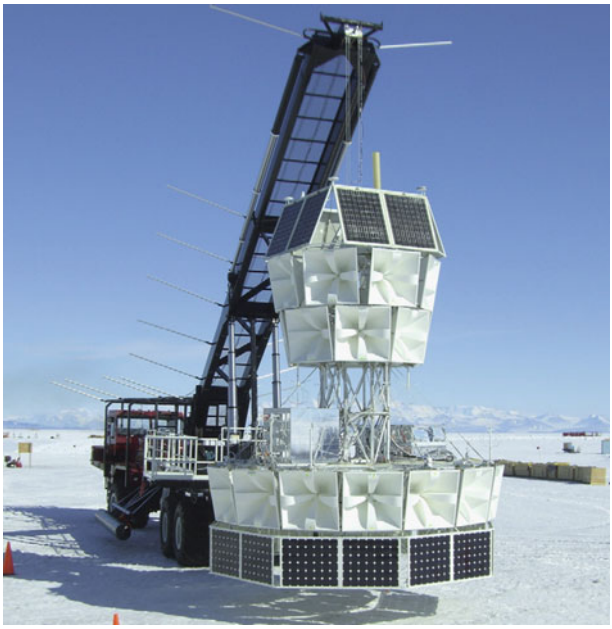


Fig. 13
Picture of the ANITA payload on the launch vehicle at Williams Field near McMurdo Station, Antarctica (Nichol et al. 2011)

going into an electron-induced shower and the remaining 20% into a hadronic shower. Coherent radio Cherenkov emission is produced by an electromagnetic shower in a dielectric medium because of charge asymmetry from positron annihilation in flight and Compton scattering that results in ~20% electron excess. An electromagnetic shower from a neutrino interaction in ice appears as a disc of particles several cm in diameter (Molière radius) and about 1 cm thick, propagating at the speed of light. This disc produces Cherenkov radiation that is coherent at wavelengths greater than the diameter of the disc and can be detected as a pulse at microwave frequencies. The duration of the pulse is ~100 ps and the mean frequency is 0.7 GHz. ANITA is designed to detect this radiation in the 200–1,200 MHz frequency band with dual polarization, broadband antenna clusters arranged with overlapping FOV, and a highly selective RF impulse trigger. ANITA uses two rings of quad-ridged horn antennas with 16 antennas in each ring (see *Fig. 13*). Pulses must be detected by at least four antennas in coincident upper ring/lower ring pairs. ANITA has flown twice, accumulating an exposure of 37.3 days. Thus far, no neutrino detection has been reported (Gorham et al. 2009).

ANITA has reported the detection of 16 UHECR air showers, 40% above 10^{19} eV (Hoover et al. 2010). The radio signal from these air showers results from geosynchrotron emission produced when the electrons and positrons in the electromagnetic particle cascade are bent in opposite directions by the geomagnetic field. The particles spiral around field lines and emit synchrotron radiation until they lose energy and leave the shower. The longitudinal scale of the shower disc in air is on the order of a meter, comparable to the wavelength of radio below a few hundred MHz, and the cascade produces forward-beamed synchrotron emission which is

at least partly coherent. The power of the radio impulse increases quadratically with particle energy and at the highest energies the pulse can be detected at large distances by reflection off the ice.

CREST (Cosmic Ray Electron Synchrotron Telescope) is designed to measure the spectrum of multi-TeV electrons by the detection of the X-ray synchrotron photons generated in the magnetic field of the Earth using a large balloon-borne instrument. As noted earlier, electrons in the TeV range are expected to result from local sources because the distant-source contribution is suppressed by synchrotron and inverse Compton losses during propagation. When the trajectories of high-energy cosmic-ray electrons are deflected by the geomagnetic field, they emit synchrotron radiation (Stephens and Balasubrahmanyan 1985). Multi-TeV electrons have Lorentz factors between 2×10^6 to 2×10^7, and by traversing a roughly 0.5 Gauss magnetic field, the characteristic synchrotron photon energies range from ~50 keV to tens of MeV. These photons are emitted in a very narrow cone, with an opening angle of approximately $1/\gamma$. The result is a line of photons many hundreds of meters long at balloon altitude, whose average energy is a strong function of the energy of the primary electron. Since CREST needs to intersect only a portion of the kilometers-long trail of photons generated by the high-energy electron, the method gives a much larger effective area than the physical size of the detector. The CREST instrument is composed of an array of 1,024 BaF_2 crystals each read out by a PMT. A layer of plastic scintillator on the top of the crystal vetos charged particles. The first CREST Antarctic flight is planned for December 2011.

UHECRs have been detected to energies greater than 10^{20} eV, but their flux is minuscule, on the order of one per km^2 per century, and instruments with exposures approaching a million km^2 sr yr are needed to exploit the potential field of charged-particle astronomy. Presently UHECRs are measured using large ground detector arrays or ground-based fluorescence cameras with areas of thousands of km^2, see Chap. 24, "Indirect Detection of Cosmic Rays." John Linsley suggested in 1982 that viewing the atmosphere from space on dark nights would enable much larger detection volumes. Air showers consist of huge numbers of charged particles produced by a cascade process in which the energy of the primary UHECRs is dissipated in the atmosphere. This, in turn, results in the emission of UV light from fluorescence of excited atmospheric nitrogen as well as scattered Cherenkov light. The fluorescence light from a cosmic-ray-induced air shower appears as a luminous disk, a few meters in depth with a radius less than a kilometer, moving through the atmosphere at the speed of light. The atmosphere can thus be used as a calorimeter if the spatial and temporal development of the shower is viewed by a fast, highly pixelized camera. The UV light, limited to the 330–400 nm range by ozone absorption, is isotropic and the camera can view the shower from any direction except almost directly toward the camera (where it acts as a Cherenkov detector). This means that a single camera can view a large part of the sky, limited only by atmospheric absorption of the UV light and by physical interference and can provide a very large detecting volume. By placing the camera in space, the detecting volume can be vastly enlarged.

The fluorescence technique is the basis of the Fly's Eye and HiRes instruments and is employed by the Pierre Auger Array and the Telescope Array. Both monocular and stereo implementations are possible. Each fluorescence camera images the projection of the shower onto a plane normal to the viewing direction. The angle of the shower relative to the viewing plane is resolvable using differential timing, and in monocular operation, precision measurements of the arrival times of UV photons from different parts of the shower track must be used. Resolving distance is more difficult and requires precise measurements of pixel crossing times or identification of the intersection point of the shower axis and the ground by viewing the Cherenkov

light spot. Stereo observation completely resolves both of these ambiguities. In stereo, fast timing provides supplementary information to reduce systematics and improve the resolution of the arrival direction of the UHECRs. The stereo view also confers the crucial advantage that differences in atmospheric absorption or scattering of the UV light can be determined by viewing the same shower through different regions of atmosphere, greatly reducing systematics.

Two space-based implementations of this concept have been extensively studied. JEM-EUSO (Japanese Exposure Module Extreme Universe Space Observatory) is a monocular instrument proposed for flight on the ISS (Takahashi et al. 2009). OWL (Orbiting Wide-angle Light Collectors) is a free-flying mission that would use a pair of co-orbiting satellites in near-equatorial orbit to provide a stereo view (Stecker et al. 2004). Both telescopes use large optical aperture (low f number) systems. JEM-EUSO uses a 2.5 m diameter Fresnel-lens-based refracting telescope and OWL uses a reflecting Schmidt camera with an entrance aperture of 3 m (see ❯ *Fig. 14*). Both incorporate a finely segmented focal-plane array with several hundred thousand pixels. Each focal-plane pixel corresponds to ~1 km^2 on the ground and the required angular resolution is 3 orders of magnitude above the diffraction limit. As an example of the potential of this technique, OWL would view about 10^6 km^2 of atmosphere with a full instantaneous aperture of 2×10^6 km^2 sr reached at 6×10^{19} eV. JEM-EUSO would have an effective instantaneous acceptance of about 4×10^5 km^2 sr when pointed toward nadir, increasing in tilted mode. Corrected for an estimated 50% reconstruction efficiency to compare to OWL, the acceptance is 2×10^5 km^2 sr. Viewing only the dark side of the Earth and conservatively accounting for moon, man-made light, and clouds, the average OWL acceptance would be 2×10^5 km^2 sr. This is many times greater than the largest ground arrays (7×10^3 km^2 sr for the Pierre Auger Observatory), although the advantage is reduced by the longer viewing times of ground-based

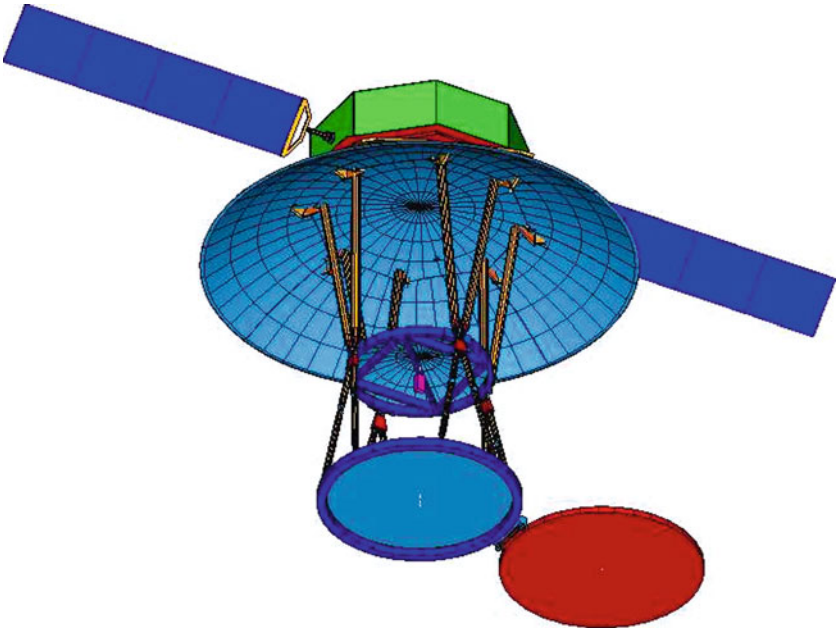


Fig. 14
One OWL telescope shown fully deployed with its light shield omitted

observatories compared to a typical space mission. OWL views about 10^{13} metric tons of atmosphere and so has the potential to detect neutrino events, with a detection rate depending on flux and interaction cross section.

4 Conclusion

Modern astronomy, astrophysics, and heliophysics require sophisticated instruments measuring high-energy gamma-ray and X-ray photons and particles ranging from electrons and positrons to heavy nuclei. For flight on balloons or satellites, instruments must be carefully designed to maximize their performance while minimizing use of limited resources such as weight and power. They must also be capable of autonomous operation and able to operate for long periods without maintenance. This challenge has led to a wide range of sophisticated and highly capable instruments that are providing the information needed to understand high-energy astrophysical engines ranging from the Sun to active galactic nuclei as well as the dynamics of particle populations in the galaxy and the heliosphere. The instruments described in this chapter are only a representative subset of what has been achieved. It is hoped that the information in this chapter and in this handbook may lead to the development of even more powerful scientific tools.

References

Ahn HS et al (2007) The cosmic ray energetics and mass (CREAM) instrument. Nucl Instr Methods A 579:1034–1053

Akimov VV et al (1970) Measurements of the primary cosmic ray spectra in the 10^{11}–10^{14} eV energy range from proton-1, 2, 3 satellites. Proceedings of the 11th international conference on cosmic rays, vol 1. Budapest, p 517

Atwood WB et al (2009) The large area telescope on the fermi gamma-ray space telescope mission. Astrophys J 697:1071–1102

Barthelmy SD et al (2005) The burst alert telescope (BAT) on the SWIFT midex mission. Space Sci Rev 120:143–164

Bieber JW (1999) Antiprotons at solar maximum. Phys Rev Lett 83(4):674–677

Fisk LA (1971) Solar modulation of galactic cosmic rays. J Geophys Res 76:221–226

Gehrels N et al (2004) The swift gamma-ray burst mission. Astrophys J 611:1005–1020

Gehrels N, Ramirez-Ruiz E, Fox DB (2009) Gamma-ray bursts in the swift era. Ann Rev Astron Astroph 47:567–617

Giacconi R et al (1962) Evidence for X rays from sources outside the solar system. Phys Rev Lett 9:439–443

Gorham PW et al (2009) New limits on the ultrahigh energy cosmic neutrino flux from the ANITA experim. Astropart Phys 32:10–41

Gruber DE et al (1996) The high energy X-ray timing experiment on XTE. Astron Astrophys Suppl Ser 120:641–644

Guzik TG et al (2004) The ATIC long duration balloon project. Adv Space Res 33:1763–1770

Hoover S et al (2010) Observation of ultrahigh-energy cosmic rays with the ANITA balloon-borne radio interferometer. Phys Rev Lett 105:151101

Jahoda K et al (2006) Calibration of the Rossi X-ray timing explorer proportional counter array. Astrophys J Suppl Ser 163:401–423

Koglin JE et al (2005) NuSTAR hard X-ray optics. Proc SPIE 59000:59000X-1

Kounine A et al (2010) Status of the AMS experiment. arXiv:1009.5349

Levine AM et al (1996) First results from the all-sky monitor on the Rossi X-ray timing explorer. Astrophys J 469:L33–L36

Mason GM et al (1998) The ultra-low-energy isotope spectrometer (ULEIS) for the ACE spacecraft. Space Sci Rev 86:409–448

Mitchell JW et al (2004) The BESS program. Nucl Phys B (Proc Suppl) 134:31–38

Müller D et al (2004) Transition radiation detectors in particle astrophysics. Nucl Instr Methods A 522:9

Nichol JR et al (2011) Radio detection of high-energy particles with the ANITA experiment. Nucl Instr Methods A 626–627:S30–S35

Obermeier A (2011) A direct measurement of cosmic rays to very high energies: implications for galactic propagation and sources. PhD thesis, Radboud University, Nijmegen, The Netherland, ISBN 978-90-9025962-8

Picozza P et al (2007) PAMELA a payload for antimatter matter exploration and light-nuclei astrophysics. Astropart Phys 27:296–315

Rauch BF et al (2009) Cosmic ray origin in OB associations and preferential acceleration of refractory elements: Evidence from abundances of elements $_{26}$Fe through $_{34}$Se. Astrophys J 697: 2083–2088

Stecker FW et al (2004) Observing the ultrahigh energy universe with OWL eyes. Nucl Phys B 136C:433–438

Stephens SA, Balasubrahmanyan VK (1985) High energy gamma ray observatories for the study of cosmic ray electrons above 10^{14} eV. Nucl Instr Methods A 241:257–264

Stone EC et al (1998a) The advanced composition explorer. Space Sci Rev 86:1–22

Stone EC et al (1998b) The solar isotope spectrometer for the advanced composition explorer. Space Sci Rev 86:357–408

Stone EC et al (1998c) The cosmic-ray isotope spectrometer for the advanced composition explorer. Space Sci Rev 86:285–356

Strong AW, Moskalenko IV, Ptuskin VS (2007) Cosmic-ray propagation and interactions in the galaxy. Ann Rev Nucl Part Sci 57:285–327

Takahashi Y et al (2009) The JEM-EUSO mission. New J Phys 11:065009

Torii S et al (2011) Calorimetric electron telescope mission. Search for dark matter and nearby sources. Nucl Instr Methods A 630:55–57

Yamamoto A et al (2011) Search for cosmic-ray antiproton origins and for cosmological antimatter with BESS. Adv Space Res (in press)

24 Indirect Detection of Cosmic Rays

Ralph Engel
Karlsruher Institut für Technologie, Karlsruhe, Germany

1 Introduction 594

2 Phenomenology of Extensive Air Showers 596
2.1 Photon-Induced Showers 599
2.2 Hadron-Induced Showers 600
2.3 Neutrino-Induced Showers 603

3 Measurement Techniques and Observables 604
3.1 Particle Detector Arrays 604
3.2 Atmospheric Cherenkov Light Detectors 607
3.3 Fluorescence Telescopes 610
3.4 Radio Signal Detection 614

4 Examples of Air Shower Detectors 616
4.1 KASCADE 616
4.2 The Milagro Gamma-Ray Observatory 618
4.3 Tunka 620
4.4 H.E.S.S. 620
4.5 The Pierre Auger Observatory 621

5 Open Problems and Future Experiments 623

6 Conclusion 625

Acknowledgments 625

References 626

Further Reading 632

C. Grupen, I. Buvat (eds.), *Handbook of Particle Detection and Imaging*,
DOI 10.1007/978-3-642-13271-1_24, © Springer-Verlag Berlin Heidelberg 2012

Abstract: Cosmic rays entering the Earth's atmosphere interact with the nuclei of air and produce cascades of secondary particles. At high energy ($E \gtrsim 10^{14}$ eV), these particle cascades are called extensive air showers and provide a means of measuring flux, arrival direction distribution, and elemental composition of cosmic rays. In this article, an introduction to the phenomenology of extensive air showers and parametrizations for their description is given. The concepts and key ideas of detecting extensive air showers are presented and methods for reconstructing the properties of the primary particles are discussed. Several representative experiments are reviewed to illustrate practical implementations of air shower detection concepts. The article concludes with a discussion of open problems and future detection techniques that are currently under development.

1 Introduction

At energies below ~10^{15} eV, cosmic rays can be measured directly with balloon- or satellite-borne detectors, see Chap. 23, "Astrophysics and Space Instrumentation." However, the steeply falling flux of particles makes it very difficult to extend direct measurements even by one decade in energy. Fortunately, starting at an energy of about 10^{14} eV, the indirect detection of cosmic rays becomes feasible. Instead of detecting the cosmic particles at the top of the atmosphere, the cascades of secondary particles they produce after interacting with nuclei in the atmosphere are measured. Such cascades of secondary particles are produced by cosmic rays of all energies, but those with $E \gtrsim 10^{14}$ eV contain so many particles that they reach the surface of the Earth as particle showers. Extending over areas from 10 m to several km, they are called extensive air showers.

Extensive air showers were discovered in the late 1930s, realizing that the high coincidence rate between particle detectors separated by up to 300 m could only be explained by many particles arriving as group together at the surface of the Earth. The discovery is attributed to Pierre Auger (Auger et al. 1939) but also several other scientists came independently to similar conclusions, see Kolhörster et al. (1938) and discussion in Linsley (1998). Already the first estimates showed that the primary particles of these showers have an energy of ~10^{15} eV, beyond the reach of even modern accelerators. As we know now, the energy spectrum of cosmic rays extends to energies beyond 10^{20} eV. The first air shower of about 10^{20} eV has been detected in 1962 (Linsley 1963). However, with an expected rate of much less than one particle per km^2 and century, only a few particles of such high energy have been recorded so far.

A compilation of measurements of the total flux of cosmic rays (all-particle energy spectrum) is shown in *Fig. 1*. The measured flux has been scaled by $E^{2.5}$ to make the characteristic features of the energy spectrum clearly visible. A discussion of these features is beyond the scope of this article. Here we will give only a very brief introduction, the interested reader is referred to the reviews (Haungs et al. 2003; Blümer et al. 2009; Beatty and Westerhoff 2009; Nagano 2009; Kotera and Olinto 2011; Letessier-Selvon and Stanev 2011). The break in the power law from $\mathrm{d}N/\mathrm{d}E \propto E^{-2.7}$ to about $E^{-3.1}$ at $E \approx 3 \times 10^{15}$ eV is called the *knee*. It is assumed that, at least up to this energy, cosmic rays are accelerated in galactic objects such as expanding supernova remnants (Blandford and Eichler 1987; Hillas 2005). However, the origin of the knee is not yet understood. It could be related to an increased leakage of cosmic rays from the

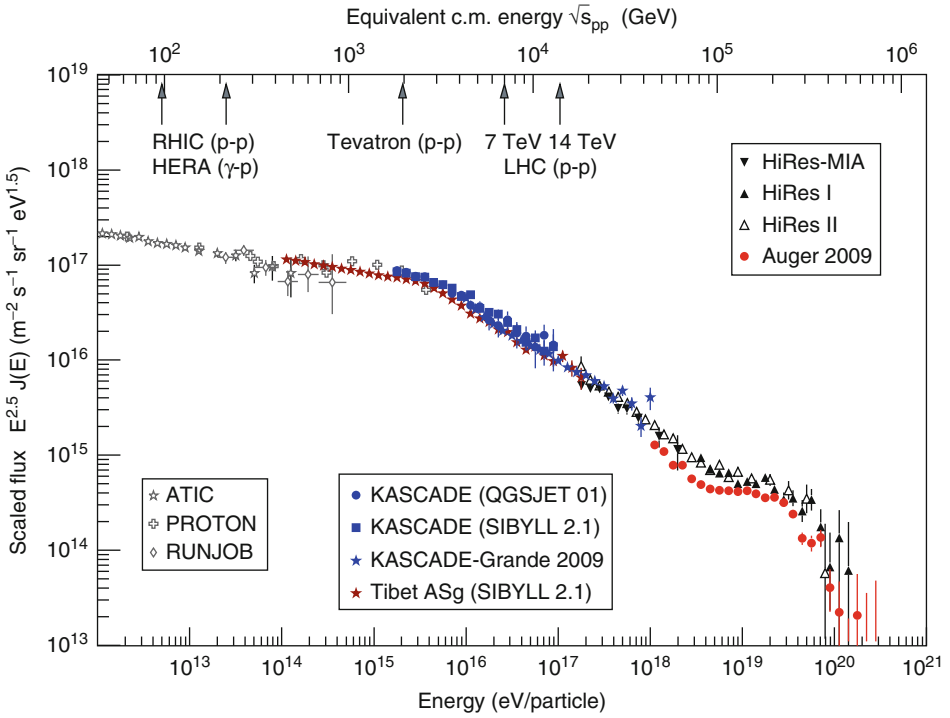


Fig. 1

All-particle flux of cosmic rays arriving at Earth, scaled by $E^{2.5}$. The equivalent center-of-mass energy of the collision with air, for protons as cosmic ray particles, is given on the upper horizontal axis. Direct measurements are shown from the balloon experiments ATIC and RUNJOB and the PROTON satellites. All high-energy data are based on indirect measurements. See Blümer et al. (2009) for refs. to the data

Galaxy, a change from one source population to another one, or even an indication for new particle physics (Hörandel 2004). In the energy range of the *ankle*, 10^{18}–10^{19} eV, the spectrum becomes harder again. The ankle is typically interpreted as the imprint of a change from galactic sources to extragalactic ones. This seems to be natural since protons of $10^{18.5}$ eV or higher energy are not magnetically confined in the Galaxy if the magnetic field strength is of the order of $B \sim 3\,\mu$G. Alternatively, the transition to extragalactic sources could take place at lower energy, e.g., $10^{17.5}$ eV, and the ankle would then be the signature of energy loss through e^+e^- pair production in the photon field of the microwave background (Berezinsky et al. 2006). The recently found flux suppression at energies above 7×10^{19} eV seems to confirm the predictions of Greisen (Greisen 1966) and Zatsepin and Kuzmin (Zatsepin and Kuzmin 1966) (GZK cutoff), who calculated the energy loss of particles in the microwave background due to pion production. However, the flux suppression could also be related to the maximum injection energy of the sources and, by chance, appear to be similar to that expected from the GZK cutoff (Allard et al. 2008).

All the flux measurements at high energy are based on indirect detection methods. Large effective areas can be reached since only a sparse coverage of the corresponding area with particle detectors is needed to detect air showers by coincidence or other means. For example, in the KASCADE detector array, 1.3% of the $200 \times 200\,\mathrm{m}^2$ area is actually covered by particle detectors (Antoni et al. 2003). Aiming at showers of ultra-high energy, even a coverage of only 5×10^{-4}% is sufficient (Abraham et al. 2004).

In this article we will discuss indirect detection methods used in cosmic ray physics. After reviewing features of extensive air showers that are of relevance to the different detection techniques in Sect. 2, we will introduce the concepts of the most important detection techniques in Sect. 3. The presentation will be focused on the general aspects, applicability ranges, and performance parameters of the methods rather than the large variety of realizations found in experiments. Representative detector installations will be discussed in Sect. 4 including some highlights of recents measurements. Before concluding in Sect. 6 a critical discussion of open problems and an outlook to forthcoming cosmic ray detectors that are under construction or currently planned is given in Sect. 5.

A more detailed presentation of detection techniques of air showers can be found in the text book of Blümer et al. (2011) and dedicated review articles (Nagano and Watson 2000; Haungs et al. 2003; Blümer et al. 2009).

2 Phenomenology of Extensive Air Showers

Extensive air showers are particle cascades developing in air (78.1% N_2, 20.9% O_2, and 0.9% Ar by volume), which can be considered as calorimeter of very low target density. The relation between vertical atmospheric depth, X_v, and altitude h can be expressed approximately as

$$X_v = \int_h^\infty \rho(h')\,\mathrm{d}h' = X_{\mathrm{ref}}\,\mathrm{e}^{-h/h_{\mathrm{ref}}}, \tag{1}$$

with $\rho(h)$ being the air density, X_{ref} the depth at sea level ($1{,}030\,\mathrm{g/cm}^2$), and $h_{\mathrm{ref}} = 8.4\,\mathrm{km}$ ($6.4\,\mathrm{km}$ for $X_v < 200\,\mathrm{g/cm}^2$) (Gaisser 1990). Some characteristic parameters of air are given in *Table 1* as function of altitude.

When a cosmic ray particle enters the atmosphere, the first interaction with an air nucleus takes place at altitudes between 15 and 35 km, depending on the mass of the particle and angle of incidence. The interaction points are exponentially distributed in depth:

$$\frac{\mathrm{d}P}{\mathrm{d}X} = \frac{1}{\lambda_{\mathrm{int}}}\mathrm{e}^{-X/\lambda_{\mathrm{int}}} \qquad \text{with} \qquad \lambda_{\mathrm{int}} = \frac{24{,}160\,\mathrm{mb\,g/cm}^2}{\sigma_{\mathrm{inel}}}, \tag{2}$$

where λ_{int} is the interaction length and σ_{inel} denotes the inelastic cross section with air. The interaction lengths of protons and pions in air are 85 and $120\,\mathrm{g/cm}^2$ at 100 GeV and decrease slowly with energy, reaching 60 and $75\,\mathrm{g/cm}^2$ at the energy of the knee. To a good approximation, the interaction length of iron nuclei is about a factor four times smaller.

The secondary particles of the first interaction are mainly pions, kaons, and nucleons. These particles interact themselves again if their interaction length is shorter than the corresponding decay length. For example, the interaction length of pions is about $120\,\mathrm{g/cm}^2$ at 100 GeV, corresponding to 900 m at sea level. Therefore neutral pions ($c\tau = 25\,\mathrm{nm}$; $c\gamma\tau = 18.6\,\mu\mathrm{m}$) decay immediately into two photons, and charged pions ($c\tau = 7.8\,\mathrm{m}$; $c\gamma\tau = 5.6\,\mathrm{km}$) of this energy interact again. Pions decay ($\pi^\pm \to \mu^\pm + \nu_\mu/\bar{\nu}_\mu$) only once their energy is degraded to about

Table 1

Parameters of air that are of relevance to air shower physics. The values are given for the US standard atmosphere (National Aeronautics and Space Administration 1976) relative to sea level

Altitude (km)	Vertical depth (g/cm^2)	Local density (10^{-3} g/cm^3)	Molière unit (m)	Electron Cherenkov threshold (MeV)	Cherenkov angle (°)
40	3	3.8×10^{-3}	2.4×10^4	386	0.076
30	11.8	1.8×10^{-2}	5.1×10^3	176	0.17
20	55.8	8.8×10^{-2}	1.0×10^3	80	0.36
15	123	0.19	478	54	0.54
10	269	0.42	223	37	0.79
5	550	0.74	126	28	1.05
3	715	0.91	102	25	1.17
1.5	862	1.06	88	23	1.26
0.5	974	1.17	79	22	1.33
0	1,032	1.23	76	21	1.36

30 GeV. The stable and relatively long-lived secondary hadrons (baryons, charged pions, and kaons) form the *hadronic shower component*. This hadronic shower core feeds all other shower components. High-energy photons from the decay of π^0 are the dominant source for the *electromagnetic shower component*. The decay of charged pions and kaons gives rise to the *muonic shower component*. In addition, up to 10% of the low-energy muons are produced by the em. shower component. Conversely, muon interaction and decay lead again to em. particles.

In the early years of cosmic ray physics, shower properties were calculated solving cascade equations, see Rossi and Greisen (1941), for example. Now it is common to simulate air showers in much more detail with the Monte Carlo method. Commonly used simulation packages are AIRES (Sciutto 1999, 2010), CORSIKA (Heck et al. 1998), CONEX (Bergmann et al. 2007), COSMOS (Kasahara et al.), and SENECA (Drescher and Farrar 2003). The latter three combine the numerical solution of cascade equations with Monte Carlo simulation techniques to increase the simulation speed. In addition to being a very efficient method to handle the large number of secondary particles in a shower, the Monte Carlo method allows the correct treatment of shower-to-shower fluctuations.

In ❯ *Fig. 2* the particle tracks of photon-, proton-, and iron-induced air showers of 10^{13} eV are shown. To illustrate the differences between the showers the electromagnetic, muonic, and hadronic components are shown separately. The em. component of showers is rather independent of the primary particle type, and the number of muons and hadrons can be used for estimating the type/mass of the primary particle.

In the following we will give an overview of analytic results describing shower properties that are used to derive information on the energy and mass or particle type of the primary particle. Up-to-date predictions from Monte Carlo simulations will be shown in ❯ Sect. 3. Additional information on the physics of air showers can be found in text books (Gaisser 1990; Stanev 2003; Grieder 2010) and recent review articles (Anchordoqui et al. 2004; Engel et al. 2011).

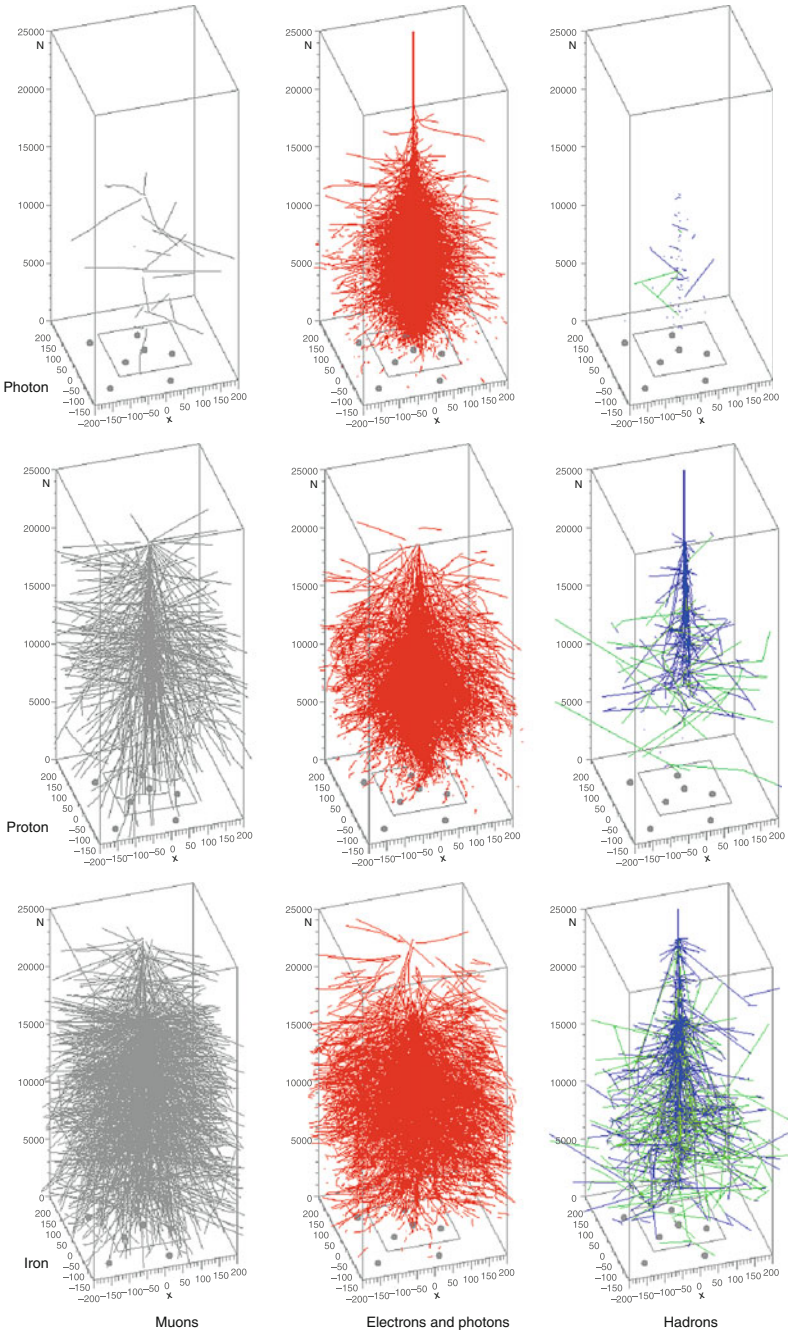


Fig. 2

Tracks of secondary particles of air showers induced by a photon, proton, and iron nucleus. The height of the graphs corresponds 25 km, and the width is 400 m. The simulations were done with CORSIKA

2.1 Photon-Induced Showers

The properties of em. showers follow from the interplay of e^+e^- pair production by photons and the emission of photons in electron bremsstrahlung. (In the following electrons and positrons are referred to as "electrons" as is typically done in air shower physics.) The radiation length X_0 in air is 37 g/cm^2, which corresponds to about 300 m at sea level. The electron energy loss can be written as $\mathrm{d}E/\mathrm{d}X = -\alpha(E) - E/X_0$, with α being the ionization energy loss that, in general, depends on both the electron energy and the density of air (Sternheimer et al. 1984). The critical energy, E_c, at which the ionization energy loss of electrons equals the radiative losses, is 86 MeV.

Due to the scale-invariance of the secondary particle distributions in em. interactions at high energy, em. cascades can be treated analytically in the limit of constant α to obtain the characteristics of the mean shower profile (Rossi and Greisen 1941; Lipari 2008). The number of particles at shower maximum is approximately proportional to the primary energy E_0, and the depth of shower maximum, $X_{\max}$, depends logarithmically on E_0:

$$\langle X_{\max}\rangle \approx X_0 \ln\left(\frac{E_0}{E_c}\right) + \frac{1}{2}X_0. \tag{3}$$

An often used parametrization of the mean longitudinal shower size profile, i.e., number of charged particles, is (Greisen (1956), also see derivation in Lipari (2008))

$$N_e(X) = \frac{0.31}{\sqrt{\ln(E_0/E_c)}} \exp\left\{\left(1 - \frac{3}{2}\ln s\right)\frac{X}{X_0}\right\}, \tag{4}$$

where s is the *shower age* parameter. The exact definition of s is related to the analytic treatment of cascade equations (Greisen 1956), a good approximation is $s = 3X/(X + 2X_{\max})$. Showers at maximum have the age $s = 1$, those with age $s < 1$ still have to reach their maximum, and those with $s > 1$ are at a development stage past their maximum (absorption phase). The electron energy distribution follows the age-dependent power law $\mathrm{d}N_e/\mathrm{d}E \propto E^{-(1+s)}$ for $E \gg E_c$.

Energy conservation leads to the track length integral

$$E_0 = \frac{E_c}{X_0}\int N_e(X)\,\mathrm{d}X \tag{5}$$

in the approximation of constant ionization energy loss $\alpha = E_c/X_0$ and absence of hadronic interactions (photoproduction) or muon pair production.

Mainly Coulomb scattering of electrons off air atoms leads to the lateral spread of the shower particles. The average RMS of the deflection angle of an electron, $\Delta\theta$, can be calculated in multiple scattering theory in the small-angle approximation:

$$\langle\Delta\theta^2\rangle = \left(\frac{E_s}{E}\right)^2 \frac{\Delta X}{X_0}, \qquad E_s = m_e c^2\sqrt{\frac{4\pi}{\alpha_{\mathrm{em}}}} \approx 21\,\mathrm{MeV}, \tag{6}$$

with m_e and α_{em} being the electron mass and the fine-structure constant (Molière 1948). The length scale of the lateral distribution of low-energy particles in a shower is characterized by the *Molière unit* $r_1 = (E_s/E_c)X_0 \approx 9.3\,\mathrm{g/cm}^2$, see *Table 1*. Particles with $E \gg E_c$ have a lateral spread reduced by the factor E_c/E. Therefore most of the em. particles at large lateral distance have been produced by high-energy electrons or photons close to the shower axis at a depth only 2–3 radiation lengths higher up in the atmosphere.

An often used analytical expression for the lateral spread is that of Greisen (Greisen 1956), who parametrized the solutions of cascade equations obtained by Nishimura and Kamata (Nishimura 1965):

$$\frac{\mathrm{d}N_e}{r\,\mathrm{d}r\,\mathrm{d}\varphi} = N_e(X)\frac{\Gamma(4.5-s)}{2\pi r_1^2\Gamma(s)\Gamma(4.5-2s)}\left(\frac{r}{r_1}\right)^{s-2}\left(1+\frac{r}{r_1}\right)^{s-4.5}, \tag{7}$$

now known as Nishimura–Kamata–Greisen (NKG) function. Various improvements to this parametrization have been developed, see Capdevielle et al. (2002).

At very high energy, two additional processes change the characteristics of em. showers, see Risse and Homola (2007) for a review.

At about $E \gtrsim 10^{18}$ eV, subsequent interactions of photons or electrons with air can no longer be considered as independent and the scattering amplitudes have to be added coherently (Landau and Pomeranchuk 1953; Migdal 1956). This effect is known as Landau–Pomeranchuk–Migdal (LPM) suppression of new particle production in certain kinematic regions (Stanev et al. 1982; Klein 1999). Shower-to-shower fluctuations of em. showers increase and the depth of maximum is shifted deeper into the atmosphere.

Magnetic pair production and bremsstrahlung in the Earth's magnetic field (Erber 1966; Stanev and Vankov 1997) becomes important for photons with $E \gtrsim 10^{19.5}$ eV. Such interactions typically take place a thousand kilometers above the atmosphere. Mainly due to magnetic bremsstrahlung, a shower of more than 100 secondary photons and a few electrons is formed, which interact in the atmosphere simultaneously. Recent simulations of this effect can be found in Cillis et al. (1999), Vankov et al. (2003), Homola et al. (2005, 2007). As the primary energy is shared by many secondary particles, the LPM effect hardly influences such showers. Due to the superposition of many lower-energy em. showers, shower-to-shower fluctuations of converted primary photons are significantly reduced.

2.2 Hadron-Induced Showers

The difference between em. and hadronic showers can be understood qualitatively within the model of Matthews (Matthews 2005) of a simplified cascade (Heitler–Matthews model).

Suppose the hadronic interaction of a particle of energy E produces n_{tot} new hadronic particles, each with energy E/n_{tot}. Let us further assume that two third of these particles are charged pions (multiplicity n_{ch}) and one third neutral pions (multiplicity n_{neut}). Neutral pions decay immediately into em. particles ($\pi^0 \to 2\gamma$). Charged pions interact with air nuclei if their energy is greater than the typical decay energy E_{dec} and decay at lower energy, producing one observable muon per pion.

Then the number of muons in a shower is given by

$$N_\mu = \left(\frac{E_0}{E_{\mathrm{dec}}}\right)^{\beta}, \qquad \beta = \frac{\ln n_{\mathrm{ch}}}{\ln n_{\mathrm{tot}}} \approx 0.86\ldots0.93. \tag{8}$$

The number of muons produced in an air shower increases almost linearly with primary energy and depends on the air density (through E_{dec}) and the charged and total particle multiplicities of hadronic interactions. ❯ Equation 8 can be improved to account for different particle types, different secondary particle energies, and an energy-dependent secondary particle multiplicity, but the overall functional form does not change (Matthews 2005; Alvarez-Muniz et al. 2002). Realistic values for β and E_{dec} have to be determined with simulations. They depend on the

modeling of hadronic multiparticle production, the zenith angle of the shower, and the energy threshold for muon detection. Examples for different hadronic interaction models can be found in Alvarez-Muniz et al. (2002).

The energy transferred from the primary hadron to the em. shower component can be calculated within the same model:

$$E_{\mathrm{em}} = \left[1 - \left(\frac{2}{3}\right)^{n}\right] E_0, \tag{9}$$

where n is the number of interactions charged pions undergo before decaying. Here the assumption has been made that 1/3 of the energy is transferred to photons through π^0 decay in each interaction. For typical shower energies the number of generations of charged pions is about five to six (Meurer et al. 2006) and increases slightly with primary shower energy. Correspondingly the fraction of energy transferred to the em. component increases from about 70–80% at 10^{15} eV to 90–95% at 10^{20} eV. In contrast to the strong absorption seen for the em. shower component, muons reach ground with only minor absorption.

The em. shower component fed by the hadronic core behaves similarly to a photon-induced shower except that it does not correspond to the full energy of the primary hadron and that the elongation rate of the depth of shower maximum is somewhat smaller. The elongation rate of showers, D_{10}, is the amount by which the depth of maximum of a shower increases per decade of energy (note that sometimes also $D_e = D_{10}/\ln 10$ is used in literature). Approximately the following relation holds:

$$\begin{aligned} D_{10}^{\mathrm{had}} &\approx (1 - B_\lambda - B_n)\ln(10) X_0 \ln(E_0/E_c) \\ &\approx (1 - B_\lambda - B_n) D_{10}^{\mathrm{em}}, \end{aligned} \tag{10}$$

which is called *elongation rate theorem* (Linsley and Watson 1981). The coefficients $B_\lambda = -\mathrm{d}\lambda_{\mathrm{int}}/\mathrm{d}\ln E$ and $B_n = \mathrm{d}\ln(n_{\mathrm{tot}})/\mathrm{d}\ln E$ depend on the characteristics of hadronic multiparticle production. They are a measure of the increase of the interaction cross section and the amount of scaling violation of the secondary particle distributions. One particular implication of the elongation rate theorem is the fact that constant hadronic interaction cross sections and perfect scaling (e.g., energy-independent secondary particle distributions) lead to the maximum elongation rate of about 85 g/cm^2 per decade. Again Eq. 10 can be improved by adding higher-order contributions but the results are still very similar (Alvarez-Muniz et al. 2002).

Hadrons in an air shower exhibit a wide lateral distribution since secondary hadrons are produced at a typical, almost energy-independent transverse momentum of $p_\perp \sim$ 300–400 MeV/c, leading to a large angle of low-energy hadrons relative to the shower axis. Still the lateral distribution of photons and electrons in hadronic showers is very similar to that of purely em. showers. The em. component is mainly fed by neutral pions of high energy, for which the ratio between transverse and longitudinal momentum is very small. Hence the lateral spread of the bulk of the em. shower particles is still determined by multiple Coulomb scattering. Only at lateral distances $r \gg r_1$ hadronically produced photons begin to be important. These photons are produced by low-energy π^0 for which the longitudinal momentum is not much larger than the transverse one (Lafebre et al. 2009). The lateral distribution of muons is wider than that of em. particles as most of them are produced in the decay of low-energy pions for which ~350 MeV/c transverse momentum leads to a large angle to the shower axis (Drescher and Farrar 2003; Meurer 2006).

To illustrate the typical parameters of hadron-induced showers at high energy, mean longitudinal and lateral profiles of the different shower components are shown in Fig. 3 for

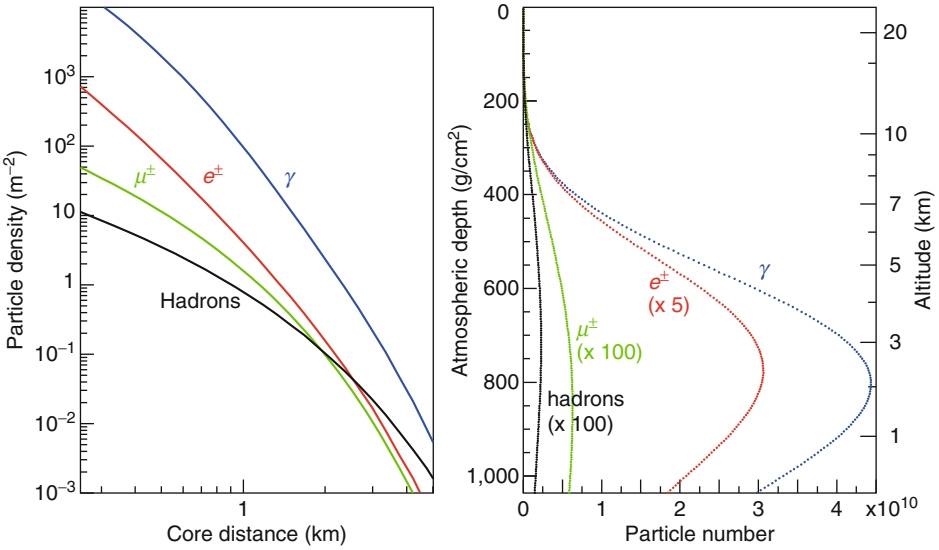


Fig. 3

Lateral and longitudinal shower profiles for vertical, proton-initiated showers of 10^{19} eV simulated with CORSIKA. The lateral distribution of the particles at ground is calculated for 870 g/cm^2, the depth of the Auger Observatory. The energy thresholds of the simulation were 0.25 MeV for γ, $e^{\pm}$ and 0.1 GeV for muons and hadrons. From Engel et al. (2011)

a proton-initiated shower of 10^{19} eV. The shower maximum is reached at about 1.5 km above sea level. The number of charged particles at shower maximum can be used to estimate the primary energy by multiplying it by 1.66 GeV, a relation that holds for a wide range of energies. The harder lateral distributions of muons and hadrons are clearly visible. Particles can be detected up to several kilometers from the shower core.

The *superposition model* can be used to extend the results discussed so far to showers initiated by nuclei. The binding energy of the nucleons in a nucleus is much smaller than the typical interaction energies, allowing the seemingly crude approximation that a nucleus with mass A and energy E_0 be considered as A independent nucleons with energy E_0/A. The superposition of A independent, nucleon-induced showers gives then

$$
\begin{aligned}
N_{\max}^{(A)}(E_0) &\approx A \cdot N_{\max}^{(p)}(E_0/A) \approx N_{\max}^{(p)}(E_0), \\
X_{\max}^{(A)}(E_0) &\approx X_{\max}^{(p)}(E_0/A), \\
N_{\mu}^{(A)}(E_0) &\approx A \cdot \left(\frac{E_0/A}{E_{\mathrm{dec}}}\right)^{\beta} = A^{1-\beta} \cdot N_{\mu}^{(p)}(E_0).
\end{aligned}
\tag{11}
$$

While the number of charged particles at shower maximum is almost independent of the primary hadron, the number of muons and the depth of maximum depend on the mass of the primary particle. The heavier the shower-initiating particle, the more muons are expected for a given primary energy. For example, iron showers contain about 40% more muons than proton showers of the same energy and reach their maximum 80–100 g/cm^2 higher in the atmosphere.

One of the important aspects of the superposition model is the fact that, averaged over many showers, the distribution of nucleon interaction points in the atmosphere coincides with that of more realistic calculations accounting for nucleus interactions and breakup into remnant nuclei (Engel et al. 1992). Therefore the superposition model gives a good description of many features of air showers as long as inclusive observables are concerned such as the mean depth of shower maximum and the number of muons. However, it is not applicable to observables related to correlations or higher-order moments (Battistoni et al. 1997; Kalmykov and Ostapchenko 1989).

Detailed shower simulations confirm qualitatively the energy and mass dependence for hadronic showers discussed here (Knapp et al. 1996, 2003; Engel et al. 2011). There is a considerable uncertainty of the predicted shower parameters that stems from our limited knowledge of hadronic multiparticle production. Model assumptions are needed for extrapolating accelerator measurements to higher energies or phase-space regions of secondary particles that are not measured in collider or fixed-target experiments.

2.3 Neutrino-Induced Showers

Neutrinos interact through neutral and charged current reactions with the quarks of nuclei in air. Their interaction cross section with a nucleon can be parametrized as (Glück et al. 1999)

$$\sigma_{\mathrm{CC}}^{\nu N} = \begin{cases} 1.10 \times 10^{-36} (E_\nu/\mathrm{GeV})^{0.454}\,\mathrm{cm}^2 & : \quad 10^5 \lesssim E_\nu/\mathrm{GeV} \lesssim 10^8 \\ 5.20 \times 10^{-36} (E_\nu/\mathrm{GeV})^{0.372}\,\mathrm{cm}^2 & : \quad 10^8 \lesssim E_\nu/\mathrm{GeV} \lesssim 10^{12} \end{cases} ,$$

$$\sigma_{\mathrm{NC}}^{\nu N} = \begin{cases} 3.55 \times 10^{-36} (E_\nu/\mathrm{GeV})^{0.467}\,\mathrm{cm}^2 & : \quad 10^5 \lesssim E_\nu/\mathrm{GeV} \lesssim 10^8 \\ 3.14 \times 10^{-36} (E_\nu/\mathrm{GeV})^{0.349}\,\mathrm{cm}^2 & : \quad 10^8 \lesssim E_\nu/\mathrm{GeV} \lesssim 10^{12} \end{cases} , \tag{12}$$

with a similar, but slightly smaller cross section for antineutrinos. Although their interaction cross section rises with energy the interaction probability in the atmosphere is only of the order of 10^{-5} for vertical incidence. This interaction probability can be much larger in theories with extensions to the Standard Model (Randall and Sundrum 1999; Domokos and Kovesi-Domokos 1999; Jain et al. 2000; Feng and Shapere 2002; Anchordoqui et al. 2010). One promising method of detecting neutrino-induced showers is that of searching for young, nearly horizontal air showers (Zas 2005).

The interaction of a neutrino with a nucleus is to a good approximation a point-like scattering off a quark in which either a leading charged lepton (charged current interaction) or a high-energy neutrino is produced (neutral current interaction). The energy carried away by the leading lepton is of the order of 70–90%. Only for electrons or τ leptons (through the decay products) this energy might be detectable. The struck quark fragments together with the remnant of the nucleon produce many secondary hadrons. Production of high-energy charm particles is enhanced in comparison to other hadronic interactions because of the mass and coupling of the exchanged W and Z bosons. In certain kinematic configurations one can expect to see a shower with two maxima, one coming from the hadronic shower produced by the quark fragmentation and another one from the shower initiated by the decay products of a leading τ lepton (Moura and Guzzo 2007).

Neutrino interactions can be simulated with standard air shower codes if an external generator, for example HERWIG (Corcella et al. 2001), is used for the first interaction.

3 Measurement Techniques and Observables

A large variety of different detection techniques are applied in air shower experiments. Whereas in the early years of cosmic ray physics typically only one of these techniques was used per experiment, it is now common to combine several of them in hybrid detectors to take advantage of measuring several observables simultaneously, which leads to a reduction of statistical and systematic uncertainties.

3.1 Particle Detector Arrays

Surface detector arrays consist of a set of particle detectors that are typically arranged on a regular pattern. Depending on the energy range the experiment is optimized for, the distance between the detector stations can vary from ~15 m (KASCADE (Antoni et al. 2003), Tibet AS-γ (Amenomori et al. 1990)) up to more than 1,000 m (Telescope Array (Kawai et al. 2008), Auger Observatory (Abraham et al. 2004)).

Showers are detected by searching for time coincidences of signals in neighboring detector stations. The arrival direction can then be determined from the time delay of the shower front reaching the different detectors, see *Fig. 4*. The shower appears like a disk of particles that is a few m thick in the center, increasing up to a few hundred m at large lateral distances. Only at small lateral distances the curvature of the shower front can be approximated by a sphere. The angular resolution of the reconstructed arrival direction depends on the distance and accuracy of time synchronization between the detector stations and the number of particles detected per station (for defining the arrival time of the shower front). Air shower arrays reach angular resolutions of typically 1–2° for low-energy showers and better than 0.5° for large showers.

The core position of the shower is found by fitting the signal $S(r)$ of the detector stations with a suited lateral distribution function (LDF). Preferentially the lateral distribution is determined from data directly using vertical showers. Because the NKG function (Eq. 7)

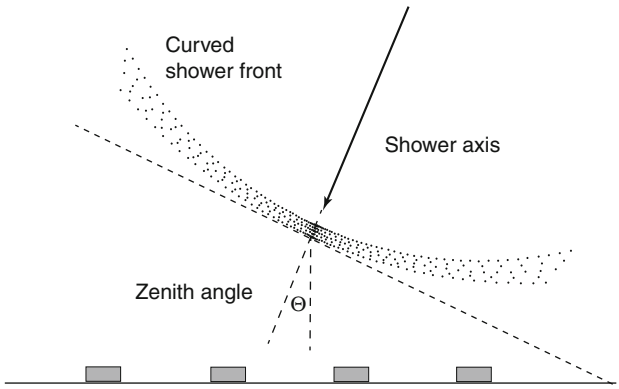


Fig. 4
Detection principle and geometry reconstruction of air showers with surface detector arrays

was developed to describe em. showers only, either modified versions of the NKG function such as

$$S(r) = C\left(\frac{r}{r_s}\right)^{-1.2}\left(1+\frac{r}{r_s}\right)^{-(\eta-1.2)}\left(1+\left(\frac{r}{1000\,\mathrm{m}}\right)^{2}\right)^{-\delta}, \tag{13}$$

(used for the scintillator array AGASA), or

$$S(r) = \tilde{C}\left(\frac{r}{r_s}\right)^{-\beta}\left(1+\frac{r}{r_s}\right)^{-\beta}, \tag{14}$$

and other phenomenological parametrizations such as

$$S(r) = \tilde{C}\, r^{-\left(\beta+\frac{r}{4000\,\mathrm{m}}\right)} \tag{15}$$

(used in water Cherenkov arrays such as Haverah Park and Auger) are applied. The parameters η, δ, β, and r_s are determined from data or simulations. If several shower components are measured separately (charged particles, muons), different LDFs are used. A list of often-used LDF parametrizations can be found in Haungs et al. (2003). There is a strong correlation between the arrival direction, shower curvature, core position, and asymmetry in the detector signal for non-vertical showers that has to be treated with care. The reached stat. uncertainty of the core positions varies from less than 1 m in the ideal case of a dense array and high-energy shower to up to more than 50 m for ultra-high-energy cosmic ray detectors.

To reconstruct the energy of a shower either the number of detected particles at ground is calculated by integrating the lateral distribution or a signal density at a specific lateral distance is determined. The latter is illustrated in *Fig. 5* for the Auger surface detector array. The measured signals of the detector stations of one particular event are reconstructed with

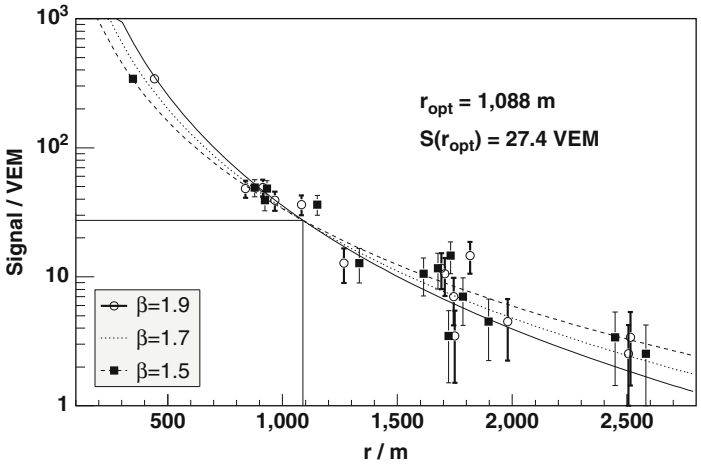


Fig. 5

Example of the determination of the optimum distance for measuring the particle density of an air shower in the Pierre Auger Observatory (see text). The detector signal is expressed in units of the signal expected for vertical muons (vertical equivalent muons, VEM). From Newton et al. (2007)

the LDF of Eq. 14 with different values of β. A fixed point is found at a core distance of about r_{opt} = 1,100 m (Newton et al. 2007). The signal (i.e., particle density) obtained for this distance is independent on the details of the LDF used for reconstruction and, hence, can be used as a robust estimator for determining the shower energy through comparison with Monte Carlo reference showers or cross-calibration with other calorimetric energy measurements. The optimum distance depends mainly on the spacing of the detectors and is not related to shower-to-shower fluctuations. An ideal detector configuration is reached if also the shower-to-shower fluctuations and the composition dependence of the lateral particle density exhibit a minimum at the optimum distance. This has been the case for the AGASA array with a detector distance of 1,000 m and a minimum of the shower fluctuations at 10^{19} eV in the range of 600–800 m (Dai et al. 1988). Typical reconstruction resolutions for the signal at optimum distance or total particle number at ground are in the range of 10–20%.

The most promising surface-detector approach is the separate measurement of the number of electrons and muons. The corresponding predictions for air showers simulated with the hadronic interaction models EPOS (Werner et al. 2006; Pierog and Werner 2009) and SIBYLL (Ahn et al. 2009) (interactions with $E > 80$ GeV) and FLUKA (Ballarini et al. 2006; Ferrari et al. 2005) (interactions with $E \leq 80$ GeV) are shown in *Fig. 6*. The simulation results confirm the predictions of the superposition model (Eq. 11) with a relative difference in the muon number between iron and proton showers of ~40%. The difference in the number of electrons is mainly related to the shallower depth of shower maximum of iron showers relative to proton showers. The uncertainties stemming from the simulation of hadronic multiparticle production in the showers affect the strongest the predictions for protons.

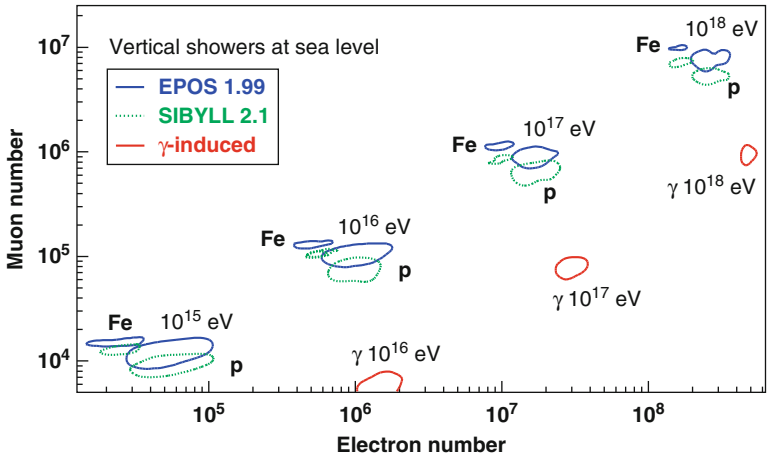


Fig. 6

Predicted correlation between the number of muons and electrons of vertical showers at sea level. The simulations were done with CORSIKA using the same cutoff energies for the secondary particles as in *Fig. 3*. The curves encircle approximately the one-sigma range of the fluctuations. From Engel et al. (2011)

With the energy transferred to the em. shower component being closely related to the number of muons at ground one can devise an almost model-independent estimator for the primary energy:

$$E_0 = E_{\mathrm{em}} + E_{\mathrm{had}} \approx \tilde{E}_c N_e^{(\max),e} + E_{\mathrm{dec}} N_\mu, \tag{16}$$

where $\tilde{E}_c > E_c$ is a typical energy scale one has to assign to electrons to compensate for the non-detected photons. In practical applications, the energy is parametrized as $\ln E = a \ln N_e + b \ln N_\mu + c$, with a, b, c being parameters determined from simulations. A similar expression can be written for $\ln A$ to find the primary mass, see Haungs et al. (2003), Hörandel (2007). Depending on the distance of the observation level to the depth of the typical shower maximum, fluctuations in the particle numbers can be large and need to be accounted for in energy and composition reconstruction.

Other composition estimators are based on the fact that muons dominate the early part of the time signal in the detector stations (rise time method (Walker and Watson 1982; Ave et al. 2003; Abraham et al. 2009)) and that the depth of shower maximum is related to the curvature of the shower front as well as the steepness of the lateral distribution (Ave et al. 2003; Dova et al. 2004). Recently also the signal asymmetry for inclined showers ($\theta < 60^\circ$) in azimuthal angle about the shower axis has been exploited (Dova et al. 2009).

In the early years of air shower measurements, air shower arrays have been the detectors with the largest acceptance leading to a number of fundamental discoveries. For example, using an array of hodoscope counters, Kulikov and Khristiansen discovered the knee in the spectrum in the electron number of showers in 1958 (Kulikov and Khristiansen 1958). Only a few years later the first shower with an energy of about 10^{20} eV was measured with the Volcano Ranch detector, an array of 20 scintillation detectors covering 12 km^2 (Linsley 1963). Bigger detectors followed in the attempt to find the upper end of the cosmic ray spectrum (SUGAR (Bell 1976), Haverah Park (Edge et al. 1973), Yakutsk (Glushkov et al. 1976), and AGASA (Chiba et al. 1992)).

Investigations of the flux and composition of primary cosmic rays in the knee energy range have been done with a number of particle detector arrays making important contributions (e.g., CASA-MIA (Glasmacher et al. 1999), EAS-TOP (Aglietta et al. 2004b), KASCADE – see ❯ Sect. 4.1, and GRAPES (Tanaka et al. 2008)). The combination of information from electromagnetic and muon detectors has been very important for these measurements. Alternatively, surface arrays can be operated in coincidence with deep underground muon detectors providing a complementary way of deriving composition information (e.g., EAS-TOP with MACRO (Aglietta et al. 2004b,c), IceTop with IceCube (Stanev et al. 2009)).

First attempts of detecting showers induced by gamma rays with surface arrays failed because of the high energy thresholds and the sparse active area of these instruments. Only recently detection of gamma-ray sources or source regions was achieved with the very-high-altitude detectors Tibet AS-γ (Amenomori et al. 1990) and ARGO-YBJ (Aielli et al. 2006), and the very densely instrumented Milargo Observatory (see ❯ Sect. 4.2).

3.2 Atmospheric Cherenkov Light Detectors

The large number of Cherenkov photons emitted by the charged shower particles when traversing a medium with refractive index $n > 1$ can be used for efficient detection of air showers in a wide range of energies. Imaging atmospheric Cherenkov telescopes (IACTs) can detect showers above an energy threshold of about 50 GeV (Aharonian et al. 2008; Hinton and Hofmann 2009)

but are limited in reach to very high energies due to the small effective area. Non-imaging Cherenkov detectors can be set up similar to an array of particle detectors, offering the possibility to instrument very large areas at ground and reach to very high energy (Budnev et al.; Ivanov et al. 2009). Typically only the Cherenkov light of the abundant secondary particles in an air shower is detected, but also the direct Cherenkov light of the primary particle can be measured (Kieda et al. 2001; Aharonian et al. 2007).

The production of Cherenkov radiation is discussed in detail in Chap. 18, "Cherenkov Counters", here we recall some important features of relevance to air shower detection. It is convenient to express the threshold of particle energy E for Cherenkov light emission in terms of the Lorentz γ factor:

$$\gamma \geq \frac{n(h)}{\sqrt{n(h)^2 - 1}}, \tag{17}$$

with $E = \gamma m$ and m being the particle mass. The height dependence of the refractive index $n(h)$ is a function of the local air density and satisfies approximately

$$n(h) = 1 + 0.000283 \frac{\rho_{\text{air}}(h)}{\rho_{\text{air}}(0)}, \tag{18}$$

where ρ_{air} is the density of air. The energy threshold for electrons and the Cherenkov angle θ_{Ch} in air, $\cos\theta_{\text{Ch}} = 1/(\beta n(h))$, are given in *Table 1* as function of height. Typical values at $h = 10$ km are $\theta_{\text{Ch}} = 0.8°$ (12 mrad) and a threshold of $\gamma = 72$, corresponding to $E = 37$ MeV for electrons and $E = 7.6$ GeV for muons.

The Cherenkov light cone of a particle at 10 km height has a radius of about 120 m at ground. This means that most of the light is expected within a circle of this radius. Due to multiple Coulomb scattering the shower particles do not move parallel to the shower axis. The angular distribution follows in first approximation an exponential,

$$\frac{\mathrm{d}N_\gamma}{\mathrm{d}\theta} = \frac{1}{\theta_0} \mathrm{e}^{-\theta/\theta_0}, \qquad \theta_0 = 0.83 E_{\text{th}}^{0.67}, \tag{19}$$

with E_{th} being the Cherenkov energy threshold (Hillas 1982a,b). Typical values of θ_0 are in the range $6 \ldots 8°$. The interplay of the altitude-dependent Cherenkov angle and the emission height leads to a typical lateral distribution of photons at ground, see *Fig. 7*. The absorption and scattering of Cherenkov light in the atmosphere limits the detectable wavelength range to about 300–450 nm, where the upper limit follows from the λ^{-2} suppression of large wavelengths. One possible parametrization of the lateral distribution of the Cherenkov light is (Fowler et al. 2001)

$$C(r) = \begin{cases} C_{120} \cdot \exp(a[120\text{ m} - r]); & 30\text{ m} < r \leq 120\text{ m} \\ C_{120} \cdot (r/120\text{ m})^{-b}; & 120\text{ m} < r \leq 350\text{ m} \end{cases}, \tag{20}$$

with the parameters C_{120}, a, and b.

Clear, moonless nights are required for taking data with air Cherenkov detectors resulting in an effective duty cycle of 10–15%. Also continuous monitoring of the atmospheric conditions including the density profile of the atmosphere are necessary (Bernlöhr 2000).

Arrays of photodetectors are used in non-imaging Cherenkov experiments to sample the lateral distribution of light in dark and clear nights. After reconstructing the core position, the measured parameter C_{120} and the slope are linked to the properties of the primary particle. Simulations show that the density of photons at 120 m from the core is almost directly proportional to the energy of the shower and that the slope is related to the depth of shower

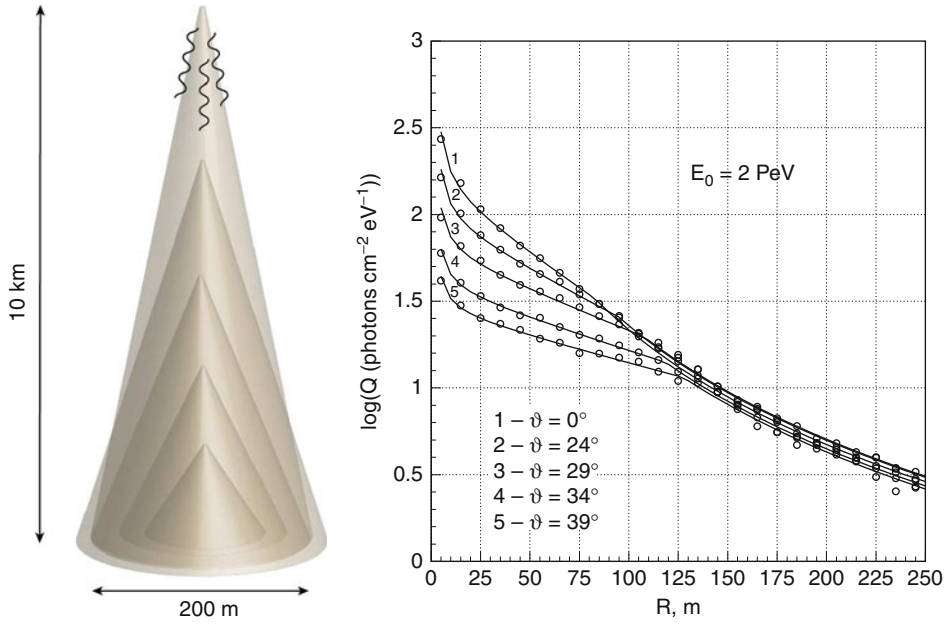


Fig. 7

***Left*: Illustration of the relation between production height and Cherenkov opening angle for producing the observed Cherenkov light distribution at ground. *Right*: Simulated lateral distributions of Cherenkov light produced by proton-induced showers of different zenith angle (Korosteleva et al. 2003). The simulations were done for a height of 2,000 m above sea level**

maximum (Hillas 1982b). Examples of surface arrays applying this non-imaging technique of shower detection via Cherenkov light are AIROBICC (Karle et al. 1995), EAS-TOP (Aglietta et al. 2004a), BLANCA (Fowler et al. 2001), Tunka (Budnev et al. 2009), and Yakutsk (Ivanov et al. 2009). The latter two are currently in operation, with Tunka being extended from an array of originally 25 stations to 133.

Recently much progress has been made in applying the imaging Cherenkov method to the detection of high-energy gamma rays. Two or more large Cherenkov telescopes are placed at a typical distance of about 100 m, allowing the reconstruction of shower direction and energy with high accuracy from stereoscopic images. The detection principle is illustrated in ❯ *Fig. 8*. Using shape parameters, photon-induced showers can be discriminated from the 10^5 times more abundant hadronic showers. While hadronic showers of GeV and TeV energies are characterized by a rather irregular structure due to the subshowers initiated by π^0 decay, photon-induced showers have a smooth overall shape. A moment analysis of the elliptical images in terms of the Hillas parameters (Hillas 1996) provides cuts to select gamma-ray showers. The technique of imaging atmospheric Cherenkov telescopes was developed and established with the monocular Whipple telescope (Mohanty et al. 1998). The largest atmospheric Cherenkov telescopes currently in operation are H.E.S.S. (Bernlöhr et al. 2003; Cornils et al. 2003), MAGIC (Ferenc 2005; Borla Tridon et al. 2010), and VERITAS (Weekes et al. 2002, 2010).

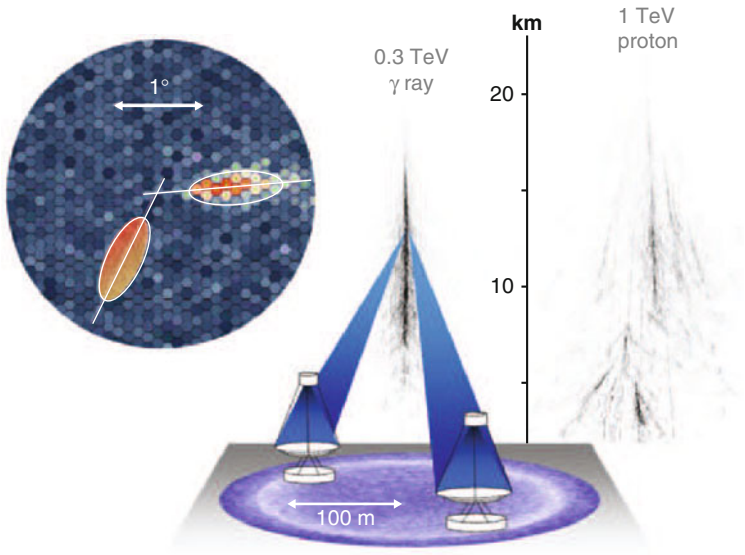


Fig. 8
Illustration of the stereo-detection principle of imaging atmospheric Cherenkov telescopes (Hinton and Hofmann 2009). The superimposed camera images are shown on the *left-hand side*. The intersection of the shower axes in this combined image corresponds to the arrival direction of the shower

3.3 Fluorescence Telescopes

If the shower energy exceeds $E \gtrsim 10^{17}$ eV, fluorescence light produced by nitrogen molecules in the atmosphere can be used to measure directly the longitudinal profile of air showers. Nitrogen molecules are excited by the charged particles of an air shower traversing through the atmosphere. The de-excitation proceeds through different channels of which two transitions of electronic states, called 2P and 1N for historical reasons, lead in combination with the change of the vibrational and rotational states of the molecule to several fluorescence emission bands. The spectral distribution of the fluorescence light is shown in *Fig. 9*. Most of the fluorescence light emission is found in the wavelength range from 300 to 400 nm. The lifetime of the excited states of nitrogen is of the order of 10 ns.

The number of emitted fluorescence photons would follow directly from the ionization energy deposited by the shower particles in the atmosphere if there were no competing de-excitation processes. Collisions between molecules are the dominant non-radiative de-excitation processes (collisional quenching, see, for example, discussion in Keilhauer et al. (2006)). The importance of quenching increases with pressure and almost cancels the density dependence of the energy deposit per unit length of particle trajectory. This results in a weakly height-dependent rate of about 4–5 fluorescence photons produced per meter and charged particle at altitudes between 5 and 10 km. In contrast to the Cherenkov yield, the fluorescence yield cannot be predicted from theory. Therefore several experiments have been carried out to measure the yield under different atmospheric conditions, see Arqueros et al. (2008) for a review.

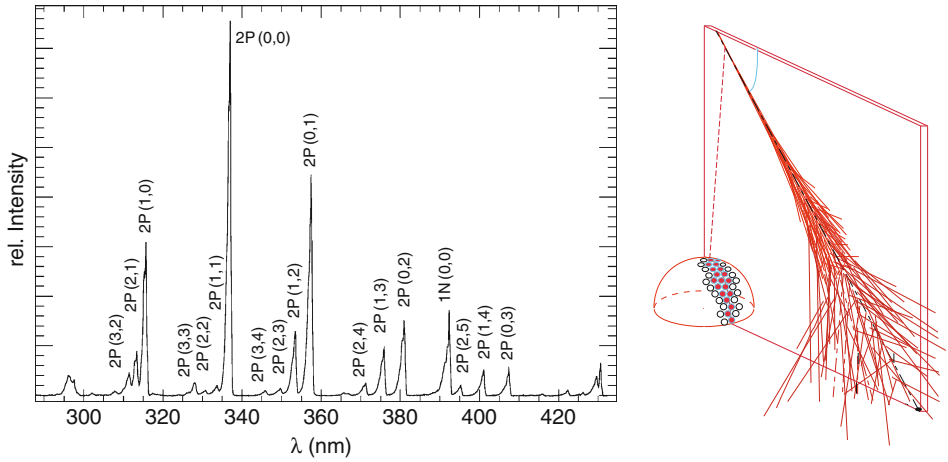


Fig. 9

***Left*: Fluorescence light spectrum of air at 20 °C and 800 hPa (Arciprete et al. 2006). The bands are labeled with the electronic transition type (2P or 1N) and the change of the vibration quantum number. *Right*: Illustration of the detection principle of fluorescence telescopes. The arrival angle of the shower can be measured with high precision in the shower–detector plane**

The reconstruction of a shower profile observed with a fluorescence telescope requires the determination of the geometry of the shower axis, the calculation of the Cherenkov light fraction, and the correction for the wavelength-dependent atm. absorption of light. In shower observations with one fluorescence telescope (monocular observation), the arrival angle perpendicular to the shower–detector plane can be determined with high precision, see *Fig. 9*. The orientation of the shower within this plane is derived from the arrival time sequence of the signals at the camera (Baltrusaitis et al. 1985; Kuempel et al. 2008). The angle ψ between the shower axis and the line of sight to the impact point of the shower core is related by

$$\chi_i(t_i) = \pi - \psi - 2\tan^{-1}\left(\frac{c(t_i - t_0)}{R_p}\right) \tag{21}$$

to the time of the signal with elevation angle χ_i (again measured in the shower-detector plane). The impact parameter R_p is given by the closest distance of the shower axis to the telescope and also determined in the time fit. The angular uncertainty of the orientation of the shower–detector plane depends on the resolution of the fluorescence camera and the length of the measured track. Typically a resolution of the order of 1° is obtained. In general, the reconstruction resolution of ψ is much worse and varies between 4.5° and 15° (for example, see Abbasi et al. (2007)). The reconstruction accuracy can be improved considerably by measuring showers simultaneously with two telescopes (stereo observation). Showers observed in stereo mode can be reconstructed with an angular resolution of about 0.6° (Abbasi et al. 2007). A similar reconstruction quality is achieved in hybrid experiments that use surface detectors to determine the arrival time of the shower front at ground (Bonifazi et al. 2005; Aglietta et al. 2007).

Knowing the geometry of the shower axis one can reconstruct the shower profile from the observed light intensities. While the highly asymmetric Cherenkov light has been subtracted from the light profile in the past (Baltrusaitis et al. 1985), new reconstruction methods take advantage of the Cherenkov light as additional shower signal (Unger et al. 2008). This is possible since universality features of air showers allow the accurate prediction of the emitted and scattered Cherenkov signal (Giller et al. 2005; Nerling et al. 2006).

The fluorescence technique allows a calorimetric measurement of the ionization energy deposited in the atmosphere. The integral over the energy deposit profile is a good estimator of the energy of the primary particle. At high energy, about 90% of the total shower energy is converted to ionization energy (Barbosa et al. 2004; Pierog et al. 2005). The remaining 10% of the primary energy, often referred to as missing energy, is carried away by muons and neutrinos that are not stopped in the atmosphere or do not interact. The missing energy correction depends on the primary particle type and energy as well as details on how hadronic interactions in air showers are modeled. However, as most of the shower energy is transferred to em. particles, this model dependence corresponds to an uncertainty of only a few percent of the total energy. In case of a gamma-ray particle as a primary, about 99% of the energy is deposited in the atmosphere.

The function proposed by Gaisser and Hillas (Gaisser and Hillas 1977) gives a good phenomenological description of individual as well as averaged longitudinal shower profiles:

$$N(X) = N_{\max} \left(\frac{X - X_1}{X_{\max} - X_1} \right)^{(X_{\max} - X_1)/\Lambda} \exp\left(-\frac{X - X_{\max}}{\Lambda} \right), \tag{22}$$

with X_1 and $\Lambda = 55\text{–}65\,\mathrm{g/cm^2}$ being parameters. It is often used to extrapolate the measured shower profiles to depth ranges outside the field of view of the telescopes.

A typical shower profile reconstructed with the fluorescence telescopes of the Auger Observatory (Abraham et al. 2010) is compared to simulated showers in ❯ *Fig. 10*. Both the mean depth of shower maximum and the shower-to-shower fluctuations of the depth of maximum carry important composition information.

The fact that the fluorescence light is emitted isotropically makes it possible to cover large phase-space regions with telescopes in a very efficient way. The typical distance at which a shower can be detected varies from 5 to 35 km, depending on shower geometry and energy. On the other hand, fluorescence detectors can be operated only at dark and clear nights, limiting the duty cycle to about 10–15%. Furthermore, continuous monitoring of atmospheric conditions is necessary, in particular the measurement of the wavelength-dependent Mie scattering length and detection of clouds (for example, see Abbasi et al. (2006), Abraham et al. (2010a)). The density profile of the atmosphere and seasonal variations of it have to be known, too (Keilhauer et al. 2004).

In 1976 fluorescence light of air showers was detected in a proof-of-principle experiment at Volcano Ranch (Bergeson et al. 1977) which was followed by the pioneering Fly's Eye experiment in 1982 (Baltrusaitis et al. 1985). The Fly's Eye detector was operated for 10 years, beginning with a monocular setup (Fly's Eye I) to which later a second telescope was added (Fly's Eye II). Fly's Eye II was designed to measure showers in coincidence with Fly's Eye I improving the event reconstruction by stereoscopic observation. In October 1991 the shower of the highest energy measured so far, $E = (3.2 \pm 0.9) \times 10^{20}$ eV, was detected with Fly's Eye I (Bird et al. 1995). The successor to the Fly's Eye experiment (Abu-Zayyad et al. 2000; Boyer et al. 2002), the High Resolution Fly's Eye (HiRes), took data from 1997 (HiRes I) and 1999 (HiRes II) to 2006. With an optical resolution of $1^\circ \times 1^\circ$ per camera pixel, a much better reconstruction of showers

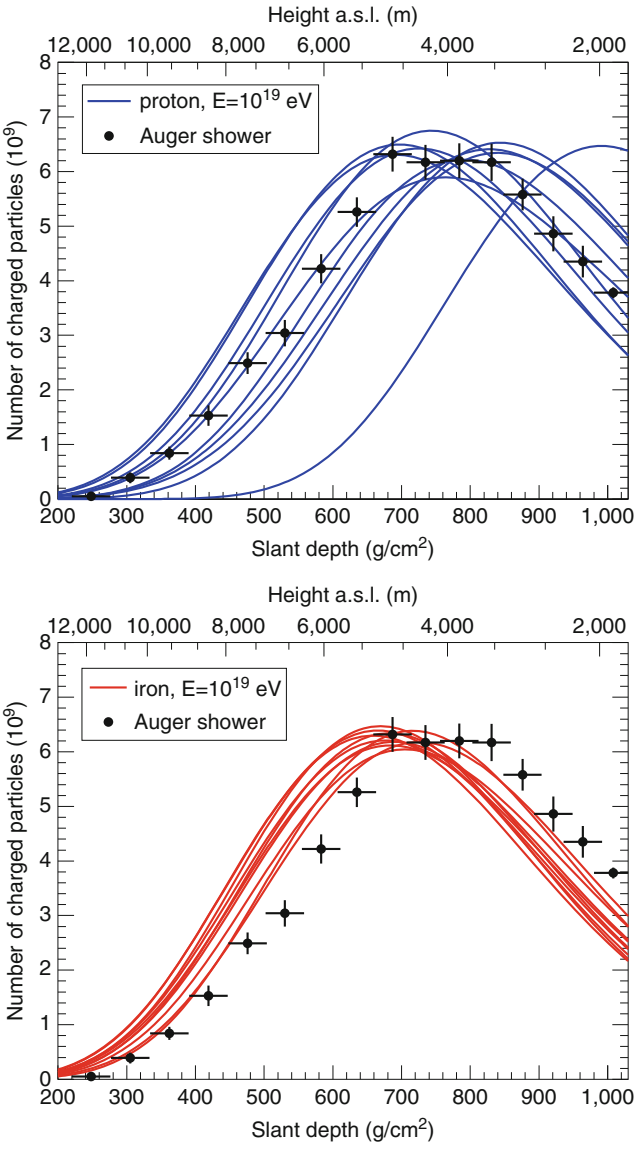


Fig. 10

Profile of one shower measured with the Pierre Auger Observatory (Blümer 2003). The reconstructed energy of this shower is about 10^{19} eV. The data are shown together with 10 simulated proton and iron showers to demonstrate the composition sensitivity of the depth of shower maximum. The showers were simulated with the SIBYLL interaction model (Engel et al. 1992; Ahn et al. 2009) and the CONEX air shower package (Bergmann et al. 2007)

was achieved. Currently, there are two fluorescence telescope systems taking data, both measuring in coincidence with a surface detector array. The Telescope Array (TA) detector (Tokuno et al. 2010) in the northern hemisphere consists of three fluorescence detector stations (Tameda et al. 2009), roughly located at the corners of a triangle of 35 km side length, that view the atmosphere above a scintillator array of 860 km^2 area. In the southern hemisphere, the Pierre Auger Observatory (Abu-Zayyad et al. 2000) is taking data with four fluorescence telescope stations (Abraham et al. 2010) and a surface array of water Cherenkov detectors covering about 3,000 km^2. For details see ❯ Sect. 4.5.

3.4 Radio Signal Detection

There are several sources of radio emission from extensive air showers. First of all there is the geo-synchrotron effect that stems from the charge separation of electrons and positrons in the shower disk while propagating through the magnetic field of the Earth (Kahn and Lerche 1966; Falcke and Gorham 2003; Huege and Falcke 2003), see ❯ *Fig. 11* (left). Another production mechanism is Cherenkov radiation that extends in wavelength into the radio range. Due to the large number of electrons in the atmosphere there is an asymmetry in the number of positrons and electrons in an air shower, the latter being about 20% more abundant. It is important to notice that these emission processes can be coherent if wavelengths larger than the typical thickness of the shower disk of a few meters, corresponding to frequencies smaller than 100 MHz, are considered. The expected electric field is then proportional to the number of electrons N_e. Hence the power radiated off by a shower scales quadratically with the number of particles and because of $N_e \propto E_0$ also quadratically with the shower energy.

The nature of the emission processes (Askaryan 1961, 1965) leads immediately to a number of qualitative predictions that have been confirmed in detailed calculations (Engel et al. 2006; Ludwig and Huege 2011; de Vries et al. 2010a). The geo-synchroton radiation is polarized transversely to the direction of motion of the particles and the local magnetic field, see ❯ *Fig. 11*. The charge excess radiation is polarized radially inward with respect to the shower axis. Typically the dominant contribution to the expected radio signal stems from the geo-synchrotron effect, followed by the charge excess contribution, and the Cherenkov signal being of the order of 10%. The radio radiation of a shower is strongly beamed in forward direction and leads to a lateral distribution with a width comparable to that of Cherenkov light.

A quantitative theory of radio emission from air showers is still under development. Different approaches are being pursued and only recently the results begin to converge (Huege et al. 2010). In macroscopic calculations the time variation of the charge excess and the current due to charge separation are parametrized and the radio signal is calculated using the retarded Liénard–Wiechert potential for the effective currents (Kahn and Lerche 1966; Scholten et al. 2008; Werner and Scholten 2008; Chauvin et al. 2010; de Vries et al. 2010b). A number of external input parameters are needed in these calculations to describe the path length and mean separation of $e^{\pm}$ in showers. Depending on the degree of detail of the implementation of shower features this approach can be used to predict the radio signal only at large distance from the core and shower disk, and for not too high frequencies. In contrast, adding up the radio signal from each shower particle during Monte Carlo simulation of a shower promises to account for all details of shower evolution and corresponding fluctuations (DuVernois et al. 2005; Engel et al. 2006; Kalmykov et al. 2009). This approach is numerically challenging and very time consuming. A good compromise seems to be the calculation of the expected radio signal

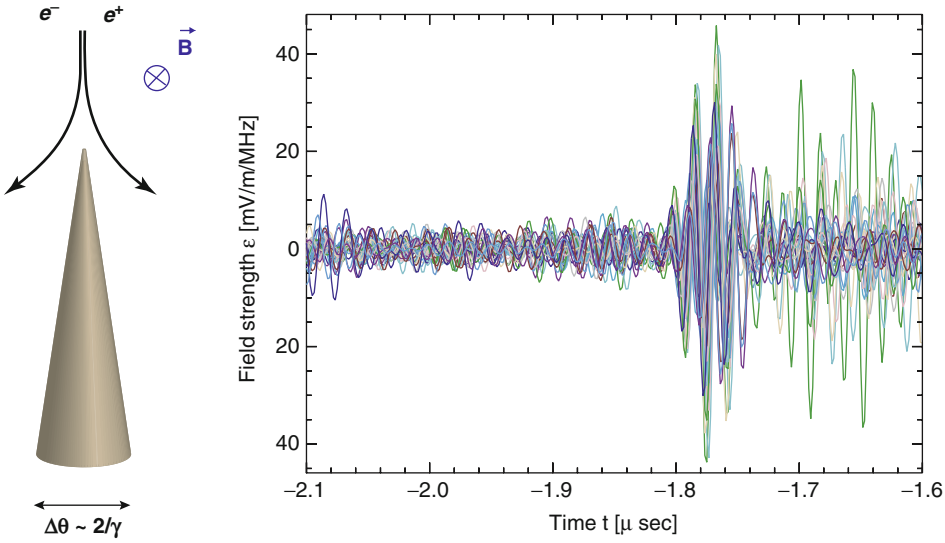


Fig. 11

***Left*: Illustration of synchrotron radiation of an e^+e^- pair in the geomagnetic field. The radiation is beamed and the opening angle of the cone is about $1/\gamma$, with γ being the Lorentz factor. *Right*: Radio pulse measured with LOPES in the frequency range 40–80 MHz (Apel et al. 2010). Different lines show the signal from different radio antennas. The incoherent signal after the radio pulse (starting at –1.7 μs) stems from the particle detectors in the KASCADE array**

with histogrammed showers, combining the radio signal of many particles before summing it up (Huege et al. 2007; Ludwig and Huege 2011).

The simulations predict a minimum of fluctuations in the lateral distribution of the electric field amplitude at a distance of about 100 m from the shower core. The field strength in this range correlates well with the primary energy of the shower. The slope of the lateral distribution is predicted to be directly related to the depth of shower maximum, independent of the shower energy (Huege et al. 2008; Kalmykov et al. 2009; de Vries et al. 2010), see *Fig. 12* (right). This relation offers the measurement of the mass composition of cosmic rays.

First radio pulses were measured by Jelley et al. (Jelley et al. 1965) already in 1965 to verify the prediction by Askaryan that air showers should produce electromagnetic pulses in the radio frequency range in the atmosphere (Askaryan 1961; Askaryan et al. 1965). A review of the early attempts of exploiting this signal for cosmic ray measurements is given in Allan (1971). Data analysis was hampered by the limited power of electronic signal processing and atmospheric monitoring at this time. Less than a decade ago new attempts of utilizing the radio signal of air showers were started (Falcke and Gorham 2003). The largest data sets currently available come from the LOPES (Falcke et al. 2005) and CODALEMA (Ardouin et al. 2005) experiments, both being triggered by scintillator arrays and measuring radio pulses in the 30–80 MHz range. The technology of triggering on the radio signal directly is currently under development (Ardouin et al. 2011; Dallier 2011). The data confirm both the approximately linear scaling

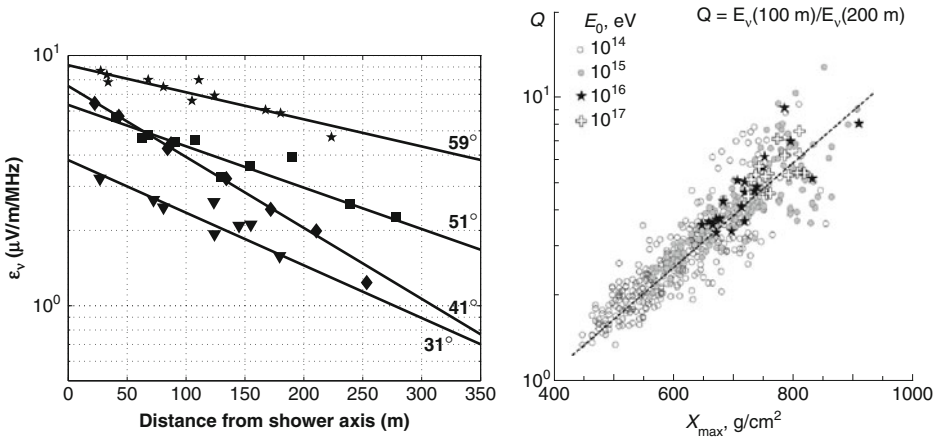


Fig. 12

***Left*: Lateral distribution of the electric field amplitude in showers measured with CODALEMA (Ardouin et al. 2006). *Right*: Simulations predict a direct relation between the depth of shower maximum and the slope of the lateral distribution of the electric field, here calculated for a frequency of ν = 60 MHz (Kalmykov et al. 2009)**

of the electric field with energy and the geo-synchroton mechanism as the dominant source of the radio signal in the MHz range (Falcke et al. 2005; Ardouin et al. 2009). The observed lateral distribution of the electric field amplitude ϵ exhibits approximately an exponential distance dependence and can be parametrized by (Allan 1971)

$$\epsilon_\nu = \left(\frac{20\,\mu\mathrm{V}}{\mathrm{m}\cdot\mathrm{MHz}}\right)\cdot\left(\frac{E_0}{10^{17}\mathrm{eV}}\right)\cdot\sin\alpha\cdot\cos\theta\cdot\exp\left(-\frac{R}{R_0(\nu,\theta)}\right), \tag{23}$$

with E_0 the primary particle energy, α the angle between shower axis and the geomagnetic field, θ the zenith angle, and R the antenna distance to the shower axis. The distance scale $R_0(\nu,\theta)$ is a bandwidth-dependent parameter being in the range of 125–200 m for most events (Apel et al. 2010). Examples of lateral distributions are shown in *Fig. 12* (left).

4 Examples of Air Shower Detectors

In the following some typical air shower detector installations are reviewed. The large diversity of such detectors makes it impossible to discuss examples of all the different detection techniques.

4.1 KASCADE

KASCADE (Karlsruhe Shower Core and Array Detector) is a muti-detector complex combining a classic air shower array for the electromagnetic and muonic components of showers with a

central calorimeter and a muon tracking detector (Antoni et al. 2003). The KASCADE detector is located in Karlsruhe, Germany (49.1° N, 8.4° E), at an altitude of 110 m above sea level. The layout of the detector complex is shown in *Fig. 13*.

The scintillation detectors of the air shower array are housed in 252 stations on a rectangular grid with 13 m spacing. The detector stations contain liquid scintillators of 0.78 m^2 for measuring charged particles with a detection threshold of about 5 MeV. (There are two scintillation detectors in each station of the outer clusters and four per inner station.) The stations of the outer detector clusters also contain plastic scintillators of 3.24 m^2 that are shielded by a layer of 10 cm of lead and 4 cm of iron for muon detection with a threshold of about 230 MeV. The central detector of 320 m^2 contains a hadron sampling calorimeter (eight layers of iron slabs and liquid scintillators) with a threshold of 50 GeV (Engler et al. 1999) and a muon tracking detector

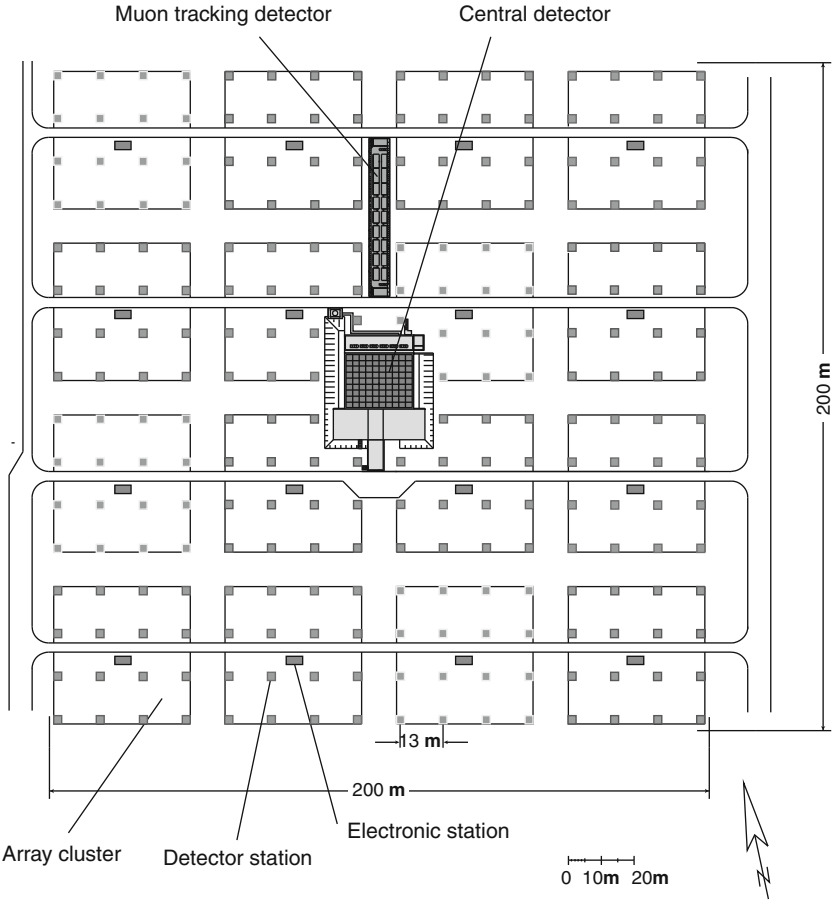


Fig. 13

Layout of the KASCADE detector with an effective area of 200 × 200 m^2 (Antoni et al. 2003). The detector stations of the array are grouped in 16 clusters for triggering and readout

(multiwire proportional chambers and a layer of limited streamer tubes) with an energy threshold of 2.4 GeV (Bozdog et al. 2001; Antoni et al. 2004). The muon tracking detector north of the central detector is built up of three layers of limited streamer tubes shielded by a layer of soil with a detection area of 128 m^2 for vertical muons (800 MeV detection threshold) (Doll et al. 2002).

The KASCADE detector has been operated since 1996. In 2003 an array of 37 scintillators with a spacing of about 137 m was added (KASCADE-Grande), increasing the array size to 0.5 km^2 (Apel et al. 2010). Regular data taking finished in 2009. Both air shower arrays are currently serving as trigger facility for other experiments such as LOPES (Falcke et al. 2005).

Important results obtained with the KASCADE detector include the flux and groundbreaking composition measurement in the knee energy range (Antoni et al. 2003, 2005), showing unambiguously that the composition changes toward a heavier one with increasing energy, and tests of hadronic interaction models (Antoni et al. 1999, 2001; Apel et al. 2006, 2007). The analysis of KASCADE-Grande data is in progress.

The air shower simulation package CORSIKA (Cosmic Ray Simulations for KASCADE) (Heck et al. 1998) was developed in the course of designing KASCADE and later continuously improved for analyzing the data taken with KASCADE and KASCADE-Grande. Nowadays CORSIKA has become the standard tool of almost all air shower experiments worldwide.

4.2 The Milagro Gamma-Ray Observatory

The Milagro detector was the first large-area water Cherenkov detector being built for continuously monitoring the sky for gamma-ray sources. It was located at an altitude of 2,630 m in the Jemez Mountains in New Mexico, USA (36° N), and took data between 2000 and 2007.

The layout of the Milagro detector is shown in *Fig. 14*. The core of the detector is a 60 m × 80 m water pond of 8 m depth. This pond was instrumented with two layers of PMTs with the photocathode of 20 cm diameter facing upward. The top layer of 450 PMTs was 1.2 m under the water surface and served the detection of showers (shower layer). The bottom layer of 273 PMTs, being 6 m below the water level, was used to detect the passage of muons for discriminating hadron- and photon-induced showers (muon layer). The pond was covered with a light-tight barrier.

With this setup it was possible to achieve full coverage of the detector area, allowing an efficient detection of all particles of an air shower that reach the surface of the water reservoir. Not only electrons and muons contribute to the light signal, also the much more abundant photons convert in the water and give rise to Cherenkov light from the secondary electrons. The threshold for detection of gamma-ray showers was saturated already at 1 TeV. At the same time more than 95% of the hadronic showers could be rejected based on the signal of the deep layer of PMTs.

To increase the sensitivity to higher-energy gamma rays the water pond was surrounded by 170 individual water Cherenkov detectors covering an area of 200 × 200 m^2. These outrigger detectors were water tanks of 1 m height and 3 m diameter, with a single PMT viewing the water from the top. The outrigger array increased the effective area of the detector by locating the core position of high-energy showers that triggered the water reservoir with the core position outside of the pond. The array also contributed to an improved reconstruction of the arrival direction angle of high-energy showers.

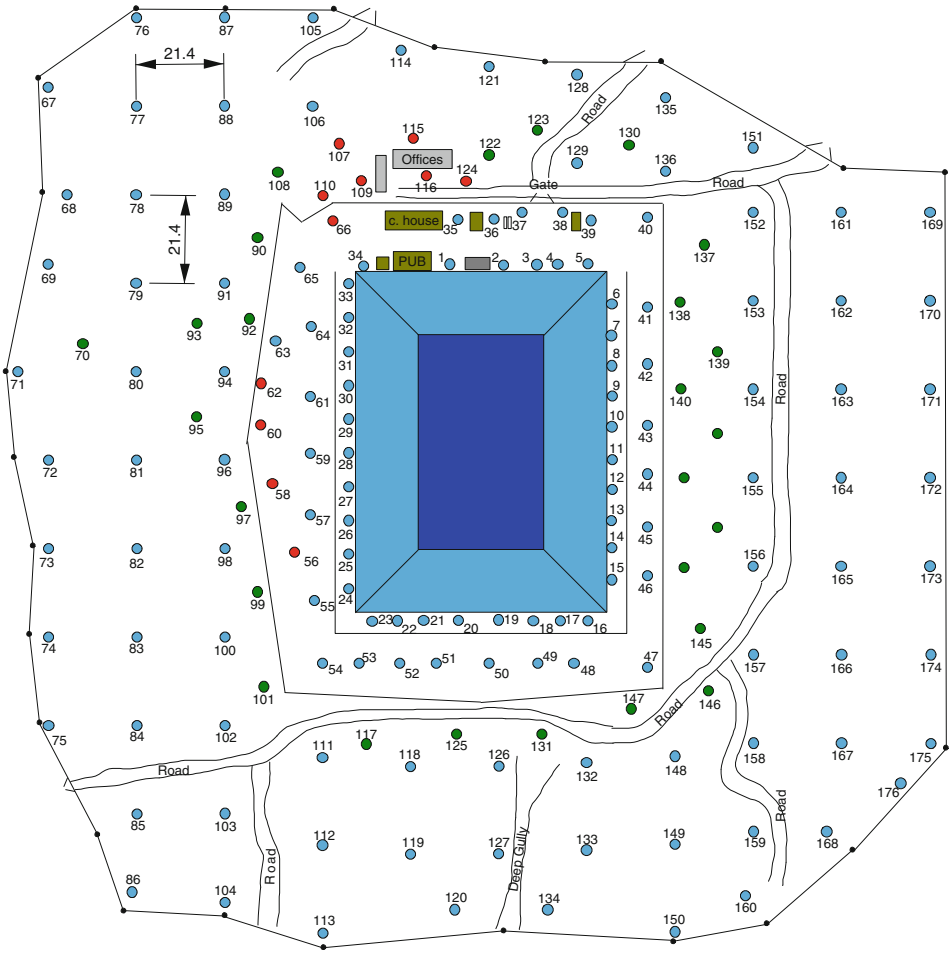


Fig. 14

Layout of the Milagro Gamma-Ray Observatory (Atkins et al. 2004). The central water detector is surrounded by 170 water Cherenkov tanks with an outer spacing of ~21 m to aid reconstruction of large showers

All events were reconstructed in real time at the site with typical trigger rates of 2 kHz. A gamma-ray signal would show up as an excess of events from the direction of the source or source region. The angular resolution of reconstructed showers was ~0.75°.

Important results from the Milagro experiment include the detection of diffuse emission from the Galactic plane at about 10 TeV (Atkins et al. 2005); the discovery of more than 10 sources of TeV γ rays (Abdo et al. 2007), including a new class of sources with large angular extent (Abdo et al. 2008a); and the discovery of an unexpected anisotropy of charged cosmic rays on the angular scale of ~10° (Abdo et al. 2008b).

4.3 Tunka

Tunka is a classic non-imaging air Cherenkov detector for observing the Cherenkov light flashes of hadronic showers in clear moonless nights. It is located at an altitude of 680 m in the Tunka valley near Lake Baikal, Russia (51°48′ N, 103°04′ E). The initial setup of 25 detector stations (Tunka-25, 0.1 km^2) was extended to 133 stations in 2009, covering now an area of 1 km^2 (Antokhonov et al. 2011).

Each of the 133 detector stations contains an upward-facing PMT with a photocathode of 20 cm diameter. Stations are grouped into 19 hexagonal clusters of seven detectors each. The signal is digitized by FADCs at 200 MHz. Each station is equipped with a remote-controlled cap that is closed during daytime to protect the PMTs.

One year of operation corresponds to about 400 h data taking time under ideal conditions. Both the lateral distribution derived from the time-integrated signal and the width of the time trace recorded by the stations can be used to derive composition information for the primary particles through the dependence on the depth of shower maximum.

The data of Tunka-25 have been analyzed (Budnev et al. 2009) and the mean depth of shower maximum has been derived in the knee energy range that is not accessible to fluorescence telescopes. The results give important, independent support to the composition estimate derived from ground-based particle arrays (e.g., KASCADE (Antoni et al. 2005) and EAS-TOP data (Aglietta et al. 2004b)).

4.4 H.E.S.S.

The H.E.S.S. (High Energy Stereoscopic System) is a high-energy gamma-ray telescope of the third generation. It consists of four imaging atmospheric Cherenkov telescopes located at the Khomas Highland (1,800 m a.s.l.; 23°16′ S, 16°30′ E) in Namibia. Key features of gamma-ray telescopes of the third generation are large mirror areas, cameras with very fine pixelation and large total field of view, and the stereoscopic observation. The large mirror size is required for reaching energy thresholds of ~100 GeV and below. The fine pixelation of the camera allows a very good discrimination between photon- and hadron-induced showers. At the same time a large field of view is of great advantage. The image of an individual shower is typically 1° wide and many gamma-ray sources are extended objects on the sky. Finally, stereoscopic observation improves the reconstruction quality of showers and enables a very efficient suppression of the background from single muons.

The H.E.S.S. telescopes have been designed for maximum mechanical rigidity and, at the same time, allowing full steering and automatic remote alignment of the mirror system of 107 m^2 per telescope, see *Fig. 15*. Each telescope has a focal length of 15 m and a mirror diameter of 13 m. The mirrors are built up of 382 individual reflectors of 60 cm diameter and and can be aligned individually to minimize the point spread function of the telescope. Each camera of the telescopes – weighing almost 1 t – consists of 960 pixels, each viewing 0.16° of the sky. The total field of view is about 5°. The telescopes can be pointed to a source with an angular precision of about 2.5″ and a slew rate of 100°/min.

The resulting performance of the H.E.S.S. array is the following. Showers can be reconstructed with an angular resolution better than 0.1°. Applying cuts on the Hillas parameters of the images, 50% or more gamma-ray events and less than 0.1% hadronic events are accepted. The detection threshold for gamma rays is about 100 GeV. The effective collection area exceeds

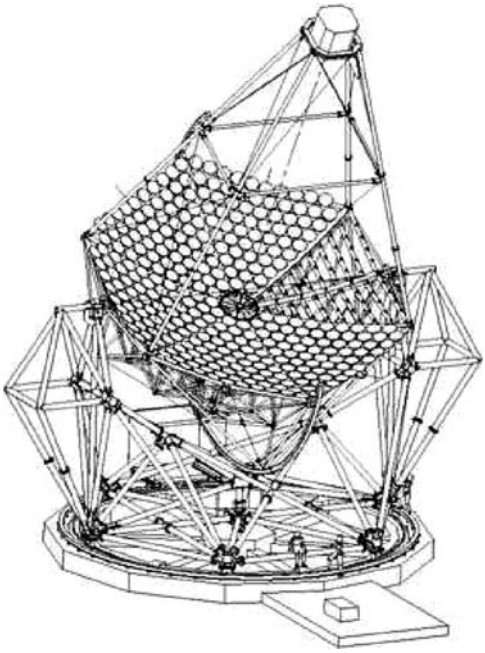


Fig. 15
Drawing of an H.E.S.S. telescope (Bernlöhr et al. 2003) showing the steel frame, mirror elements, camera, and steering mechanics. The reflectors are not shown in one segment to display the support beams

1 km^2 for showers above 10 TeV. The sensitivity of H.E.S.S. can be illustrated with the Crab Nebula as benchmark source. The Crab Nebula can be detected in about 30 s at 5σ confidence level.

To lower the energy threshold to 70 GeV and to increase the sensitivity of detection by a factor of 2, a giant 28 m telescope is currently being built in the center of the original H.E.S.S. array.

Many ground-breaking measurements and discoveries were made with H.E.S.S. One highlight is the galactic plane survey that resulted in more than 40 sources in a band of $60^\circ < l < 280^\circ$ of Galactic longitude l, many of them being spatially extended (Aharonian et al. 2006), and also the discovery of a new class of sources (Aharonian et al. 2006). Other outstanding results are the spatially resolved observation of galactic supernova remnants (e.g., RX J1713.7-3946 (Aharonian et al. 2004, 2006)), the measurement of the local electron spectrum (Aharonian et al. 2008), and the discovery of diffuse gamma-ray emission from the galactic center ridge (Aharonian et al. 2006).

4.5 The Pierre Auger Observatory

The Pierre Auger Observatory (35.3° S, 69.3° W), located at an altitude of 1,450 m in the Pampa Amarilla near the town Malargue, Argentina, is the largest air shower detector built so far. It has

been designed to investigate cosmic rays with energies exceeding 10^{19} eV, combining a surface array of particle detectors with fluorescence telescopes for hybrid detection (Abraham et al. 2004).

The layout of the observatory is shown in *Fig. 16*. An array of 1,600 water Cherenkov detectors on a triangular grid of 1.5 km spacing is covering an area of about 3,000 km^2. Each surface detector station contains 12 t of purified water that is viewed by three down-facing PMTs. The PMT signals are digitized with 40 MHz and stored in a ring buffer. The surface detector (SD) stations are solar-powered with backup batteries and communication is realized as custom-built wireless LAN in the ISM band. A central data acquisition system combines the trigger information of the detector stations and controls the transfer of signal traces from the detector stations.

The fluorescence detector (FD) consists of four fluorescence stations, each housing six fluorescence telescopes with a $30^\circ \times 30^\circ$ field of view per telescope (Abraham et al. 2010). The fluorescence telescopes operate independently of the surface array. However, time traces of relevant surface detectors are read out also for triggers coming from the fluorescence telescopes. Each telescope has an 11 m^2 mirror that focusses the light on a camera of 440 PMTs. The PMT signals are digitized at a rate of 10 MHz.

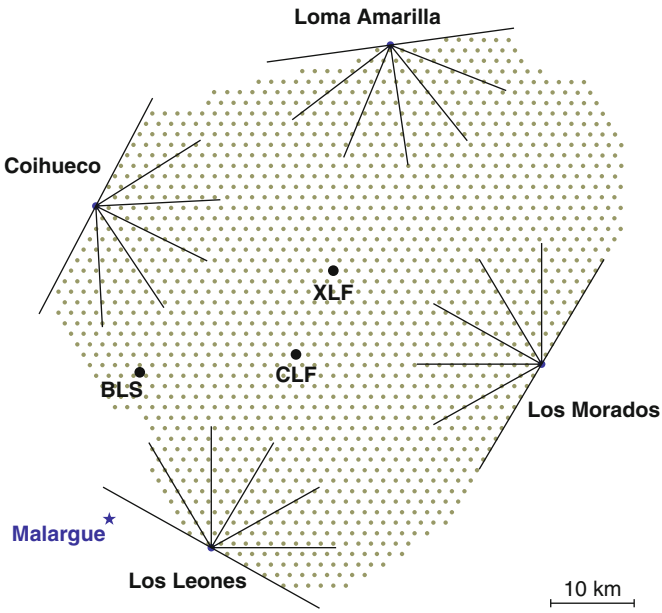


Fig. 16

Layout of the Auger Observatory in Argentina. Shown are the locations of the 1,600 surface detector stations. The field of view of the fluorescence telescopes is indicated by lines. Also marked are the locations of the two laser facilities in the array (CLF and XLF) and the balloon launching station (BLS)

A number of atmospheric monitoring devices are employed to ensure high-quality data (Abraham et al. 2010a). These are steerable LIDAR stations (BenZvi et al. 2007), infrared cameras for cloud detection, and weather stations at each of the fluorescence telescopes, as well as two UV lasers (Fick et al. 2006) in the surface detector array. Furthermore balloon-borne radio soundings of the atmosphere are performed.

The calorimetric measurement of the longitudinal shower profiles obtained with the fluorescence telescopes is used for calibrating the surface detector array that collects data with almost a 100% duty cycle. While the detection threshold of the surface array saturates at $10^{18.5}$ eV (Abraham et al. 2010), the fluorescence telescopes can detect showers with good quality also at energies as low as 10^{18} eV. The angular reconstruction accuracy depends on the shower energy and arrival angle. It is typically ~1.5° and improves to 0.7° for the highest-energy showers. Due to the height of the water Cherenkov detectors of 1.2 m, the Auger Observatory has also a good sensitivity to horizontal neutrino-induced showers (Zas 2005).

Several enhancements to the baseline design of the Auger Observatory are currently under construction. HEAT (High Elevation Auger Telescopes) comprises three additional fluorescence telescopes complementing the existing Coihueco telescopes by viewing higher elevations for reconstructing showers more reliably at ~10^{17} eV. The energy threshold of the surface detector array is reduced similarly by AMIGA (Auger Muons and Infill for the Ground Array) (Medina et al. 2006), which will consist of pairs of surface detector stations and muon detectors on a triangular grid of 750 m, covering an area of 24 km^2. Furthermore, the Auger Engineering Radio Array (AERA) with 24 MHz antennas (phase I) has been commissioned and is taking data to study the radio signal of air showers (Dallier 2011).

Highlights of results from the Auger Observatory include the confirmation of the theoretically expected flux suppression at energies higher than 6×10^{19} eV (Abraham et al. 2008a), the discovery of an anisotropic arrival direction distribution of cosmic rays in the suppression region (Abraham et al. 2007), and the measurement of the depth of shower maximum (Abraham et al. 2010b), indicating a mixed- or heavy-mass composition of cosmic rays at the highest energies. Furthermore, the searches for neutrino- and photon-induced air showers lead to the exclusion of many exotic models for the sources of ultra-high-energy cosmic rays (Abraham et al. 2008b, c).

5 Open Problems and Future Experiments

All indirect detection techniques of cosmic rays depend on our understanding of extensive air showers. In most cases shower measurements can only be interpreted by comparing the data to simulated reference showers. With no calculable theory of hadronic multiparticle production available so far, hadronic interactions have to be described by phenomenological models. The limited understanding of hadronic multiparticle production constitutes currently the main contribution to the systematic uncertainty of composition measurements (Knapp et al. 2003).

While there have been methods developed to derive the energy of the primary particle of a shower with only a small dependence on the modeling of hadronic interactions, there still seems to be a systematic difference of the order of 20% in the energy assignment between different experiments. New measurements of the absolute fluorescence yield will reduce these systematic uncertainties.

Regarding the measurement of the elemental composition of cosmic rays, the dependence on simulated air showers cannot be reduced much by better measurement techniques. Therefore, in the foreseeable future, significant progress can only be achieved by accelerator measurements of hadronic interactions of relevance to air shower physics to improve the reliability of shower simulations (Engel et al. 2011).

Model-independent, calorimetric methods of shower energy measurement rely on the detection of the em. shower component. While almost all energy is transferred to the em. component in showers of the highest energies, only a fraction of less than 70% of the energy is carried by em. particles in the knee energy range. Therefore it is necessary to measure also the muonic component of intermediate-energy showers to be able to estimate the primary particle energy reliably.

Current work toward new detection methods aims at developing techniques offering very large apertures to increase the statistics at the high-energy end of current experiments. Promising new techniques are, for example, the measurement of the coherent radio signal of air showers either with ground-based arrays (Falcke and Gorham 2003) or balloon-borne instruments (Hoover et al. 2010). Other investigations focus on searching for microwave radiation from air showers due to molecular bremsstrahlung, offering shower imaging similar to the observation of fluorescence light (Gorham et al. 2008) but with a much higher duty cycle. Similarly, measuring radar reflection from the plasma trail of ultra-high-energy air showers is a promising detection technique that offers very large apertures (Gorham 2001) but still has to be proven to work. Also the feasibility of air shower observations in the infrared wavelength range is studied (Conti et al. 2011).

Space-borne fluorescence detectors promise even larger apertures than those achievable with giant ground arrays (Santangelo and Petrolini 2009). The first step toward such a detector is planned with JEM-EUSO (Extreme Universe Space Observatory on board of the Japanese Experiment Module) (Gorodetzky 2011). JEM-EUSO is a fluorescence telescope equipped with Fresnel lenses of 2.65 m diameter, viewing the atmosphere from the orbit of the International Space Station at a height of ~400 km. The instantaneous aperture of such an instrument would be larger by a factor 56 (280) than that of the Auger Observatory if the camera is operated in downward (tilted) mode.

There are several new ground-based detectors planned for gamma-ray observations. The HAWC (High Altitude Water Cherenkov) telescope, the successor to the Milagro Observatory, is currently built at an altitude of 4,100 m at the volcano Sierra Negra in Mexico (Sinnis 2010). It will comprise 900 water Cherenkov tanks and have 10–15 times the sensitivity of Milagro with a wide field of view (~2 sr).

Also aiming at monitoring of the TeV gamma-ray sky with a large field of view is the multipurpose detector LHAASO (Cao et al. 2010) that will also measure charged cosmic rays over a wide energy range. To be built at an altitude of 4,300 m, LHAASO is planned to combine an array of water Cherenkov detectors with an active area of 90,000 m^2, a particle detector array of 1 km^2, a muon detector array of about 40,000 m^2, and two imaging air Cherenkov telescopes similar to MAGIC (Ferenc 2005; Borla Tridon et al. 2010).

The next-generation gamma-ray telescope will be the Cherenkov Telescope Array (CTA) (Hofmann and Martinez 2010). To cover the full sky, CTA is planned to consist of two arrays of Cherenkov telescopes, one in the northern and one in the southern hemisphere. A significant increase in sensitivity will be achieved by deploying large numbers (50–100) of Cherenkov telescopes at different distances. At the same time an extension of the energy range in comparison to existing telescopes will be accomplished by using telescopes of different sizes.

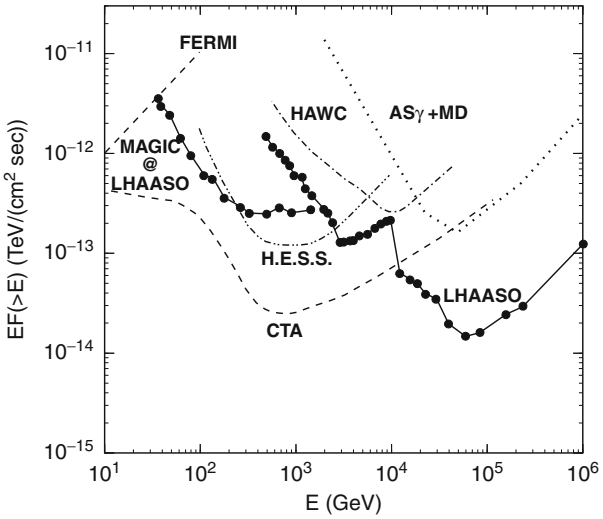


Fig. 17
Sensitivity of existing and planned gamma-ray experiments (Cao et al. 2010)

The sensitivities of existing and planned gamma-ray detectors in the TeV energy region are compared in *Fig. 17*.

6 Conclusion

Indirect methods of measuring cosmic rays allow us to detect particles with energies from a few hundred TeV up to the highest energies observed in the universe (~10^{20} eV). By observing the cascade of secondary particles produced by cosmic rays when interacting with nuclei of air one can derive information on the arrival direction, energy, and mass composition of the primary particle. Sparse arrays of particle detectors or imaging telescopes are sufficient for measuring key observables because of the large number of shower particles and the large lateral extent of the particle cascade at ground. Still the steeply falling flux of cosmic rays makes it very difficult to cover a wide range in energy with a single detector setup. Applying the indirect detection techniques developed for charged cosmic rays to showers induced by gamma rays has proven to be an efficient way of extending the classical, satellite-borne gamma-ray astronomy into and beyond the TeV energy range.

Acknowledgments

The author thanks his colleagues from the Pierre Auger, KASCADE-Grande, and LOPES Collaborations with whom he has worked on various subjects covered in this contribution. He is also grateful to Claus Grupen and Andreas Haungs for valuable comments on the manuscript.

References

Abbasi R et al. (HiRes Collab.) (2006) Techniques for measuring atmospheric aerosols at the high resolution fly's eye experimen. Astropart Phys 25:74–83

Abbasi RU et al. (HiRes Collab.) (2007) Search for point-like sources of cosmic rays with energies above 1018:5 eV in the HiRes-I monocular dataset. Astropart Phys 27:512–520

Abdo AA et al. (Milagro Collab.) (2007) TeV gamma-ray sources from a survey of the Galactic Plane with Milagro. Astrophys J 664:L91–L94

Abdo AA et al. (Milagro Collab.) (2008a) A measurement of the spatial distribution of diffuse TeV gamma ray emission from the Galactic Plane with Milagro. Astrophys J 688:1078–1083

Abdo AA et al. (Milagro Collab.) (2008b) Discovery of localized regions of excess 10-TeV cosmic rays. Phys Rev Lett 101:221101

Abraham J et al. (Pierre Auger Collab.) (2004) Properties and performance of the prototype instrument for the Pierre Auger observatory. Nucl Instrum Meth A52:50–95

Abraham J et al. (Pierre Auger Collab.) (2007) Correlation of the highest energy cosmic rays with nearby extragalactic objects. Science 318:938–943

Abraham J et al. (Pierre Auger Collab.) (2008a) Observation of the suppression of the flux of cosmic rays above 4×10^{19} eV. Phys Rev Lett 101:061101

Abraham J et al. (Pierre Auger Collab.) (2008b) Upper limit on the diffuse flux of UHE tau neutrinos from the Pierre Auger observatory. Phys Rev Lett 100:211101

Abraham J et al. (Pierre Auger Collab.) (2008c) Upper limit on the cosmic-ray photon flux above 10^{19} eV using the surface detector of the Pierre Auger observatory. Astropart Phys 29:243–256

Abraham J et al. (Pierre Auger Collab.) (2009) Studies of Cosmic Ray Composition and Air Shower Structure with the Pierre Auger Observatory. In: Proceedings of 31th International Cosmic Ray Conference, Lodz

Abraham J et al. (Pierre Auger Collab.) (2010a) A study of the effect of molecular and aerosol conditions in the atmosphere on air fluorescence measurements at the Pierre Auger observatory. Astropart Phys 33:108–129

Abraham J et al. (Pierre Auger Collab.) (2010b) Measurement of the Depth of Maximum of Extensive Air Showers above 10^{18} eV. Phys Rev Lett 104:091101

Abraham J et al. (Pierre Auger Collab.) (2010) Trigger and aperture of the surface detector array of the Pierre Auger observator. Nucl Instrum Meth A613:29–39

Abraham JA et al. (Pierre Auger Collab.) (2010) The fluorescence detector of the Pierre Auger observatory. Nucl Instrum Meth A620:227–251

Abu-Zayyad T et al. (HiRes Collab.) (2000) The prototype high-resolution Fly's Eye cosmic ray detector. Nucl Instrum Meth A450:253–269

Aglietta M et al. (EAS-TOP and MACRO Collab.) (2004a) The cosmic ray proton, helium and CNO fluxes in the 100-TeV energy region from TeV muons and EAS atmospheric Cherenkov light observations of MACRO and EAS-TOP. Astropart Phys 21:223–240

Aglietta M et al. (EAS-TOP Collab.) (2004b) The cosmic ray primary composition in the "knee" region through the EAS electromagnetic and muon measurements at EAS-TOP. Astropart Phys 21:583–596

Aglietta M et al. (MACRO Collab.) (2004c) The primary cosmic ray composition between 1016 eV and 10^{15} eV from extensive air showers electromagnetic and TeV muon data. Astropart Phys 20:641–652

Aglietta M et al. (Pierre Auger Collab.) (2007) Anisotropy studies around the galactic centre at EeV energies with the Auger observatory. Astropart Phys 27:244–253

Aharonian F et al. (H.E.S.S. Collab.) (2005) A new population of very high energy gamma-ray sources in the Milky Way. Science 307:1938–1942

Aharonian F et al. (H.E.S.S. Collab.) (2006) A detailed spectral and morphological study of the gamma-ray supernova remnant RX J1713.7-3946 with H.E.S.S. Astron Astrophys 449:223–242

Aharonian F et al. (H.E.S.S. Collab.) (2006) Discovery of very-high-energy gamma-rays from the galactic centre ridge. Nature 439:695–698

Aharonian F et al. (H.E.S.S. Collab.) (2007) First ground based measurement of atmospheric Cherenkov light from cosmic rays. Phys Rev D75:042004

Aharonian F et al. (H.E.S.S. Collab.) (2008) The energy spectrum of cosmic-ray electrons at TeV energies. Phys Rev Lett 101:261104

Aharonian F et al. (HESS Collab.) (2006) The HESS survey of the inner galaxy in very high-energy gamma-rays. Astrophys J 636:777–797

Aharonian F, Buckley J, Kifune T, Sinnis G (2008) High energy astrophysics with ground-based gamma ray detectors. Rept Prog Phys 71:096901

Aharonian FA et al. (HESS Collab.) (2004) High-energy particle acceleration in the shell of a supernova remnant. Nature 432:75–77

Ahn EJ, Engel R, Gaisser TK, Lipari P, Stanev T (2009) Cosmic ray interaction event generator SIBYLL 2.1. Phys Rev D 80:094003

Aielli G et al. (Argo-YBJ Collab.) (2006) Layout and performance of RPCs used in the Argo-YBJ experiment. Nucl Instrum Meth A562:92–96

Allan HR (1971) Radio emission from extensive air showers. Prog Element Part Cos Ray Phys 10:171

Allard D, Busca NG, Decerprit G, Olinto AV, Parizot E (2008) Implications of the cosmic ray spectrum for the mass composition at the highest energies. J Cosmol Astropart Phys 0810:033

Alvarez-Muniz J, Engel R, Gaisser TK, Ortiz JA, Stanev T (2002) Hybrid simulations of extensive air showers. Phys Rev D66:033011

Amenomori M et al. (1990) Development and a performance test of a prototype air shower array for search for gamma-ray point sources in the very high-energy region. Nucl Instrum Meth A288:619

Anchordoqui L et al. (2004) High energy physics in the atmosphere: phenomenology of cosmic ray air showers. Ann Phys 314:145–207

Anchordoqui LA et al. (2010) Using cosmic neutrinos to search for non-perturbative physics at the Pierre Auger Observatory. Phys Rev D82:043001

Antokhonov B et al. (TUNKA Collab.) (2011) The new Tunka-133 EAS Cherenkov array: status of 2009. Nucl Instrum Meth A628:124–127

Antoni T et al. (2004) A large area limited streamer tube detector for the air shower experiment KASCADE-Grande. Nucl Instrum Meth A533:387–403

Antoni T et al. (KASCADE Collab.) (1999) Test of high-energy interaction models using the hadronic core of EAS. J Phys G: Nucl Part Phys 25:2161

Antoni T et al. (KASCADE Collab.) (2001) Test of hadronic interaction models in the forward region with KASCADE event rates. J Phys G27:1785–1798

Antoni T et al. (KASCADE Collab.) (2003a) Preparation of enriched cosmic ray mass groups with KASCADE. Astropart Phys 19:715–728

Antoni T et al. (KASCADE Collab.) (2003b) The Cosmic ray experiment KASCADE. Nucl Instrum Meth A513:490–510

Antoni T et al. (KASCADE Collab.) (2005) KASCADE measurements of energy spectra for elemental groups of cosmic rays: results and open problems. Astropart Phys 24:1–25

Apel W et al. (KASCADE-Grande Collab.) (2010) The KASCADE-Grande experiment. Nucl Instrum Meth A620:202–216

Apel W et al. (LOPES Collab.) (2010) Lateral distribution of the radio signal in extensive air showers measured with LOPES. Astropart Phys 32:294–303

Apel WB et al. (KASCADE Collab.) (2006) Comparison of measured and simulated lateral distributions for electrons and muons with KASCADE. Astropart Phys 24:467–483

Apel WD et al. (KASCADE Collab.) (2007) Test of interaction models up to 40 PeV by studying hadronic cores of EAS. J Phys G34:2581–2593

Arciprete F et al. (2006) AIRFLY: Measurement of the air fluorescence radiation induced by electrons. Nucl Phys Proc Suppl 150:186–189

Ardouin D et al. (2011) First detection of extensive air showers by the TREND self-triggering radio experiment. Astropart Phys 34:717–731

Ardouin D et al. (CODALEMA Collab.) (2005) Radio-detection signature of high-energy cosmic rays by the CODALEMA experiment. Nucl Instrum Meth A555:148–163

Ardouin D et al. (CODALEMA Collab.) (2006) Radioelectric field features of extensive air showers observed with CODALEMA. Astropart Phys 26:341–350

Ardouin D et al. (CODALEMA Collab.) (2009) Geomagnetic origin of the radio emission from cosmic ray induced air showers observed by CODALEMA. Astropart Phys 31(3):192–200

Arqueros F, Hoerandel JR, Keilhauer B (2008) Air fluorescence relevant for cosmic-ray detection – summary of the 5th fluorescence workshop, El Escorial 2007. Nucl Instrum Meth A597:1–22

Askaryan GA (1961) Excess negative charge of an electron shower and its coherent radio emission. J Exp Theor Phys 14:441–443

Askaryan GA (1965) Coherent radio emission from cosmic showers in air and in dense media. J Exp Theor Phys 48:658–659

Atkins R et al. (Milagro Collab.) (2004) TeV gamma-ray survey of the northern hemisphere sky using the Milagro observatory. Astrophys J 608: 680–685

Atkins RW et al. (Milagro Collab.) (2005) Evidence for TeV gamma-ray emission from the Galactic plane. Phys Rev Lett 95:251103

Auger P, Ehrenfest P, Maze RR, Freon A (1939) Extensive cosmic-ray showers. Rev Mod Phys 11:288–291

Ave M et al. (2003) Mass composition of cosmic rays in the range $2 \times 10^{17} - 3 \times 10^{18}$ eV measured with the Haverah Park array. Astropart Phys 19: 61–75

Ave M, Knapp J, Marchesini M, Roth M, Watson AA (2003) Time structure of the shower front as measured at Haverah Park above 10^{19} eV. In: Proceedings of 28th International Cosmic Ray Conference, Tsukuba, p 349

Ballarini F et al. (2006) The FLUKA code: an overview. J Phys Conf Ser 41:151–160

Baltrusaitis RM et al. (Fly's Eye Collab.) (1985) The Utah Fly's Eye detector. Nucl Instrum Meth A240:410–428

Barbosa HMJ, Catalani F, Chinellato JA, Dobrigkeit C (2004) Determination of the calorimetric energy in extensive air showers. Astropart Phys 22: 159–166

Battistoni G, Forti C, Ranft J, Roesler S (1997) Deviations from the super-position model in a dual parton model applied to cosmic ray interactions with formation zone cascade in both projectile and target nuclei. Astropart Phys 7:49–62

Beatty JJ, Westerhoff S (2009) The highest-energy cosmic rays. Ann Rev Nucl Part Sci 59: 319–345

Bell CJ (1976) A recalculation of the upper end of the cosmic ray energy spectrum. J Phys G Nucl Phys 2:867–880

BenZvi SY et al. (2007) The lidar system of the Pierre Auger observatory. Nucl Instrum Meth A574:171–184

Berezinsky V, Gazizov AZ, Grigorieva SI (2006) On astrophysical solution to ultra high energy cosmic rays. Phys Rev D74:043005

Bergeson HE et al. (1977) Measurement of light emission from remote cosmic ray showers. Phys Rev Lett 39:847–849

Bergmann T et al. (2007) One-dimensional hybrid approach to extensive air shower simulation. Astropart Phys 26:420–432

Bernlöhr K (2000) Impact of atmospheric parameters on the atmospheric Cherenkov technique. Astropart Phys 12:255–268

Bernlöhr K et al. (H.E.S.S. Collab.) (2003) The optical system of the HESS imaging atmospheric Cherenkov telescopes, Part 1: layout and components of the system. Astropart Phys 20:111–128

Bird DJ et al. (Fly's Eye Collab.) (1995) Detection of a cosmic ray with measured energy well beyond the expected spectral cutoff due to cosmic microwave radiation. Astrophys J 441: 144–150

Blandford R, Eichler D (1987) Particle acceleration at astrophysical shocks: a theory of cosmic ray origin. Phys Rept 154:1–75

Blümer J (Pierre Auger Collab.) (2003) Cosmic rays at the highest energies and the Pierre Auger observatory. J Phys G29:867–879

Blümer J, Engel R, Hörandel JR (2009) Cosmic rays from the knee to the highest energies. Prog Part Nucl Phys 63:293–338

Blümer J, Engler J, Hörandel JR (2011) Detectors for astroparticle physics. Springer monographs, in press, 2011

Bonifazi C et al. (Pierre Auger Collab.) (2005) Angular resolution of the Pierre Auger Observatory. In: Proceedings of 29th International Cosmic Ray Conference, Pune, p 17

Borla Tridon D, Schweizer T, Goebel F, Mirzoyan R, Teshima M (MAGIC Collab.) (2010) The MAGIC-II gamma-ray stereoscopic telescope system. Nucl Instrum Meth A623:437–439

Boyer J, Knapp B, Mannel E, Seman M (2002) FADC-based DAQ for HiRes Fly's Eye. Nucl Instrum Meth A482:457–474

Bozdog H et al. (2001) The detector system for measurement of multiple cosmic muons in the central detector of KASCADE. Nucl Instrum Meth A465:455–471

Budnev NM et al. (2009) The cosmic ray mass composition in the energy range 10^{15}–10^{18} eV measured with the Tunka Array: Results and Perspectives arXiv:0902.3156 [astro-ph.HE]

Cao Z (ARGO-YBJ & LHAASO Collab.) (2010) The ARGO-YBJ Experiment Progresses and Future Extension, arXiv:1006.4298 [hep-ex]

Capdevielle J et al. (2002) Lateral-distribution functions for giant air showers. Nuovo Cim C25: 393–424

Chauvin J, Riviere C, Montanet F, Lebrun D, Revenu B (2010) Radio emission in a toy model with point-charge-like air showers Astropart Phys 33:341–350

Chiba N et al. (AGASA Collab.) (1992) Akeno giant air shower array (AGASA) covering 100-km^2 area. Nucl Instrum Meth A311:338–349

Cillis AN, Fanchiotti H, Garcia Canal CA, Sciutto SJ (1999) Influence of the LPM effect and dielectric suppression on particle air showers. Phys Rev D59:113012

Conti E, Sartori G, Viola G (2011) Measurement of the near-infrared fluorescence of the air for the detection of ultra-high-energy cosmic rays. Astropart Phys 34:333–339

Corcella G et al. (2001) HERWIG 6.5: an event generator for Hadron Emission Reactions With Interfering Gluons (including supersymmetric processes), JHEP 01 (2001) 010

Cornils R et al. (H.E.S.S. Collab.) (2003) The optical system of the HESS imaging atmospheric Cherenkov telescopes, Part 2: Mirror alignment and point spread function. Astropart Phys 20:29–143

Dai HY, Kasahara K, Matsubara Y, Nagano M, Teshima M (1988) On the energy estimation of ultrahigh-energy cosmic rays observed with the surface detector array. J Phys G14:793–805

Dallier R (Pierre Auger Collab.) (2011) Measuring cosmic ray radio signals at the Pierre Auger Observatory. Nucl Instrum Meth A630:218–221

de Vries KD, Scholten O, Werner K (2010a) Macroscopic Geo-Magnetic Radia-tion Model: Polarization effects and finite volume calculations, arXiv:1010.5364 [astro-ph.HE]

de Vries KD, v. d. Berg AM, Scholten O, Werner K (2010b) The lateral distribution function of coherent radio emission from extensive air showers: determining the chemical composition of cosmic rays, arXiv:1008.3308 [astro-ph.HE]

Doll P et al. (2002) Muon tracking detector for the air shower experiment KASCADE. Nucl Instrum Meth A488:517–535

Domokos G, Kovesi-Domokos S (1999) Strongly interacting neutrinos and the highest energy cosmic rays. Phys Rev Lett 82:1366–1369

Dova MT et al. (2009) Time asymmetries in extensive air showers: a novel method to identify UHECR species. Astropart Phys 31:312–319

Dova MT, Mancenido ME, Mariazzi AG, McCauley TP, Watson AA (2004) The mass composition of cosmic rays near 1018 eV as deduced from measurements made at Volcano Ranch. Astropart Phys 21:597–607

Drescher HJ, Farrar GR (2003) Air shower simulations in a hybrid approach using cascade equations. Phys Rev D67:116001

Drescher HJ, Farrar GR (2003) Dominant contributions to lateral distribution functions in ultrahigh energy cosmic ray air showers. Astropart Phys 19:235–244

DuVernois MA, Cai B, Kleckner D (2005) Geosynchrotron radio pulse emission from extensive air showers: simulations with AIRES. In: Proceedings of 29th International Cosmic Ray Conference (ICRC 2005), vol 8. Pune, India, 3–11 Aug 2005, pp 311–314

Edge DM, Evans AC, Garmston HJ (1973) The cosmic ray spectrum at energies above 10^{17} eV. J Phys A 6:1612–1634

Engel J, Gaisser TK, Stanev T, Lipari P (1992) Nucleus-nucleus collisions and interpretation of cosmic ray cascades. Phys Rev D46: 5013–5025

Engel R, Heck D, Pierog T (2011) Extensive air showers and hadronic interactions at high energy. Ann Rev Nucl Part Sci 61, in print

Engel R, Kalmykov NN, Konstantinov AA (2006) Simulation of radio signals from 1-TeV to 10-TeV air showers using EGSnrc. Int J Mod Phys A21S1:65–69

Engler J et al. (1999) A warm-liquid calorimeter for cosmic-ray hadrons. Nucl Instrum Meth A427:528–542

Erber T (1966) High-energy electromagnetic conversion processes in intense magnetic fields. Rev Mod Phys 38:626–659

Falcke H et al. (LOPES Collab.) (2005) Detection and imaging of atmospheric radio flashes from cosmic ray air showers. Nature 435:313–316

Falcke H, Gorham P (2003) Detecting radio emission from cosmic ray air showers and neutrinos with a digital radio telescope. Astropart Phys 19:477–494

Feng JL, Shapere AD (2002) Black hole production by cosmic rays. Phys Rev Lett 88:021303

Ferenc D (MAGIC Collab.) (2005) The MAGIC gamma-ray observatory. Nucl Instrum Meth A553:274–281

Ferrari A, Sala PR, Fasso A, Ranft J (2005) FLUKA: a multi-particle transport code (Program version 2005), CERN-2005-010

Fick B et al. (2006) The central laser facility at the Pierre Auger observatory. J Instrum 1:P11003

Fowler F, Fortson L, Jui C, Kieda D, Ong R et al. (2001) A Measurement of the cosmic ray spectrum and composition at the knee. Astropart Phys 15:49–64

Gaisser TK (1990) Cosmic rays and particle physics. Cambridge University Press, Cambridge

Gaisser TK, Hillas AM (1977) Reliability of the method of constant intensity cuts for reconstructing the average development of vertical showers. In: Proceedings of 15th International Cosmic Ray Conference, vol 8. Plovdiv, p 358

Giller M, Kacperczyk A, Malinowski J, Tkaczyk W, Wieczorek G (2005) Similarity of extensive air showers with respect to the shower age. J Phys G31:947–958

Glasmacher MAK et al. (CASA-MIA Collab.) (1999) The cosmic ray composition between 1014 eV and 1016 eV. Astropart Phys 12:1–17

Glück M, Kretzer S, Reya E (1999) Dynamical QCD predictions for ultrahigh energy neutrino cross sections. Astropart Phys 11:327–334

Glushkov A, Diminshtein Q, Efimov N, Kaganov L, Pravdin M (1976) Measurements of energy spectrum of primary cosmic rays in the energy range above 1017 eV. Izv Akad Nauk Ser Fiz 40: 1023–1025

Gorham PW (2001) On the possibility of radar echo detection of ultra-high energy cosmic ray and neutrino induced extensive air showers. Astropart Phys 15:177–202

Gorham PW et al. (2008) Observations of microwave continuum emission from air shower plasmas. Phys Rev D78:032007

Gorodetzky P (JEM-EUSO Collab.) (2011) Status of the JEM EUSO telescope on international space station. Nucl Instrum Meth A626–627:S40–S43

Greisen K (1956) The extensive air showers. Prog Cosmic Ray Phys 3:1–141

Greisen K (1966) End to the cosmic ray spectrum? Phys Rev Lett 16:748–750
Grieder PKF (2010) Extensive air showers: high energy phenomena and astrophysical aspects – a tutorial, reference manual and data book. Springer, Berlin
Hörandel JR (2004) Models of the knee in the energy spectrum of cosmic rays. Astropart Phys 21: 241–265
Hörandel JR (2007) Cosmic rays from the knee to the second knee: 10^{14} eV to 10^{18} eV. Mod Phys Lett A22:1533–1552
Haungs A, Rebel H, Roth M (2003) Energy spectrum and mass composition of high-energy cosmic rays. Rept Prog Phys 66:1145–1206
Heck D, Knapp J, Capdevielle J, Schatz G, Thouw T (1998) CORSIKA: a Monte Carlo code to simulate extensive air showers, Wissenschaftliche Berichte, Forschungszentrum Karlsruhe FZKA 6019
Hillas AM (1982a) Angular and energy distributions of charged particles in electron photon cascades in air. J Phys G8:1461–1473
Hillas AM (1982b) The sensitivity of Cherenkov radiation pulses to the longitudinal development of cosmic ray showers. J Phys G8:1475–1492
Hillas AM (1996) Differences between gamma-ray and hadronic showers. Space Sci Rev 75: 17–30
Hillas AM (2005) Can diffusive shock acceleration in supernova remnants account for high-energy galactic cosmic rays? J Phys G31:R95–R131
Hinton JA, Hofmann W (2009) Teraelectronvolt astronomy. Ann Rev Astron Astrophys 47: 523–565
Hofmann W, Martinez M (CTA Consortium) (2010) Design concepts for the cherenkov telescope array, arXiv:1008.3703 [astro-ph.IM]
Homola P et al. (2005) Simulation of ultra-high energy photon propagation in the geomagnetic field. Comput Phys Commun 173:71
Homola P et al. (2007) Characteristics of geomagnetic cascading of ultra-high energy photons at the southern and northern sites of the Pierre Auger observatory. Astropart Phys 27:174–184
Hoover S et al. (ANITA Collab.) (2010) Observation of ultra-high-energy cosmic rays with the ANITA balloon-borne radio interferometer. Phys Rev Lett 105:151101
Huege T, Falcke H (2003) Radio emission from cosmic ray air showers: coherent geosynchrotron radiation. Astron Astrophys 412:19–34
Huege T, Ludwig M, Scholten O, de Vriew KD (2010) The convergence of EAS radio emission models and a detailed comparison of REAS3 and MGMR simulations, arXiv:1009.0346 [astro-ph.HE]
Huege T, Ulrich R, Engel R (2007) Monte Carlo simulations of geosynchrotron radio emission from CORSIKA-simulated air showers. Astropart Phys 27:392–405
Huege T, Ulrich R, Engel R (2008) Energy and composition sensitivity of geosynchrotron radio emission from cosmic ray air showers. Astropart Phys 30:96–104
Ivanov AA, Knurenko SP, Sleptsov IY (2009) Measuring extensive air showers with Cherenkov light detectors of the Yakutsk array: the energy spectrum of cosmic rays. New J Phys 11:065008
Jain P, McKay DW, Panda S, Ralston JP (2000) Extra dimensions and strong neutrino nucleon interactions above 1019 eV: Breaking the GZK barrier. Phys Lett B484:267–274
Jelley J et al. (1965) Radio pulses from extensive cosmic-ray air showers. Nature 205:327–328
Kahn FD, Lerche I (1966) Radiation from cosmic ray air showers. Proc Roy Soc Lond A-289:206
Kalmykov NN, Konstantinov AA, Engel R (2009) Radio emission from extensive air showers as a method for cosmic-ray detection. Phys Atom Nucl 73:1191–1202
Kalmykov NN, Ostapchenko SS (1989) Comparison of characteristics of the nucleus nucleus interaction in the model of quark-gluon strings and in the superposition model. Sov J Nucl Phys 50:315–318
Karle A et al. (1995) Design and performance of the angle integrating Cherenkov array AIROBICC. Astropart Phys 3:321–347
Kasahara K et al., COSMOS http://cosmos.n.kanagawa-u.ac.jp/cosmosHome
Kawai H et al. (TA Collab.) (2008) Telescope array experiment. Nucl Phys Proc Suppl175–176: 221–226
Keilhauer B, Blümer J, Engel R, Klages HO (2006) Impact of varying atmospheric profiles on extensive air shower observation: fluorescence light emission and energy reconstruction. Astropart Phys 25:259–268
Keilhauer B, Blümer J, Engel R, Klages HO, Risse M (2004) Impact of varying atmospheric profiles on extensive air shower observation: Atmospheric density and primary mass reconstruction. Astropart Phys 22:249–261
Kieda DB, Swordy SP, Wakely SP (2001) A high resolution method for measuring cosmic ray composition beyond 10-TeV. Astropart Phys 15:287–303
Klein S (1999) Suppression of bremsstrahlung and pair production due to environmental factors. Rev Mod Phys 71:1501–1538
Knapp J, Heck D, Schatz G (1996) Comparison of hadronic interaction models used in air shower simulations and of their influence

on shower development and obsevables, in Wissenschaftliche Berichte FZKA 5828, Forschungszentrum Karlsruhe

Knapp J, Heck D, Sciutto SJ, Dova MT, Risse M (2003) Extensive air shower simulations at the highest energies. Astropart Phys 19:77–99

Kolhörster W, Matthes I, Weber E (1938) Gekopplete Höhenstrahlen. Naturwiss 26:576

Korosteleva E, Kuzmichev L, Prosin V (EAS-TOP Collab.) (2003) Lateral distribution function of EAS Cherenkov light: experiment QUEST and CORSIKA simulation. In: Proceedings of 28th International Cosmic Ray Conference, Tsukuba, pp 89–92

Kotera K, Olinto A (2011) The astrophysics of ultra-high energy cosmic rays, arXiv:1101.4256 [astro-ph.HE]

Kuempel D, Kampert KH, Risse M (2008) Geometry reconstruction of fluorescence detectors revisited. Astropart Phys 30:167–174

Kulikov GV, Khristiansen GB (1958) On the size spectrum of extensive air showers. J Exp Theor Phys 35:441–444

Lafebre S, Engel R, Falcke H, Hörandel J, Huege T, Kuijpers J, Ulrich R (2009) Universality of electron-positron distributions in extensive air showers. Astropart Phys 31:243–254

Landau LD, Pomeranchuk I (1953) Limits of applicability of the theory of bremsstrahlung electrons and pair production at high-energies. Dokl Akad Nauk Ser Fiz 92:535–536

Letessier-Selvon A, Stanev T (2011) Ultrahigh energy cosmic rays, to appear in Rev. Mod. Phys.

Linsley J (1963) Evidence for a primary cosmic-ray particle with energy 10^{20} *eV*. Phys Rev Lett 10:146–148

Linsley J (1998) Search for the end of the cosmic ray energy spectrum. AIP Conf Proc 433: 1–21

Linsley J, Watson AA (1981) Validity of scaling to 10^{20} eV and high-energy cosmic ray composition. Phys Rev Lett 46:459–463

Lipari P (2008) The concepts of "age" and "universality" in cosmic ray showers. Phys Rev 79:063001

Ludwig M, Huege T (2011) REAS3: Monte Carlo simulations of radio emission from cosmic ray air showers using an "end-point" formalism. Astropart Phys 34:438–446

Matthews J (2005) A Heitler model of extensive air showers. Astropart Phys 22:387–397

Medina MC et al. (2006) Enhancing the Pierre Auger observatory to the 10^{17} eV to $10^{18.5}$ eV range: capabilities of an infill surface array. Nucl Instrum Meth A566:302–311

Meurer C, Blümer J, Engel R, Haungs A, Roth M (2006) Muon production in extensive air showers and its relation to hadronic interactions. Czech J Phys 56:A211

Migdal AB (1956) Bremsstrahlung and pair production in condensed media at high-energies. Phys Rev 103:1811–1820

Mohanty G et al. (1998) Measurement of TeV gamma-ray spectra with the Cherenkov imaging technique. Astropart Phys 9:15–43

Molière GZ (1948) Theorie der Streuung schneller geladener Teilchen. II. Mehrfach-und Vielfachstreuung, Z. Naturforsch 3a, 78

Moura CA, Guzzo MM, Simulation of double-bang event in the atmosphere, hep-ph/0703145

Nagano M (2009) Search for the end of the energy spectrum of primary cosmic rays. New J Phys 11:065012

Nagano M, Watson AA (2000) Observations and implications of the ultrahigh-energy cosmic rays. Rev Mod Phys 72:689–732

National Aeronautics and Space Administration (NASA) (1976) U.S. Standard Atmosphere 1976, NASA-TM-X-74335

Nerling F, Blümer J, Engel R, Risse M (2006) Universality of electron distributions in high-energy air showers: description of Cherenkov light production. Astropart Phys 24:421–437

Newton D, Knapp J, Watson AA (2007) The optimum distance at which to determine the size of a giant air shower. Astropart Phys 26:414–419

Nishimura J (1965) Therory of cascade showers. Handbuch der Physik 46/2:1–113

Pierog T et al. (2005) Dependence of the longitudinal shower profile on the characteristics of hadronic multiparticle production. In: Proceedings of 29th International Cosmic Ray Conference, vol 7. Pune, p 103

Pierog T, Werner K (2009) EPOS model and ultra high energy cosmic rays. Nucl Phys Proc Suppl 196:102–105

Randall L, Sundrum R (1999) A large mass hierarchy from a small extra dimension. Phys Rev Lett 83:3370–3373

Risse M, Homola P (2007) Search for ultra-high energy photons using air showers. Mod Phys Lett A22:749–766

Rossi B, Greisen K (1941) Cosmic-ray theory. Rev Mod Phys 13:240–309

Santangelo A, Petrolini A (2009) Observing ultra-high-energy cosmic particles from space: S-EUSO, the super-extreme universe space observatory mission. New J Phys 11: 065010

Scholten O, Werner K, Rusydi F (2008) A macroscopic description of coherent geo-magnetic radiation from cosmic ray air showers. Astropart Phys 29:94–103

Sciutto SJ (1999) AIRES: A system for air shower simulations (version 2.2.0), astroph/9911331

Sciutto SJ (2010) The AIRES system for air shower simulations. An update, astro-ph/0106044

Sinnis G (HAWC and Milagro Collab.) (2010) Water Cherenkov technology in gamma-ray astrophysics. Nucl Instrum Meth A623:410–412

Stanev T (2003) High energy cosmic rays. Springer, Berlin

Stanev T et al. (IceCube Collab.) (2009) Status, performance, and first results of the IceTop array. Nucl Phys Proc Suppl 196:159–164

Stanev T, Vankov C, Streitmatter RE, Ellsworth RW, Bowen T (1982) Development of ultrahigh-energy electromagnetic cascades in water and lead including the Landau-Pomeranchuk-Migdal effect. Phys Rev D25:1291–1304

Stanev T, Vankov HP (1997) The nature of the highest energy cosmic rays. Phys Rev D55:1365–1371

Sternheimer RM, Berger MJ, Seltzer SM (1984) Density effect for the ionization loss of charged particles in various substances. At Data Nucl Data Tables 30:261–271

Tameda Y, Taketa A, Smith JD, Tanaka M, Fukushima M et al. (2009) Trigger electronics of the new fluorescence detectors of the telescope array experiment. Nucl Instrum Meth A609: 227–234

Tanaka H et al. (GRAPES-3 Collab.) (2008) Study on nuclear composition of cosmic rays around the knee utilizing muon multiplicity with GRAPES-3 experiment at Ooty. Nucl Phys Proc Suppl 175–176:280–285

Tokuno H et al. (TA Collab.) (2010) The telescope array experiment: status and prospects. AIP Conf Proc 1238:365–368

Unger M, Dawson BR, Engel R, Schüssler F, Ulrich R (2008) Reconstruction of longitudinal profiles of ultra-high energy cosmic ray showers from fluorescence and Cherenkov light measurements. Nucl Instrum Meth A588:433–441

Vankov HP, Inoue N, Shinozaki K (2003) Ultra-high energy gamma rays in geomagnetic field and atmosphere. Phys Rev D67:043002

Walker R, Watson AA (1982) Measurement of the fluctuations in the depth of maximum of showers produced by primary particles of energy greater than $1:5\times10^{17}$ eV. J Phys G8:1131–1140

Weekes T et al. (VERITAS Collab.) (2002) VERITAS: the very energetic radiation imaging telescope array system. Astropart Phys 17:221–243

Weekes T et al. (VERITAS Collab.) (2010) VERITAS: status summary 2009. Int J Mod Phys D19: 1003–1012

Werner K, Liu FM, Pierog T (2006) Parton ladder splitting and the rapidity dependence of transverse momentum spectra in deuteron gold collisions at RHIC. Phys Rev C74:044902

Werner K, Scholten O (2008) Macroscopic treatment of radio emission from cosmic ray air showers based on shower simulations. Astropart Phys 29:393–411

Zas E (2005) Neutrino detection with inclined air showers. New J Phys 7:130

Zatsepin GT, Kuzmin VA (1966) Upper limit of the spectrum of cosmic rays. J Exp Theor Phys Lett 4:78

Further Reading

Aharonian FA (2004) Very high energy cosmic gamma radiation: a crucial window on the extreme universe. World Scientific, Singapore

Blümer J, Engler J, Hörandel JR (2011) Detectors for astroparticle physics. Springer Monographs, Berlin

Gaisser TK (1990) Cosmic rays and particle physics. Cambridge University Press, Cambridge

Grieder PKF (2001) Cosmic rays at Earth: researcher's reference, manual and data book. Elsevier, Amsterdam

Grieder PKF (2010) Extensive air showers: high energy phenomena and astrophysical aspects – a tutorial, reference manual and data book. Springer, Berlin

Stanev T (2010) High energy cosmic rays, 2nd edn. Springer Praxis, Berlin

25 Technology for Border Security

Dudley Creagh
University of Canberra, Canberra, Australia

1 Introduction 634

2 Components of Security Systems 634
2.1 Passenger Portals 636
2.1.1 Passive Passenger Portals 637
2.1.2 Active mm-Wave Passenger Portal Technology 638
2.1.3 Active Devices: Metal Detectors 639
2.1.4 Active Devices: X-ray Passenger Portals 640
2.2 X-ray Baggage, Pallet, and Container Systems 643
2.2.1 X-ray Baggage Systems 647
2.2.2 X-ray Pallet and Air Cargo Systems 648
2.2.3 X-ray Shipping Container Examination Systems 649

3 Ancillary Technologies 649

4 Protocols: Passenger, Air Cargo/Pallet, Shipping Containers 651

5 Conclusions 651

Acknowledgments 651

References 651

C. Grupen, I. Buvat (eds.), *Handbook of Particle Detection and Imaging*,
DOI 10.1007/978-3-642-13271-1_25, © Springer-Verlag Berlin Heidelberg 2012

Abstract: This chapter discusses the physical and chemical principles that are employed in the construction of equipment for use in the border security and border protection contexts. Equipment used by airports, freight forwarding agents, and seaport authorities will be described but it should be noted that no performance figures will be given: there are obvious confidentiality and security issues precluding this.

1 Introduction

All airport passengers will be familiar with the basic elements that constitute airport security systems. Such systems comprise some kind of gateway through which the passengers pass, having first divested themselves of personal effects and outer clothing, an X-ray examination system in which these effects and their hand luggage are inspected, and some other device that is used to check whether articles in the passengers' possession have been in contact with illegal substances. In the meantime their hold baggage will have been examined by X-rays, and, perhaps, dogs trained to detect illicit substances will have smelt their bags. Similar procedures exist, but on a different scale, for both the freight forwarding and sea container applications.

Of utmost importance is the training and security screening of system operators, the training of dogs and other animals used for vapor detection, and the maintenance and calibration of the equipment used. A discussion of these issues is not in the scope of this chapter.

Governments and governmental blocs (for example, the European Union; European Parliament: Regulation (EC) No 300/2002: Common rules in the field of air safety security) regulate the use and operation of security systems. All examination systems must comply with the requirements of these regulations. Some jurisdictions (for example, the US Transport Security Administration; US Congress: Aviation Transportation and Security Act. 19 November 2001) provide certifications of the performance of individual items of equipment used in the security context.

In what follows descriptions of the principles of operation will be given for: existing and prototype equipment used as *passenger portals*, *X-ray*, *and neutron equipment* used in the passenger, freight forwarding, and sea container contexts, and particulate, liquid, and vapor phase systems used for trace detection (*Table 1*).

2 Components of Security Systems

Systems used in border security fall into two categories: those based on making physical measurements such as X-ray attenuation and those based on chemical reactions. Of these, systems using physical measurements are the most common. Usually measurements are made of some property related to the emission or absorption of electromagnetic radiation. Almost exclusively these measurements do not require that samples be taken: their interaction with the subject or object under investigation causes little change to them (the dosage given by X-ray examination systems used in whole-body scanning will be discussed later, as will the effect of the exciting laser in Raman Infrared spectrometers on samples taken for examination).

Table 1

Suppliers of technology used in the border protection context

Supplier	Contact address	Product
ACRO Security	www.acrosec.com	CHM
Ahura	www.ahurascientific.com	ETD, EBS
AS&E	www.as-e.com	XPP, XPB, XPS
APAC Security	www.apacsecurity.cn	ETD, EBS,
Banksia Scientific	www.banksiascientific.com.au	CHM
Brijot Imaging Systems	www.brijot.com	PPP
GE Homeland Security	www.gesecurity.com	ETD, EBS
GE Security	www.ge.com/de	XPB
Corporate Risk Solutions	www.risksolutions.com	XPP
Guardian	www.guardiantechintl.com	XIP
L3 Communications	www.l-3.com	APP, XPB, XPA, XPS
Nuctech	www.nuctech.com	XBS, XPB, XPA, XPS, XNS
Optosecurity	www.optosecurity.com	XIP
QR Sciences	www.qrsciences.com	NXC, XRE
Rapiscan	www.rapiscansystems.com	XPP, XPB, XPA
Smiths Detection	www.smithsdetection.com	ETD, XPB, XPA, XPS

CHM chemical, *ETD* electronic trace detection, *XRE* X-ray parcel device, *EBS* electronic bottle scanner, *PPP* passive passenger portal, *APP* active passenger portal, *XPP* X-ray passenger portal, *XBS* X-ray bottle scanner, *NXC*NQR X-ray cartons, *XP* X-ray passenger baggage, *XPA* X-ray air cargo/pallet, *XPS* X-ray shipping container, *XNS* X-ray and neutron shipping container, *XIP* X-ray image processing

Passive (systems that rely solely on radiation from the subject) and active systems for use in *whole-body scanners* in passenger portals operate in the THz or mm-wave regions of the spectrum. *Metal* detectors are active systems operating in the low to medium band of the radio frequency (RF) spectrum. X-ray systems used for passenger portals operate in the 50–150 keV range.

X-ray systems for baggage, pallet, and shipping container examination usually operate with X-ray energies in the range of 300 keV to 10 MeV. Gamma-rays, which are sometimes used, have energies around 1 MeV (present systems use ^{60}Co:E_γ = 1.17, 1.334 MeV; $T_{1/2}$ = 5.25 years). Increasingly there is interest in the use of high-energy neutron systems for baggage and freight examination. These "fast" neutrons have a maximum energy of 14 MeV.

Other systems that operate using physical measurements directly involving electromagnetic radiation include bottle and container systems, nitrogenous explosives detectors (low-frequency RF; these use (Nuclear) Quadrupole Resonance (NQR) targeting the ^{14}N nuclei in the objects under examination), and infrared radiation (IR) (passive) and Raman (IR) (active) systems, which operate in the wavelength range of 600–1,000 nm. The latter can be used to examine samples taken from passengers' effects and other items of interest.

Surveillance by *Closed Circuit Television* (*CCTV*) *systems*, whether by optical or IR cameras, is ubiquitous in modern society. Used especially with facial recognition software these provide intelligence on individuals in crowds to security operatives and police. (Facial recognition software has been employed at many major sporting events, for example, the Superbowl XXXV,

FL, USA (2001), and in community policing in the Borough of Newham, UK. The Australian Customs Service uses facial recognition in its new Smartcheck system (2008). The Information Access Division of the US National Institute for Science and Technology (NIST) has been conducting public trials in this field since 2002.) While important to the maintenance of a secure environment these systems will not be discussed here.

Systems commonly used for the electronic detection of trace (ETD) amounts of illicit substances in the solid, liquid, or vapor phases use time-of-flight mass spectrometry (*ion mobility spectrometers* (*IMS*)) to identify components that comprise the substances. Other systems using vibrating elements on which the substance is deposited by evaporation behave as microbalances to determine the molecular weight of the substances present.

The detection of improvised explosive devices may involve the use of *nonlinear junction detector* (http://www.pimall.com/) systems to identify the presence of nonohmic circuits and/or passive devices that listen for timing signals and other electronic emissions from objects under examination.

Chemical systems require the taking of samples and are specific to the class of illicit substance targeted. Different chemicals are used for the indication of the presence of drugs, nitrogenous explosives, and peroxide-based explosives, for example. A detailed discussion of the specific devices and chemical reactions that occur in them will not be given here.

It must be stressed that the taking of samples must not take place in uncontrolled conditions.

And this must not occur in the presence of passengers or anyone not directly connected with the examination.

Protocols have been established for the testing, both initially on acceptance, and in operation, of all items of testing equipment. These will be discussed later.

In the sections that follow, I shall briefly describe the characteristics of each class of equipment used in security applications, discuss the principles under which they operate, and indicate the strengths and weaknesses of the systems.

2.1 Passenger Portals

It is necessary to address the questions of why passengers are screened and what regulations exist worldwide with respect to the passage of people through border checkpoints. The principal aim is to prevent the carriage of items that may cause harm to the aircraft (explosives), harm to passengers and crew (guns, knives, and other offensive weapons), and contraband (narcotics, birds, reptiles, etc.).

Regulations that are commonly in place require that the passenger: removes all loose clothing, belts with metal buckles, electronic devices (cell phones, iPods, cameras, etc.), chunky jewelry, shoes, head covering, and wallets. This means that, when passengers present themselves at portals, they are declaring that they have nothing suspicious about their person. Technology therefore has as its aim the detection of undeclared items.

In the past "pat downs" have been used, and they remain a simple and effective alternative to the use of technology. There have been, however, objections citing passenger privacy. Hence, they do not involve touching of the groin and anal areas: areas in which concealment is possible. This limits the effectiveness of "pat downs." Another concern is that some screeners find repeated "pat downs" distasteful.

2.1.1 Passive Passenger Portals

Passive passenger portals rely on measuring the blackbody radiation from the incoming passenger and the variations in spectral emissivity of the item carried from that of the passenger (the Stephan–Boltzmann law: energy flux density = $J = \varepsilon\sigma T^4$, where: ε emissivity $(0 < \varepsilon < 1)$, $\sigma = 5.7604 \times 10^{-8}$ W m^{-2} K^{-4}, T temperature (K)). Currently available systems are sensitive to radiation from the person in the mm-wave (~100 GHz) or the THz (~250 GHz) regions. The design and construction of mm-wave and THz detectors is a rapidly growing field (Knoll 2000; Rieke 2003; Lee 2009). A wide range of multi-pixel semiconductor devices are being developed to operate at room temperature. These will be able to be incorporated in systems using the technology available in digital processing. Existing passenger portals tend to use only a line of detectors, and the image of the subject is scanned across this line onto the detector array using a tilting mirror (*Fig. 1*).

Several images are available to the operator: the Closed Circuit Television (CCTV) image, the raw image presented by the detector, and enhanced images (edge enhancement, for example) (*Fig. 2*).

These systems see only one side of the passenger. With all passenger portals privacy concerns exist: civil liberties groups insist that "see through" technologies violate the rights of the individual and contravene national and international laws. For these systems, the resolution is so poor that it seems unlikely that the concerns can be justified. Either two systems have to be deployed so that the front and back of the passenger can be viewed simultaneously, or the passenger has to turn around so that a second image can be taken.

Advantages of these systems are that: measurements can be made at a distance (stand-off capability); observations can be made while the subject is approaching the portal; objects at temperatures different from the body or having different emissivities will provide an image; the system can see what is hidden under a moderately thick layer of clothing; and examination

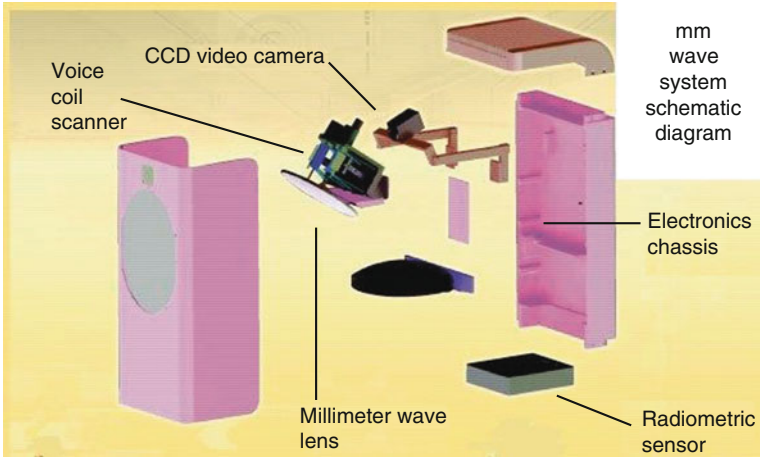


Fig. 1
Exploded schematic view of a mm-wave passenger scanner



Fig. 2
CCTV and enhanced images of a passenger with a wallet in his hip pocket

is rapid. (Screening authorities, such as those in airports, are very concerned with "throughput," the number of passengers per hour passing the portal. As well, they are concerned with false alarm rates, both positive (there is a threat) and negative (there is no threat). Both have a deleterious effect on throughput.)

Disadvantages include: lack of materials discrimination; poor resolution; inability to detect items in cavities and crevices; inability to detect objects in the foot and ankle region; and poor detection of objects not in contact with the body.

2.1.2 Active mm-Wave Passenger Portal Technology

Active passenger portals consist of a mm-wave or THz generator and antenna system and a corresponding detector system. These are scanned over the subject, and an image corresponding to the point-to-point reflectivity of the subject is formed. In a sense, this is a RADAR map of the subject.

For portals, systems that are large enough to encompass a human body, it is necessary for the system to be enclosed in a cabinet because of radiation licensing regulations. (The electromagnetic spectrum is divided into bands that are allocated to corporations and other entities by the relevant Broadcast authority. Enclosing the system minimizes the problem of encroaching on another user's band.) These systems do not have a "stand-off" capability.

If no portal is available, the passenger may be manually scanned using a handheld device. Because of their low power and the fact that they pass close to the subject the problem of encroaching on other bands in the electromagnetic spectrum may be considered by the licensing authority to be of little significance.

Figure 3 shows a cabinet passenger mm-wave scanning system.

As well, an image of a passenger is shown. Note that the outline of the image is a *computer-generated anthropomorphic shape*. Some details of his clothing are evident. A circular shape has been placed around the genital area to ensure privacy. The operator can, however, turn off this privacy filter: but he has no visual contact with the subject. The passenger is rotated through 180° to ensure that both front and rear images are produced. Three square objects of different sizes can be seen. The extremities of the passenger are not imaged.

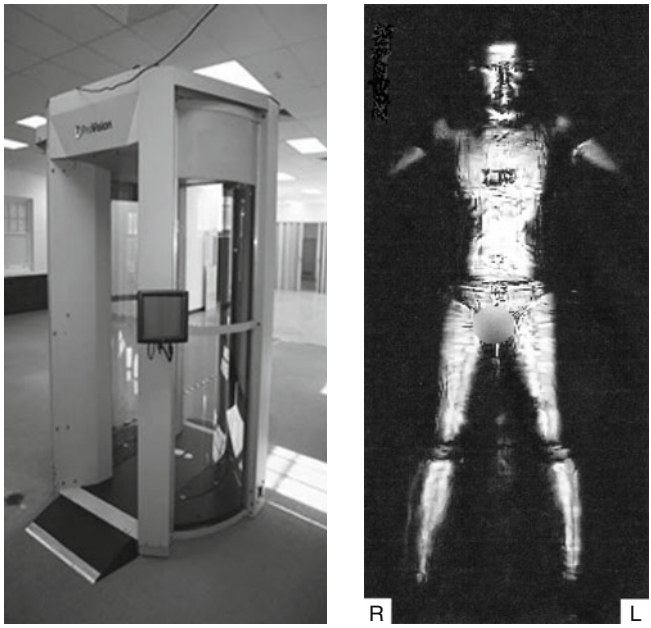


Fig. 3
Cabinet mm-wave passenger scanner (*left*) and an image of a subject taken with the system (*right*) are shown

Advantages of this system are: mm waves have no deleterious interactions with people; the spatial resolution is 2 mm; the throughput is acceptable (~600 subjects/h); the system maps the surface of the passenger, illicit items being identified by their shape and contrast; and the passenger's clothing has only a small effect on the image.

Disadvantages include: lack of materials discrimination; inability to detect items in cavities and crevices, and in the lower extremities; sensitivity to metallic items; and poor detection of objects not in contact with the body.

Wands are handheld devices and can be directed wherever the operator wishes. Their close coupling to the surface enables a stimulated emission spectrum. (Every assemblage of atoms has vibration/rotation bands associated with particular radicals (molecules) or quantized vibrational states (condensed matter). The incoming THz photon pulse excites ground-state electrons in the system under investigation to higher excited states. In relaxing to the ground state photons are emitted corresponding to the energy difference between the levels: $E_{\text{upper}} - E_{\text{ground}} = hc/\lambda$. The spectral lines are characteristic of the material examined.) Comparison of the spectrum with entries in a spectral database enables materials identification.

2.1.3 Active Devices: Metal Detectors

In its simplest form, a *metal detector* consists of an oscillator producing an alternating current that passes through a coil producing an alternating magnetic field. Electrically conductive metal

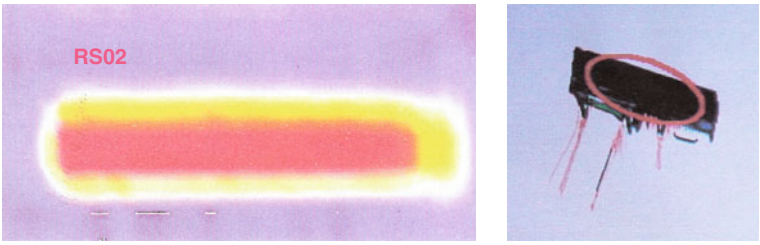


Fig. 4
Metal detector image (*left*) and X-ray image (*right*) of a Glock pistol concealed in a postal parcel. Only the barrel of the pistol is metal

close to the coil creates eddy currents in the metal, and this produces an alternating magnetic field. Another coil can be used to detect the absorption or retransmission of the radiation due to the presence of the metal. The frequency of the oscillator can be tuned to indicate the presence of particular metals, the most common being gun metal and steel. Arrays of oscillator/detector coils can be manufactured and crude imaging is possible.

Metal detectors can be portals (and systems can have a number of oscillators tuned to different frequencies), hand-held devices (tuned to a fixed frequency range), or tunnel devices, used (for example, in the examination of parcel post for firearms). *Figure 4* shows the image of the barrel of a Glock pistol concealed in a parcel with other metallic items, and beside it, an X-ray image of it, with a threat ellipse generated by the metal detector.

The principal advantage of metal detectors is that they provide a relatively inexpensive and rapid assessment of whether the passenger is carrying threat items.

The disadvantages are that: the systems are relatively insensitive to the type of metal carried, with the result that false positives occur regularly, especially if the passenger has metal prostheses and implants; they can have an adverse effect on heart pacemakers.

2.1.4 Active Devices: X-ray Passenger Portals

X-rays are ionizing radiations that can cause harm to passengers if not used correctly (see Chap. 10, "Radiation Protection.") Their use is, therefore, closely regulated by Government instrumentalities. Used in an *X-ray passenger portal* context, the X-ray energies lie in the range of 50–160 keV. The interactions of radiation with matter are summarized in *Fig. 5*. This shows the cross sections for different scattering processes by atoms, in this case, carbon. In practice the systems are either transmission systems (the X-ray passes through the subject to the detector) (absorption follows the Beer–Lambert law: $I = I_0 \exp(-\mu_l t)$, where I_0 is the incident energy, t the path length in the material, μ_l the linear absorption coefficient equal to $\mu_m \rho$ with μ_m the mass absorption coefficient and ρ the density, $\sigma = (A/N_A)\mu_m$ with A the atomic weight and N_A being Avogadro's number) or backscattered (the X-ray is scattered back toward the detector).

The dominant scattering processes are *photoelectric* (σ_{pe}) and *backscattered* (*Compton*) (σ_{inel}). Because the photon energy changes as a function of scattering angle in Compton

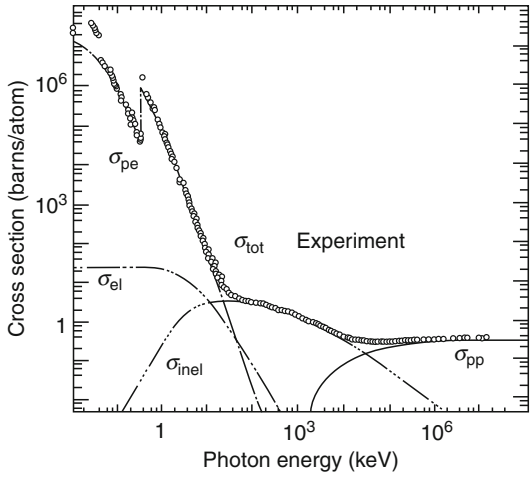


Fig. 5
Scattering of photons by free atoms: carbon

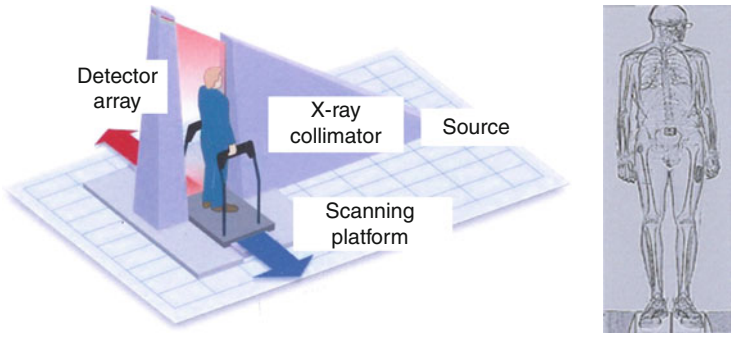


Fig. 6
Schematic diagram of a transmission X-ray portal (*left*), and a representative image of a passenger (*right*)

scattering, the backscattered photons have energies of ~0.83 of the incident photons in the 50–160 keV energy range $(E_c/E_0 = 1/[1 + (E_0/mc^2)(1 - \cos\theta)])$ (see Chap. 1, "Interactions of Particles and Radiation with Matter.")

Figure 6 shows a schematic diagram of a transmission X-ray system currently in use. The X-ray source has a maximum energy of 160 keV. The X-rays are collimated into a fan-shaped beam and detected using a linear array of scintillation detectors comprising ~1,000 elements. The subject is translated across the fan beam at a constant speed on a scanning platform. On the right, a representative image is shown.

Advantages of this system are that a complete scan of the subject is made, and the contents of vaginal, rectal, and abdominal cavities can be made. (It is standard practice in many countries

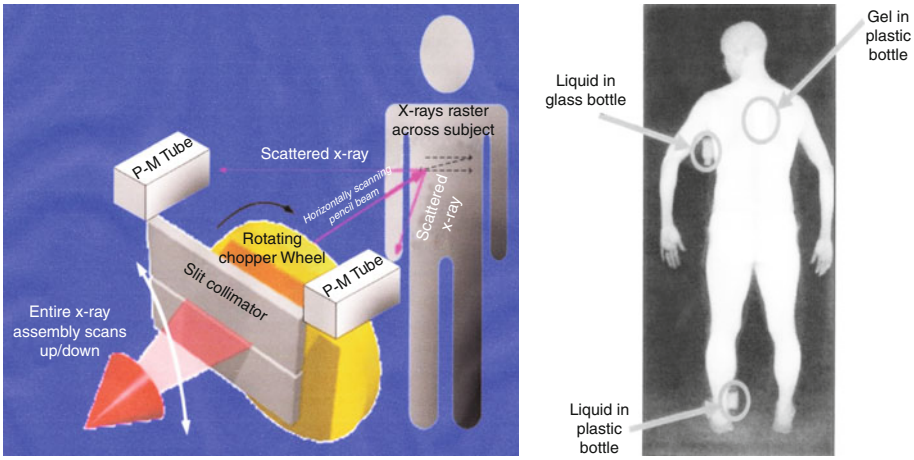


Fig. 7

Schematic diagram of a backscatter X-ray portal (*left*) and an image (*right*) of a subject carrying concealments

for border agencies to order full-body X-ray examinations for passengers suspected of carrying drugs in body cavities.)

Disadvantages of the system are that: it has low *throughput* (10 passengers/h); relatively high doses are delivered to vital organs (~3 μSv/scan); the skeletal image can obscure areas of interest; and there is no materials discrimination, although the contrast increases as the atomic number of the target material increases.

Systems using backscattered radiation operate on a somewhat different system. A collimation system is used to create a fan beam and a rotating slit system is used to form a pencil beam, which moves horizontally across the subject. The X-ray tube and rotating slit system is then translated downward so that the flying spot completely traverses the subject. The X-rays scattered back from the subject are collected on large sheets of scintillation material (often gadolinium oxysulphide), one on each side of the plane of rotation of the collimator system. Only two scintillation detectors are required. The data is assembled for display in much the same way as TV images. *Figure 7* shows a schematic diagram of such a system, and an image taken with it of a subject with concealments attached to his person. The passenger has to be scanned twice so that both his front and back can be examined. If two systems were to be used the throughput could be doubled. The operator uses contrast and shape variations to identify anomalies. As with all security applications, detection depends on the alertness and training of the operator.

Advantages of the backscatter type of portal compared to the transmission type are that they: operate at lower voltage (100 keV); do not penetrate deeply into the subject; are not affected by skeletal artifacts; deliver a lower dose per scan (0.1 μSv); have a greater *throughput* (60–100 passengers/h); and are sensitive to low-atomic-weight materials, such as in liquids and gels.

Disadvantages include: the inability to inspect cavities and crevices; poor materials discrimination; and the contrast observed depends on both the material of the object and the material behind it.

2.2 X-ray Baggage, Pallet, and Container Systems

The simplest X-ray systems used in baggage, pallet (and air cargo), and container examination consist of the following components: an X-ray source, a slit system that forms a fan beam, a detection tunnel that incorporates a means for translating the object under examination through the fan beam, a detector array that measures the changes in intensity of the X-ray beam as it passes through the X-ray fan beam, and a system for accumulating, storing, and manipulating the data into an image that can be assessed by an operator.

X-ray systems fall into three categories: those for baggage (tunnel size $<1{,}200\,\text{mm} \times 1{,}200\,\text{mm}$), pallet ($(1{,}200\,\text{mm})^2 < \text{H} \times \text{W} < (2{,}000\,\text{mm})^2$), and shipping container ($\text{H} \times \text{W} > (2{,}000\,\text{mm})^2$) applications. For tunnel sizes less than $(2{,}000\,\text{mm})^2$, *sealed X-ray tubes* are the preferred source. For shipping containers the preferred source is the *linear accelerator (LINAC)*. Both sources are sealed, highly evacuated systems in which electrons from a heated filament are accelerated by a high voltage and collide with a target (usually tungsten). The resultant change in energy causes the emission of electromagnetic radiation (bremsstrahlung). ("Bremsstrahlung" means, literally, breaking radiation. The maximum frequency is given by the energy $E_0 = eV_0 = hf = hc/\lambda$, where V_0 is the maximum applied voltage, e is the elementary charge, and h is Planck's constant. This is for a single interaction. In practice the electron comes to rest through a number of interactions with a lesser energy change, and, for each energy change, photons are produced giving rise to the spectra seen above.) For passenger baggage, V_0 is usually 160 kV; for pallet, 450 kV; and for containers, 3–9 MV. A schematic diagram of a sealed X-ray tube is shown in ❯ *Fig. 8*, together with intensity-versus-wavelength *bremsstrahlung* curves for different values of V_0.

LINAC sources are effectively loaded coaxial waveguides along the axis of which electrons generated thermionically in a filament are accelerated by an axially propagating wavefield. See ❯ Chap. 7, "Accelerators for Particle Physics" and ❯ Chap. 35, "Radiation-Based Medical Imaging Techniques: An Overview." While sealed-tube X-ray systems usually operate as constant-potential devices, LINACs are pulsed devices, with the pulse rate determined by the speed of transport of the object past the X-ray fan beam. The target for the electrons is typically a water-cooled 1 mm thick gold-plated tungsten disk. A representative spectrum from a 2.5 MeV LINAC is shown in ❯ *Fig. 9*.

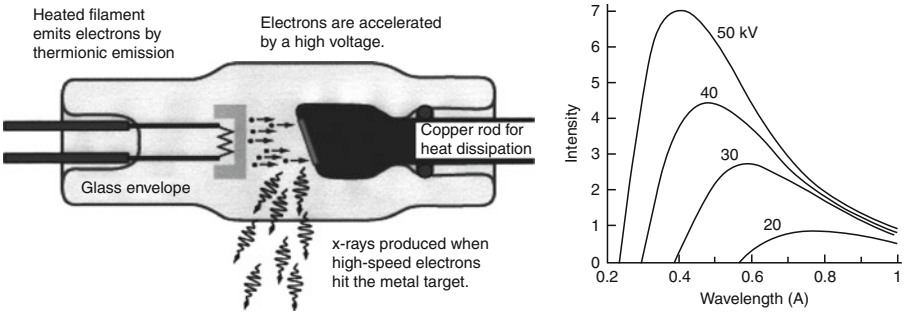


Fig. 8

Schematic diagram of a sealed-tube X-ray source (*left*) and bremsstrahlung intensity plotted against wavelength as a function of applied voltage (*right*)

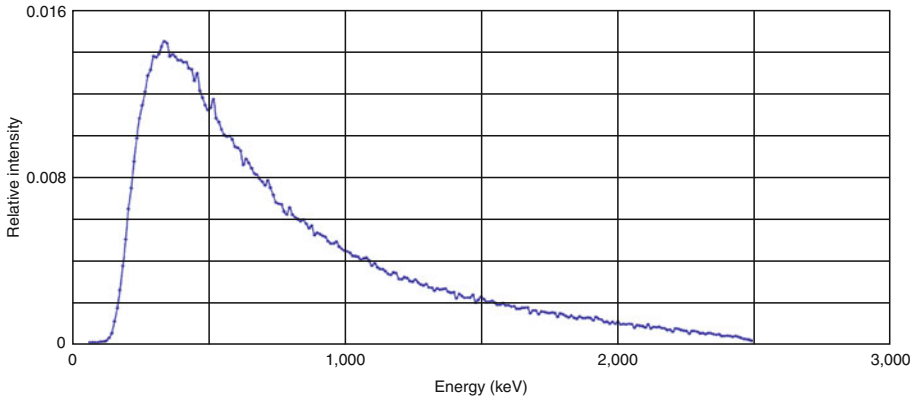


Fig. 9
Emission spectrum from a 2.5 MeV LINAC

The method of transport of the object past the object varies from application to application.

For baggage and pallet applications the source and detector systems remain stationary and the object moves through the fan beam on a conveyor belt or a driven roller system. For container systems the container is either towed through the fan beam or the container remains stationary, and the source and detector are scanned along the container. Typical scan rates are 200 mm/s.

Modern detector systems usually comprise arrays of scintillation detectors. Scintillation materials used include cesium iodide (CsI), cadmium tungstates ($CdWO_4$), and mercury cadmium telluride (HgCdTe). The scintillations caused by the incident photons are detected by solid state detectors such as avalanche photodiodes (APD). (See Chap. 15, "Scintillation Counters.") The arrays usually comprise more than 1,000 separate elements, which are grouped in units of 32 elements. Each element presents an active area of around 6 mm × 6 mm or less to the beam. The output of each of the detectors is interrogated using multiplexing techniques, and the ratio of that output is taken with the output of elements that were not impeded by the object (giving the ratio I/I_0, from which the average attenuation coefficient can be computed). (If only one type of material is present in the X-ray path length and its thickness is known, the μ_1 of the material can be calculated and compared with listings in handbooks. For a path length comprising different elements the Beer–Lambert law takes the form:

$$\frac{I}{I_0} = \exp\{-((\mu_1)_1 t_1 + (\mu_1)_2 t_2 + (\mu_1)_3 t_3 \ldots)\}.$$

Note that μ_1 can itself be either that of an atom or a chemical compound.) Each time slice forms a linear pixel array, and the overall image is the outcome of linking these time slices and displaying the result on a computer monitor. See Chap. 16, "Semiconductor Counters."

Each company has its own set of algorithms for enhancing the image. Techniques may include contrast and brightness adjustment, edge enhancement, histogramming, and so on. These algorithms are not unlike those available in photographic programs in personal computers.

Some degree of *materials discrimination* is possible if measurements along the same path length can be made at different photon energies. As can be seen in *Figs. 8* and *9*, the energy spectra are continuous spectra. To a good approximation it is possible to establish a mean value

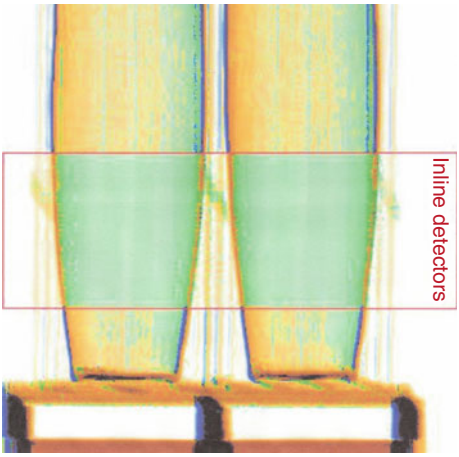


Fig. 10
Images of two urns taken using dual and inline X-ray detectors

for each of the spectra, and use these in calculations. *Figure 5* shows that the *photoelectric scattering* factor and the *Compton scattering* factor are very different in the energy range used in baggage and pallet systems. If the mean energy is changed, therefore, the value of (I/I_0) will be changed, and a new value of μ_1 will result. From the simultaneous equation it is possible to determine μ_1 and t for a two-component system.

The variation of mean energy can be effected by actually using two separate X-ray source energies (switching from 160 to 80 kV at a rate of, say, 200 Hz), or by having two separate identical detector arrays with an absorber placed in front of one of the detector arrays. Systems of these types are referred to as *Dual Energy Systems*. Of these the former is better because there is no spatial difference in the images. The effect of spatial separation on image resolution can be seen in *Fig. 10*, in which a composite single energy bank of 32 detectors (CsI in front of GdOS) is placed in a dual array system. The resolution of the inline system is superior to, and the signal-to-noise ratio of the inline system is better than, the conventional dual system by about an order of magnitude. The dynamic range is 2^{12} or 12 gray scales.

If more than one X-ray beam (referred to as "views") is used to scan the sample, the number of simultaneous equations for which solutions can be found increases, and for μ_1 and t can be found for increasingly more complicated packing in the scanned object. *Dual view, triple view, and quadruple view* systems are marketed at present. *Computed tomography* (CT) systems can give as many views as the manufacturer may wish. (In a conventional CT system, the source and detector rotate around the axis lying along the conveyor belt. As the object passes through the system, the number of images taken per revolution depends on what is considered to be an acceptable throughput.)

The image processing time determines the throughput, and this is a major determinant in the operation of airports, in both the passenger and air cargo contexts. A compromise has to be made between certainty of detection and throughput. High *throughput* causes increased false positive and false negative rates.

The additional data available in multi-view examination systems makes it possible to calculate the *effective atomic number*, $Z_{\text{eff}} = \sum_i f_i A_i (\mu_m)_i / \sum_j f_j (A_j/Z_j)(\mu_m)_j$, where f is the molar fraction, and A and Z are the atomic weight and the atomic number of the atomic species (Manohara et al. 2008), and the density ρ for any voxel (volume element) in the object. Z_{eff} and

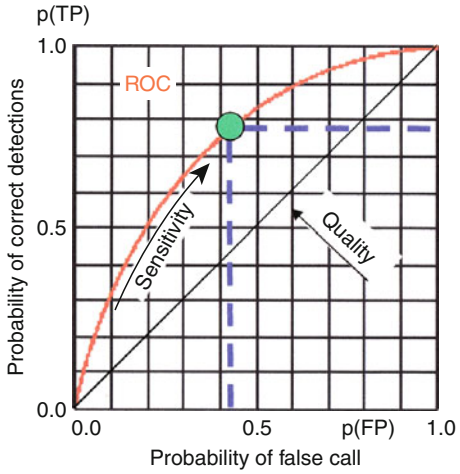


Fig. 11

Idealized Receiver Operator Characteristic (ROC) curves. The diagonal line represents the results if the operator simply guesses

ρ can be used as a discriminant between materials. For metals: $Z_{\text{eff}} > 19$. For inorganic materials: $19 > Z_{\text{eff}} > 11$. For organic materials: $11 > Z_{\text{eff}}$. Most explosive materials have parameters in the range: $8.3 > Z_{\text{eff}} > 6.5$ and $0.9 < \rho < 1.8\,\text{g/ml}$. If an object has a region that corresponds to the explosives range, there is reason to proceed to further examination of the baggage by other means.

Some X-ray examination systems are accredited by the US TSA as *explosives detection systems (EDS)*. These purport to make the assessment as to whether the object contains an explosive threat without human intervention. In practice, however, confirmation by an operator on the basis of shape and contrast is required. The performance of X-ray examination systems is usually described in terms of a *Receiver Operator Characteristic (ROC) diagram* (Fig. 11) (Green and Smets 1966).

Materials discrimination using Z_{eff} and ρ can lead to reasonably high throughputs, but the systems are prone to false positives being recorded because innocuous materials may well lie in the threat band (for example, water, chocolate). To overcome this deficiency, EDS systems can be used as a primary screening device, and questionable objects diverted into a system that has a much smaller *throughput*, but which can give a spectroscopic analysis of the questionable area of the object.

Such a system uses energy-dispersive X-ray diffraction (EDXRD) (Knoll 2000). (Most systems are crystalline, and elastic scattering occurs according to Bragg's law: $2\,d_{\text{hkl}} \sin\theta = \lambda = hc/E$. Because θ is fixed, diffraction peaks occur whenever a photon energy corresponds to a particular interplanar spacing, d_{hkl}.)

In *EDXRD systems*, the tube source (which is a bremsstrahlung source [see Fig. 9]), its collimator system, and an energy-dispersive X-ray detector are mounted so that the beam emitted from the source can pass through the suspect region of the object and be scattered into the detector, which is offset by an angle of about 5°. The elastically scattered photons from the suspect region interact with the detector giving output pulses proportional to these photon energies. And peaks occur in the spectrum that results when intensity at a particular energy is plotted against energy (Fig. 12) (Harding 2009). This shows the diffraction patterns

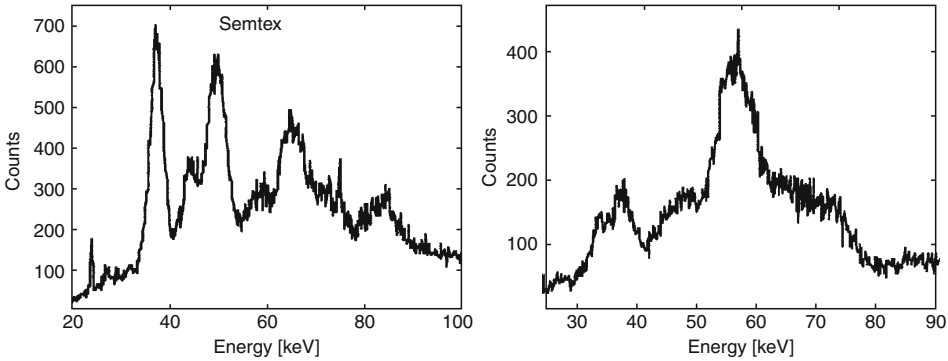


Fig. 12
Diffraction patterns for Semtex (76% PETN + 5% RDX) (*left*), chocolate (*right*)

for an explosive, Semtex ($C_3H_8N_4O_{12}(76\%) + C_3H_6N_6O_6(4.6\%)$), and chocolate. These have clearly different diffraction patterns that can be rapidly compared with entries in the JCPDS database.

The International Commission on Diffraction Data maintains the JCPDS files, which include some 500,000 entries (http://www.icdd.com). Recent developments in detector design and improvements in system sensitivity have made it possible to identify some amorphous materials and liquids as well as crystalline materials.

A system has been developed for the *X-ray and neutron examination* of air cargo ships and shipping containers (Eberhardt et al. 2005). A commercial realization of this system uses a switched 6/3 MV LINAC X-ray source with two collimators that produce two vertical fan beams at an angle of 12° to one another. The system has two array detector systems, one for each fan beam. The objects are towed through the fan beams at a constant scan rate of 200 mm/s, and conventional images are taken at 3 and 6 MeV energies in each detector system. The resulting images can be combined to produce a binocular view through the object, and the switched voltage source can give crude materials discrimination. Incorporated in the system is a conventional sealed-tube neutron generator ($^2H + {}^3H \rightarrow {}^4He + n$) that produces 10^{10} fast neutrons per second of energy ~14 MeV. This is collimated to produce a vertical fan beam which, after passing through the object, is detected by an array detector. The attenuation of the X-ray beams and the neutron beams are measured, and the output of pixels from the two arrays are mapped to form the ratio, $R(= \ln(I_0/I)_{\text{neutron}}/\ln(I_0/I)_{\text{X-ray}})$. While the X-ray attenuation cross sections are regular functions of the atomic number, the neutron cross sections are not. The ratio R can be determined for a given element and is a characteristic of that element. For nitrogenous explosive materials R lies in the range $1.18 < R < 1.28$. However, as in the case of Z_{eff}, some non-threat materials have R-values in this range. The system ultimately relies on the skill of the operator to interpret the high-resolution image seen.

2.2.1 X-ray Baggage Systems

The performance of two extreme configurations of *X-ray baggage scanners* will be described. The first contains a conventional single view dual energy system as well as a separate system that provides a backscattered image. The second is a CT system. The object is a clock radio in

which there is a concealment of TNT (Fig. 13). Note that these images would usually be in color, and the threat material would have been colored orange. The left-hand image is from a conventional system. The center image is a backscattered image. The right-hand image was taken with a CT system. The thumbnail images on the right show the shape of the TNT block and lines link these with the position of the threat. As well, the value of Z_{eff} is displayed on the monitor screen (in this case, 7.42, which corresponds to the value for TNT).

The CT system produces a clearer image, and provides the operator with the value of Z_{eff} at any location at which he places his mouse cursor. This precision is gained at the expense of *throughput*.

2.2.2 X-ray Pallet and Air Cargo Systems

These are usually dual view dual energy X-ray systems operating at typically 450 kV. The tunnel size is typically (2 m × 2 m). The images formed are similar to those shown in Fig. 13. Shown in Fig. 14 are images taken with CT air cargo scanner. In this system the source and collimator form a horizontal fan beam, which illuminates a horizontal detector array. These move together vertically. The air cargo unit load device (ULD) is loaded onto a rotating table. Initially two scans are made: as loaded, and the ULD rotated through 90°. This gives a conventional dual view presentation. If problems in interpretation exist the operator can take CT slices at the suspicious locations. Figure 14 shows one projection of the image of a ULD containing a mixed load. The left-hand image is a CT slice through cartons of predominantly oranges and meat. The cylindrical bright object near the top left of the image is a 600 ml bottle of hydrogen peroxide. The detail in the image is very good: it is possible to count the pips in the oranges, for example.

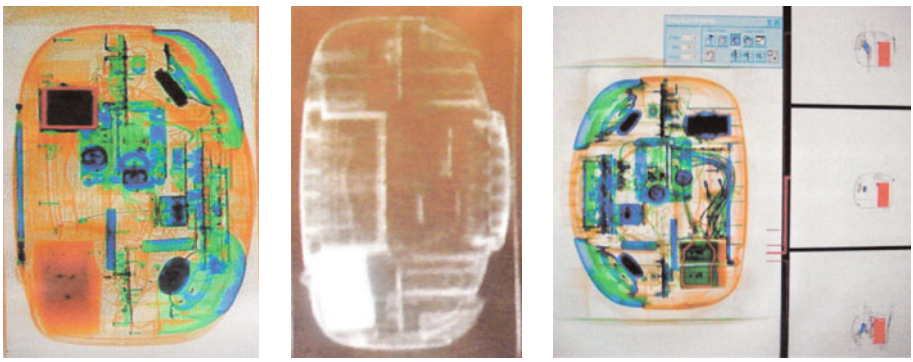


Fig. 13
TNT concealment: rectangular shape in top left of *left image*; white rectangle at the left-hand corner of the *central image*; rectangular shape at the bottom right of the *right image* (and the corresponding thumbnail images)

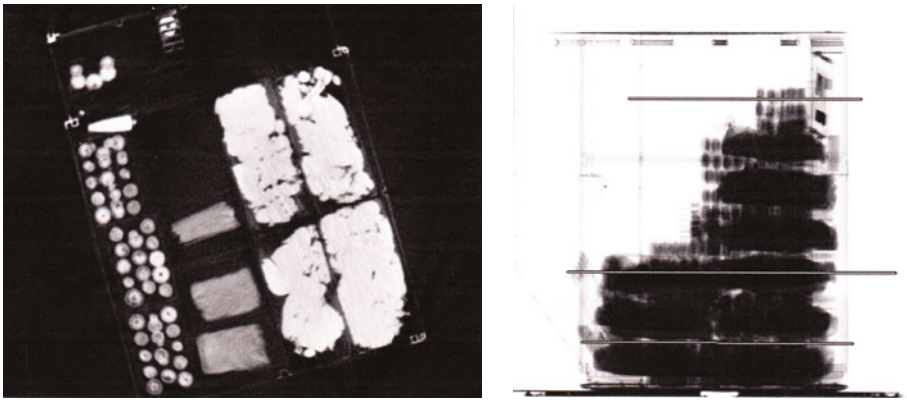


Fig. 14
The image (*right*) is the 0° projection. The image (*left*) is a section through the ULD showing cartons of oranges, meat, and a bottle of H_2O_2

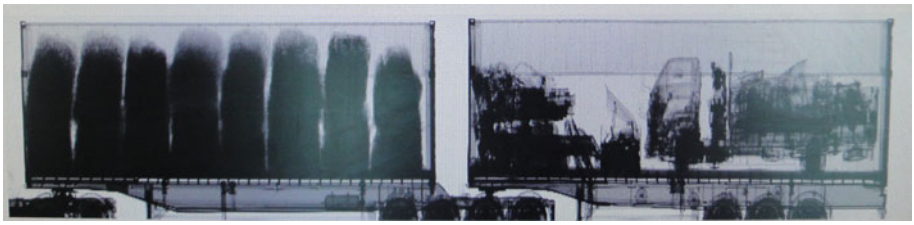


Fig. 15
Image of two containers taken in one scan of the gantry (360 s)

2.2.3 X-ray Shipping Container Examination Systems

X-ray shipping container examination systems are usually gantry systems and use single view LINAC sources. Dual view systems exist, but most large systems are single view systems. Tunnel dimensions are such that shipping containers are able to be examined (typically 4 m × 4 m). The gantry, which contains the source on one side and the detector array on the other side, can be stationary (the container is towed through the gantry, or a truck is driven through it) or moving (the gantry moves across the container at a speed of ~200 mm/s). *Figure 15* shows a typical image taken with a moving gantry system.

Identification of anomalies is effected by shape and contrast resolution.

3 Ancillary Technologies

A need exists to test items that fail the X-ray testing in the passenger context or are related to spills in the cargo context. A wide range of technology exists to undertake this testing.

Figure 16 shows some equipment commonly used in the airports and cargo environments. Brief mention of the functions of these systems will be given below.

Airline passengers and their personal items are subjected in many jurisdictions to a "swipe test" in which the operator passes a wand carrying a pad over them and then places the pad into a device that then gives an indication (usually in the form of a spectrum on a display), which indicates whether either the person or the object is a potential threat.

These systems are referred to as *Ion Mobility Spectrometers* (IMSes) in which the pad from the wand is ablated, the combustion products ionized, and the ions accelerated along a flight tube to a detector. The time of arrival of the ionized fractions is determined by the mass of the fraction. A spectrum is constructed of ion mass versus time of flight. Systems carry listings of spectra taken from a wide range of threat and non-threat materials, and the "swipe" is compared with the spectra in this library. Most IMS systems can be used as *vapor trace detectors*. Instead of using the swipe pad a sample of known volume is pumped from an enclosed space in which suspect materials might be present. This sample is ionized and then led into the IMS accelerating tube.

An alternative technology to IMS exists: *biosensing*. The products of chemical extraction of the contents of a swipe pad are introduced into the system that contains many vibrating microelements coated with antibodies, which react with the suspect chemicals. The frequency loss is then used to identify the molecules present in the sample. An adaptor can be used to detect vapors in, say, closed passenger baggage. The effectiveness of these systems is limited by the manner in which the sample was taken, the resolution of the system, and the comprehensiveness of the spectral libraries.

Systems that used either the absorption or the *stimulated emission of IR radiation* exist. Systems using the *Raman effect* (Gardiner 1989) are the most common. The object is irradiated with a laser beam (~833 nm). The emitted light falls on a grating spectrometer, the spectral lines being detected using a linear array detector. The spectra are then compared with those in an internal database. The system performs best when examining light-colored loose powders, liquid spills, and the contents of sealed translucent bottles. Dark colored materials absorb the IR radiation too strongly, and the beam cannot penetrate opaque bottles.

The most reliable system for scanning bottles, cans, and other objects is an *X-ray CT system*, the geometry of which is similar to that of the CT air cargo scanner described earlier. Images from this washing-machine-sized system give detail similar to that of *Fig. 14*. Because a dual energy source is used, values of Z_{eff} and ρ can be found for any type of concealment in a passenger's effects, including multiple concealments in an opaque metallic container, up to a mass of 1 kg.



Fig. 16
(*Left to right*): Ion mobility spectrometer; vapor trace detector; IR Raman spectrometer; X-ray bottle scanner

4 Protocols: Passenger, Air Cargo/Pallet, Shipping Containers

It is imperative that equipment is properly designed and maintained, and that operators are properly screened and trained. *Protocols* for the proper maintenance and testing of all the systems exist (American Society for the Testing of Materials: http://www.astm.org; American National Standards Institute: http://www.ansi.org; Australian Nuclear Science and Technology Organization: http://www.ansto.gov.au).

5 Conclusions

Over the past decade, there have been many advances in the technology associated with border protection. The principles of operation of the systems outlined in this chapter will not change much in the next decade, but the detailed construction of systems and their operator interfaces will undergo constant change.

One area in which advances may be made in the near future is in the development of neutron systems for the imaging of baggage and air cargo. (See ❯ Chap. 33, "The Use of Neutron Technology in Archaeological and Cultural Heritage Research.")

To cope with increasing passenger and freight traffic, improvements have been made to simplify the presentation of information to the operators at checkpoints. It must be stressed that, with the increasing complexity of examination systems comes the risk of system failure due to incompetent maintenance strategies. Risks are involved associated with inappropriate management practices such as the acquisition of systems that are not fit-for-purpose and the failure to employ operators with appropriate skills.

The human dimension must not be neglected: operators must be properly security screened, have the innate skills necessary for image interpretation, be properly trained, and be motivated to perform their tasks consistently at the highest level.

Acknowledgments

This chapter is a summary of my work in the field of Border Technology and Border Protection during the past 15 years. Many have helped me. To them I am grateful for their friendship and advice. And I give my heartfelt thanks to them for their generous support.

References

Eberhardt JE, Rainey S, Stevens RJ, Sowerby BD, Tickner JR (2005) Fast neutron radiography scanner for the detection of contraband in air cargo. Appl Rad Isot 63:179–188

Gardiner DJ (1989) Practical Raman spectroscopy. Springer, New York

Green DM, Smets JM (1966) Signal detection theory and psycophysics. Wiley, New York

Harding G (2009) Systems and methods forusing a crystallinity of a substance to identify the substance, US Patent 751954

Knoll GF (2000) Radiation detection measurement, 3 edn. Wiley, New York

Lee YS (2009) Principles of THz science and technology. Springer, New York

Manohara SR, Hanagodimath SM, Thind KS, Gerward L (2008) On the effective atomic number and electron density: for all types of materials and energies above 1 keV. Nucl Instr Meth B266:3906–3912

Rieke G (2003) Detection of light: from the ultraviolet to the sub-millimeter, 2 edn. Cambridge University Press, Cambridge

26 Accelerator Mass Spectrometry and its Applications in Archaeology, Geology, and Environmental Research

Wolfgang Kretschmer
Physikalisches Institut, Universität Erlangen-Nürnberg, Erlangen, Germany

1 Introduction ***654***

2 The Methodology of Accelerator Mass Spectrometry ***654***
2.1 Typical AMS Set-up 655
2.2 Determination of Calendar Age 657
2.3 Sample Preparation 657
2.3.1 Pretreatment of Sediment Samples 657
2.3.2 Pretreatment of Bones 658
2.3.3 Pretreatment of Archaeological Samples 658
2.3.4 Combustion to CO_2 and Reduction to Carbon 659

3 Applications of Radiocarbon Measurements to Interdisciplinary Research ***660***
3.1 Sediment Dating 660
3.2 Environmental Studies 660
3.3 Archaeological Samples 662

4 Concluding Remarks ***665***

References ***665***

C. Grupen, I. Buvat (eds.), *Handbook of Particle Detection and Imaging*,
DOI 10.1007/978-3-642-13271-1_26, © Springer-Verlag Berlin Heidelberg 2012

Abstract: Accelerator Mass Spectrometry (AMS) is an ultrasensitive method for the measurement of isotope ratios in the range of 10^{-12}–10^{-15}. Most frequently the $^{14}C/^{12}C$ ratio from biogenic samples is determined which gives information on the age of the sample of up to 50 ka with a precision of typically 40–80 years. In this paper the radiocarbon method is discussed and various applications to interdisciplinary research are presented. One application at the Erlangen AMS facility is the ^{14}C dating of sediment samples which together with simultaneous pollen analyses can establish a better chronology of climate and vegetation during Holocene in Germany. For an enhanced reliability of sediment dating different fractions like bulk sediments, pollen grains, macrofossils, and humic acids have been measured. For environmental research the ^{14}C content of aldehydes from indoor air samples can be used to disentangle the anthropogenic or biogenic origin of these compounds. Finally interesting archaeological samples from a Persian mummy are discussed.

1 Introduction

The radionuclide ^{14}C is produced in the atmosphere via the interaction of cosmic radiation with nitrogen with a nearly constant rate and decays weakly with a half-life of 5,730 years. Due to this interplay of production and decay an equilibrium concentration of ^{14}C in the atmosphere is established with an isotope ratio $^{14}C/^{12}C$ of about 10^{-12}. Radiocarbon forms $^{14}CO_2$ molecules and participates on the carbon bio-cycle similar to the stable carbon isotopes with the result that, except for some minor corrections, it is present in all living creatures and plants in equilibrium concentration. If the ^{14}C intake is discontinued, e.g., by death of the organism, its concentration decreases with the half-life mentioned above. This fact can be used for an absolute dating of carbon-containing samples by measuring either the specific activity, introduced by Libby in 1946 (Libby 1946) or by direct counting accomplished in accelerator mass spectrometry (Bennett et al. 1977). Since the production rate of radiocarbon is not exactly constant, a calibration of the time scale has to be performed by dendrochronology, where the ^{14}C content of single tree rings is determined and the calendar age is deduced from tree ring counting.

2 The Methodology of Accelerator Mass Spectrometry

The big advantage of AMS compared to decay counting is that the same precision can be obtained in a much shorter time with much smaller samples containing carbon of less than a milligram. This can be illustrated by the fact that a typical modern carbon sample of 1 g contains 6×10^{10} atoms of ^{14}C resulting in only 14 decays per minute. Therefore with decay counting a measurement time of about 50 h is necessary to achieve a statistical accuracy of 0.5%, whereas the same precision can be reached in 10 min with a modern AMS facility. However, the measurement of an isotope ratio $^{14}C/^{12}C$ of about 10^{-12} is an experimental challenge since it corresponds to the detection of one slightly heavier grain of sand in a living room filled with sand. Conventional mass spectrometry is not suitable for the measurement of such small isotope ratios since the separation of ^{14}C ions from intense fluxes of ions of the ^{14}N isobar and mass 14 molecular ions such as ^{13}CH and $^{12}CH_2$ is not possible simultaneously with a high transmission of the ion

beam. The use of a tandem accelerator with a high voltage of several MV combines an excellent reduction of interfering background by many orders of magnitude with high beam transmission and hence a high count rate.

2.1 Typical AMS Set-up

A typical AMS facility consists of a negative-ion source, an injection spectrometer, a tandem accelerator, an analyzing spectrometer, and a heavy-ion detector. These different parts serve as efficient filters for the separation of ^{14}C from the abundant background. As an example the Erlangen AMS facility (Kretschmer et al. 1997a) is shown schematically in ❯ *Fig. 1*. Negative ions are generated from the sample in a high-current sputter ion source, pre-accelerated to 50 keV, and injected into the tanden accelerator by a combination of a 90° electrostatic deflector and a 90° magnet with fast isotope switching for the sequential injection of ^{14}C, ^{13}C, and ^{12}C ions. The use of negative ions eliminates isobaric interference since ^{14}N does not form negative ions. The mass-analyzed ions are first accelerated to the positive-high-voltage terminal (5 MV) of the EN tandem accelerator where several electrons are stripped off during a passage through a thin carbon foil. Molecular background is eliminated in this stripping process since molecular ions are dissociated and the corresponding positive fragments have a different energy compared to carbon ions emerging from negative atomic ions. In the second stage of the tandem accelerator the now positively charged ions are accelerated further back to ground potential. They

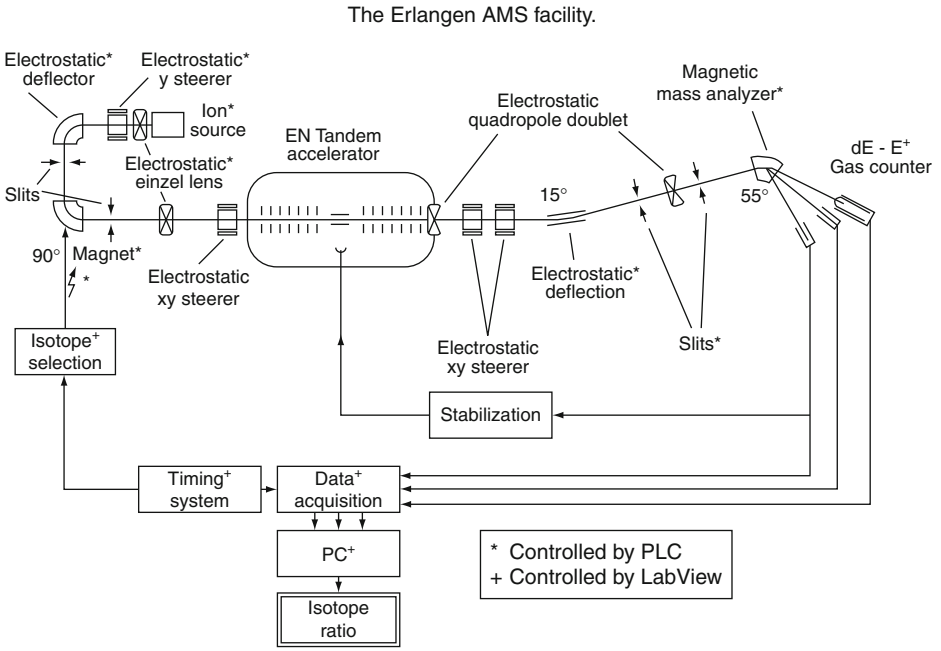


Fig. 1
Schematic view of the Erlangen AMS facility

are analyzed due to charge state, energy, and mass in a spectrometer consisting of a 15° electrostatic deflector and a 55° analyzing magnet. Finally the stable ions ^{13}C and ^{12}C are measured in Faraday cups and ^{14}C is detected in a ΔE–E gas detector which serves as a last filter for the elimination of remaining background. *Figure 2* shows ΔE–E spectra for a modern sample and for graphite demonstrating the success of background elimination: the spectrum for modern carbon shows a huge peak of ^{14}C, whereas the graphite spectrum only shows a few events. The beam transport between injection and analyzing magnet is completely performed by electrostatic devices, which ensures identical ion optics for the different carbon isotopes. The particle transmission through the facility amounts up to 80% taking into account the charge distribution in

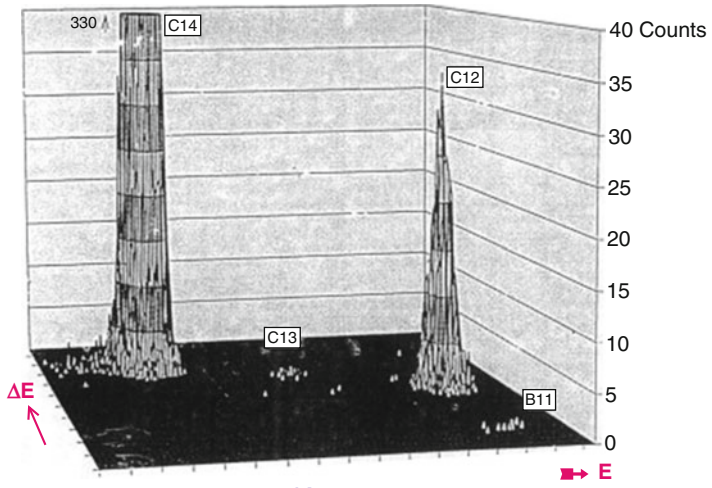


Fig. 2

ΔE–E spectra for a modern sucrose sample (*top*) and a graphite sample (*bottom*)

the high-voltage terminal. Thus even with conservative ion-source conditions of 25 μA ^{12}C current, a ^{14}C count rate of 100 Hz can be obtained for an ANU sucrose calibration sample. Machine background determined with graphite samples is measured at 0.02% modern carbon (pMC) corresponding to an apparent age of 70,000 years. The measurements are made in turns of one-minute runs which allow statistical analysis of the data and an on-line control of the system via the particle transmission and the isotope ratios. In the routine sequence of AMS measurements the quality of the whole facility is first checked with calibrated samples of known ^{14}C content. Then two to three unknown samples are measured before another calibrated sample is used. In this way an accuracy of about 0.5% has been established.

2.2 Determination of Calendar Age

The isotope ratio $^{14}C/^{12}C$ is obtained from the integrated ^{14}C events shown in *Fig. 2* and the ^{12}C current measured in the Faraday cup behind the analyzing magnet. This ratio has to be corrected for an isotopic fractionation depending on the specific assimilation process for the intake of CO_2 by the plants. This fractionation, caused by mass-dependent diffusion processes, can be deduced from the ratio of the stable carbon isotopes $^{13}C/^{12}C$ measured independently. The correction for $^{14}C/^{12}C$ is obtained under the assumption that its fractionation is twice as much as for $^{13}C/^{12}C$. After a further correction for the machine background obtained from graphite measurements, the $^{14}C/^{12}C$ ratio from the unknown sample is compared to that of a calibration sample from 1950. Assuming a constant formation rate and the exponential decay with known half-life, the so-called "radiocarbon age" in years before present (BP), i.e., before 1950, is derived. Since the assumption of constant formation rate is only approximate, the radiocarbon age of single tree rings is compared to the calendar age deduced from tree ring counting. The result of numerous dendrochronological studies is that the true age deviates up to 10% from the radiocarbon age, where the gross structure of this deviation can be related to variations of the earth's magnetic field and fine structure may be due to solar activity. Thus for a determination of the calendar age the use of calibration curves is essential.

2.3 Sample Preparation

In this section the conversion from the raw sample (soil, bones, etc.) to a sputter target suitable for an efficient formation of negative carbon ions is described. In a first step the sample has to be chemically pretreated to remove carbon compounds, which are not representative for the age of the sample. Then the remaining material is oxidized to CO_2 which is finally catalytically reduced to carbon.

2.3.1 Pretreatment of Sediment Samples

Two major components may obscure the ^{14}C results: carbonates arising from the erosion of limestone could be considerably older than the investigated layer and humic acids could be younger due to their high mobility. A fast and efficient mechanical separation of the sample material in organic and inorganic components is possible due to their different density. The

sediment material is dissolved in a zinc-chloride solution of density $\rho = 2\,g/cm^3$ and after ultracentrifugation with 3,000 cycles per minute the material is divided into the light organic fraction with $\rho = 1.3–1.5\,g/cm^3$ and the heavy inorganic fraction with $\rho = 2.65\,g/cm^3$. Subsequently the organic fraction is treated by the usual acid–alkali–acid (AAA) method, where HCl is used to remove still remaining carbonates and NaOH to remove humic acids. Finally the pretreatment is finished by heating in HCl, washing in deionized water, and drying the remaining material in an oven. The organic residue is called bulk sediment, since up to now no separation due to the size of the objects has been performed and since it represents a mixture of the remains of water plants, algae, pollen, and different macrofossils. Since it is the aim of our project to deduce the vegetation history from the pollen distribution in the sediment core, a direct dating of the pollen is highly desirable. The extraction of pollen material is mainly accomplished by additional sieving with 100 μm and 20 μm nylon meshes, since the average size of pollen ranges from 20–100 μm. The cellulose is removed with H_2SO_4 and finally for the deflocculation of amorphous organic material a treatment with NaOCl is performed. Details of the sample pretreatment are discussed in (Kretschmer et al. 1997b) and (Kretschmer et al. 1998).

2.3.2 Pretreatment of Bones

For the dating of bones we modified the procedure proposed by Longin (Longin 1971), which is based on the removal of the inorganic fraction and the extraction of collagen. In an ultrasonic treatment the bone is cleaned with deionized water. Then the carbonates can be removed in two alternative ways: either the dried and ground bone or the complete bone is treated with 0.5–1 N HCl. The grinding accelerates the procedure, but if the bone contains only a small amount of collagen it may be lost. In a next step humic acids are dissolved with 0.25–1 N NaOH and finally the collagen is extracted by dissolving the residues in acidic water (pH = 1–3) at 58 °C for 16 h. Remaining insoluble residues are separated by centrifugation and can be dated for a comparison. By drying the solution at 80 °C all collagen can be obtained in form of gelatine. There is also the possibility to get rid of some remaining contamination by an additional ultra-spin macrofiltration. In this way the gelatine fraction with molecule masses > 30 kD can be extracted and then dried as described above. After each step the residues are washed with deionized water.

2.3.3 Pretreatment of Archaeological Samples

Archaeological samples are often treated with organic conservation material. To avoid possible contamination with carbon of different age, a Soxhlet method was used as described by Bruhn et al. (2001). The Soxhlet-type apparatus, schematically shown in ❯ *Fig. 3*, consists of an upper and a lower glass container at different temperatures connected through a glass pipe for the rising solvent vapor and a glass siphon for the liquid solvent. The upper container is equipped with a reflux cooler to condense the solvent vapor and with a 45-μm-pore-size borosilicate-glass-fiber filter holding the sample. The lower container is placed in a heating block. During operation the solvent is heated to boiling, the solvent vapor rises to the upper container where

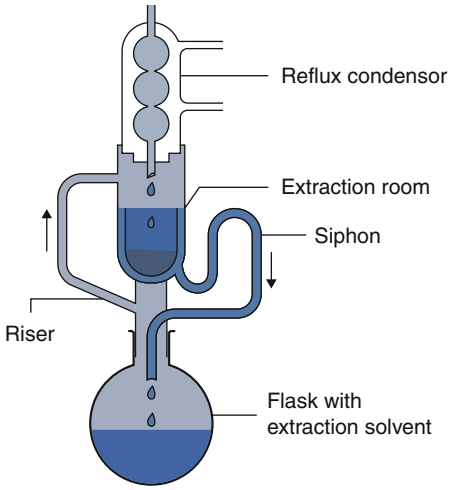


Fig. 3
Soxhlet extractor for the removal of conservation chemicals

it condenses at the cold finger and drops down to the sample which simmers in the hot solvent. When the level of the hot solvent reaches the upper bend of the siphon, the solvent and the dissolved organic contaminants flow down to the heated lower container, then the pure solvent evaporates again leaving behind the contaminants. The whole procedure is repeated frequently until the contaminants are leached out completely. Since the chemical composition of the conservation material is not known, the process is performed with different organic solvents in a special sequence, where each subsequent solvent removes its predecessor until the final one can be removed by water. Especially in the case of bone dating, a moderate boiling point (bp) of the solvents is important to avoid loss of collagen. In our investigation, we used a sequence of tetrahydrofuran (bp 65 °C, 8 h), trichloromethane (bp 61 °C, 8 h), acetone (bp 56 °C, 4 h), and methanol (bp 65 °C, 2 h). Further treatment before combustion depends on the sample.

2.3.4 Combustion to CO_2 and Reduction to Carbon

For a higher sample throughput both the combustion to CO_2 and the reduction to carbon have been simplified, similar to the Groningen target preparation (Aerts-Bijma et al. 1997). Therefore we developed a new sample preparation system consisting of a newly installed element analyzer (EA) in combination with a stable mass spectrometer (MS) and a suitable multi-sample reduction facility described in detail by Morgenroth et al. (2000). The element analyzer produces purified CO_2 in a fast flash combustion, a small part (10%) of the CO_2 is transferred to the stable isotope magnetic mass spectrometer for a high-precision measurement of $\delta^{13}C$ and $\delta^{15}N$ and 90% to the liquid-nitrogen cryo trap, where CO_2 is collected for further treatment.

3 Applications of Radiocarbon Measurements to Interdisciplinary Research

3.1 Sediment Dating

Due to a worldwide discussion about possible global climate change there is an increasing interest to predict the future development by climatic models. The key for verifying these models is situated in the different climatic archives of the earth, which supplies us with information about the climatic development of the past. Terrestrial sediments and peat profiles are one of these archives representing the main material to investigate the climate change in Central Europe. The current research program in Erlangen is concentrated on ^{14}C measurements with special focus to the investigation of sediment profiles (Kretschmer et al. 1997b). Since the last glacial period the temperature in Germany has increased by about 10 °C and vegetation has developed from only few species to a huge variety. In that time period of increased warming both animals and men have influenced the vegetation considerably. To establish a better chronology of climate and vegetation since the last glacial period in Germany the research program "change of geo-biosphere during the last 15,000 years", has been started in Germany. Sediment cores of several meters length from different locations have been taken mainly from bog sites. Radiocarbon dating of these profiles together with corresponding pollen analyses allows the deduction of vegetation history in Holocene. To obtain more reliable results, the dating of pollen grains and macrofossils is performed in addition to that of the bulk sediments. Here we discuss the results of two sediment cores from bog sites close to Klein Oelsa and Altliebel, in the Upper Lausitz northeast of Dresden (Kretschmer 1999), where also prehistoric settlements have been found. The results of other sediment cores and the corresponding pollen distributions from Southern Bavaria are discussed in Kretschmer et al. (1997b) and Kretschmer et al. (1998).

The results of ^{14}C AMS dating for both cores are shown in ❯ *Figs. 4* and ❯ *5*: The calibrated age is displayed as a function of core depth. For the Klein Oelsa core the complete age profiles for three different fractions of the sediment, for pollen, humic acids, and bulk have been determined. For a depth below 2.5 m the amount of pollen is decreasing, resulting in a very small amount of extracted carbon and therefore increasing error bars for the pollen fraction. As demonstrated in ❯ *Fig. 4* the agreement of the different age profiles with each other is perfect within the errors. The results for the Altliebel core shown in ❯ *Fig. 5* indicate a pronounced change in sedimentation rate at a depth of 1.50 m corresponding to an age of 8,200 BP. Both cores cover the time range of the last 14,000 years and together with the simultaneously performed pollen analysis (about 40 different pollen species) a reliable absolute chronology of vegetation history for the Upper Lausitz can be deduced.

3.2 Environmental Studies

Many organic environmental compounds are potentially dangerous due to their allergic, toxic, or carcinogen impact on humans. For an effective program to reduce their concentration in houses, their sources have to be detected. Our investigation is focussed on aldehyde compounds since their indoor concentration is relatively high and since they originate from biogenic or anthropogenic sources, which can be disentangled by their different ^{14}C content measured via AMS. The main assumption behind the method is that materials, derived from assimilation of atmospheric carbon dioxide, i.e., plant materials of all kinds, contain the equilibrium

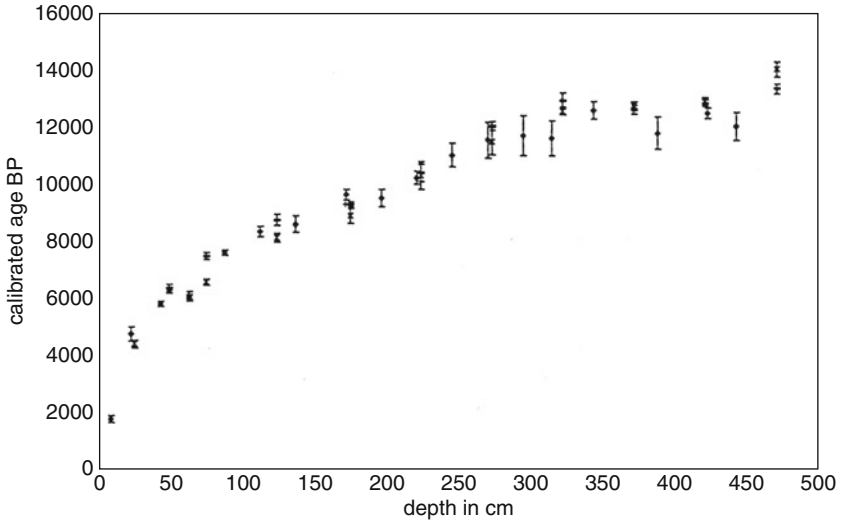


Fig. 4
Age profile for the sediment core of Klein Oelsa (NE of Dresden) using different fractions of the sediment (2σ errors): bulk sediment (-), pollen (♦), humin (x)

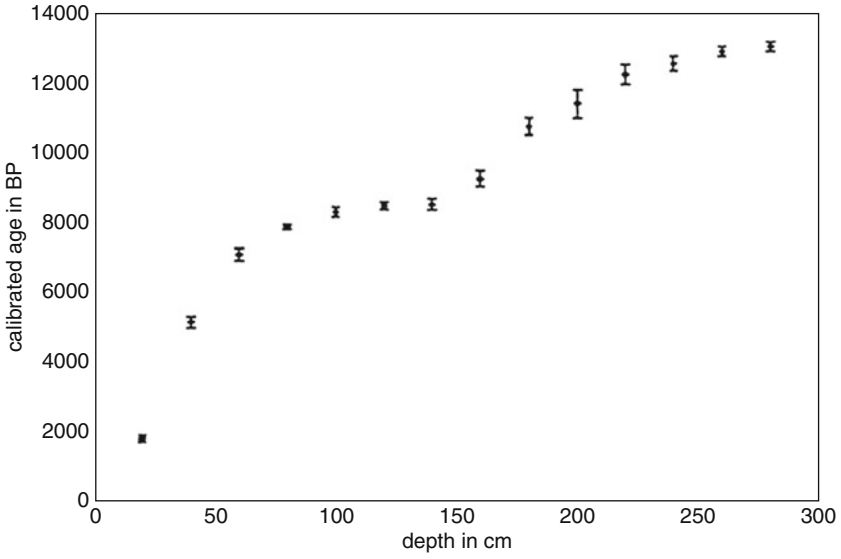


Fig. 5
Age profile for the sediment core of Altliebel (NE of Dresden) using bulk sediment (2σ errors)

concentration of ^{14}C typical for modern carbon. On the other hand, fossil materials do not contain ^{14}C due to the relatively short half-life of ^{14}C, as compared to diagenetic processes of several hundred million years. One problem with this method is the relatively low concentration

of environmental compounds in natural samples (e.g., ≈ 50 μg/m^3 for acet- and formaldehyde in indoor air), which requires either long sampling times or small sample sizes. If the resulting carbon has a mass of less than 50 μg, then the problem of contamination during each step of sample preparation is important. To minimize possible contamination we have installed a hybrid ion source which can produce negative ions from both solid samples and CO_2 gas. If gas is used then the reduction of CO_2 is no more necessary and therefore there is less contamination. We developed a compact sampling device for indoor air, using the conventional dinitrophenolhydrazine (DNPH) derivatization method. Isolation and purification of the corresponding derivative is done by high-performance liquid chromatography (HPLC). The isolated derivative is then injected into an elemental analyzer and fed via a gas handling system (Uhl et al. 2007) directly into the gas sputter source for the ^{14}C AMS measurement.

First samples were taken from a small beer tavern where smoking is allowed, and from a living room. After a sampling time of 3–6 h the carbon content of the isolated derivatives from acetaldehyde and formaldehyde is about 10–20 μg which is enough for an AMS measurement with the gas ion source.

For the living room the ^{14}C content of both formaldehyde (concentration: 59 μg/m^3) and acetaldehyde (concentration: 44 μg/m^3) is about 50 pMC within the error bars, corresponding to a 1:1 mixture of anthropogenic and biogenic sources. For the beer tavern both concentrations were higher (formaldehyde: 126,3 μg/m^3, acetaldehyde: 79,7 μg/m^3), the sources are predominantly biogen with the formaldehyde (^{14}C content: 120 ± 20 pMC) originating from cigarette smoke and the acetaldehyde (^{14}C content: 113 ± 18 pMC) originating from cigarette smoke and alcohol consumption. This was our first test of compound-specific radiocarbon analysis using a conventional derivatization method as proposed by Kato et al. (2008). The disadvantage with this method is that the corresponding derivative now contains six additional carbon atoms from DNPH which was produced petrochemically. The ^{14}C content of the aldehyde (one or two carbon atoms, respectively) has to be calculated and hence the error increases considerably. In a next step we plan to find a derivatization compound, which contains only one or even no carbon atom.

3.3 Archaeological Samples

In October 2000 a mummy was confiscated at Kharan in Balochistan, and handed over to the National Museum of Pakistan (Ibrahim 2001). It was enclosed in a double coffin, the inner one made of stone and the outer one of wood. The body of the mummy was wrapped in linen bandages and attached to a mat of straw with resins, wax, and honey. The mummy, shown in *Fig. 6*, is wearing a gold crown, and a golden Cypress tree is embossed on the chest of the princess, below the gold plate with her and her father's name written in Old Persian script. Both the inscription on the wooden coffin and the gold chest plate strongly supported the suggestion that the mummy was Ruduuna, the daughter of Xerxes, the great Persian king, who lived from 518 to 465 BC. Since no other Persian mummies have been found so far, it was supposed to be the archaeological find of the century. This prompted a great deal of scientific research involving an inspection of the Old Persian writing and the craftsmanship of the coffins, X-ray computer tomography (CT) of the whole body, and radiocarbon dating.

The ^{14}C dating was performed with the Erlangen AMS facility. First we received a small piece of textile wrapped on the body and a piece of the mat containing some grains of charcoal, and later some samples from the mummy itself, namely, bone, skin, and muscular tissue. To avoid



Fig. 6
The Persian mummy

possible contamination from conservation material, a Soxhlet method was used as described in Sect. 2.3.3.

The results of the AMS measurements (Kretschmer et al. 2004) are shown in *Table 1*, where the ^{14}C content is listed in units of percent modern carbon (pMC). The results for the first three samples are consistent with a radiocarbon content of about 113 pMC, clearly exceeding the expected value of 75 pMC. This surprising result implies that all three samples are younger than AD 1955 due to their increased ^{14}C content arising from the bomb peak caused by nuclear weapons tests. To obtain an exact date of the time of death, the remaining ambiguity (1958 or 1992) arising from the rise and the decline of the bomb peak curve could be removed by measuring the ^{14}C content of different parts of the body, which have different turnover times of carbon. Consequently we received further samples of bone, muscular tissue, and skin from the mummy. The bone should contain older carbon because of the longer turnover time of collagen in human bones (Wild et al. 2000). If the bone contains less ^{14}C than skin or muscular tissue, then the time of death is in the rise of the bomb peak, while if it is vice versa, it is in the decrease of the bomb peak. Collagen was used for the bone dating. This was obtained as described above, first with (a) and then without (b) Soxhlet extraction. The agreement of both results shows that the Soxhlet extraction is not relevant in this case since the bone is not contaminated with conservation material. This is not the case for the skin, where sample (a) with Soxhlet extraction

Table 1
AMS and $\delta^{13}C$ measurements of "mummy" samples. Index (a) and (b) refer to pretreatment with and without Soxhlet method (see text)

Sample	^{14}C content (pMC)	$\delta^{13}C$ (‰)
Textile	113.85 ± 0.47	−27.9
Straw	113.58 ± 0.47	−25.7
Charcoal	112.45 ± 0.45	−23.4
Bone (a)	115.02 ± 0.44	−19.3
Bone (b)	115.04 ± 0.51	−19.1
Muscle tissue	112.78 ± 0.49	−24.6
Skin (a)	111.08 ± 0.49	−24.4
Skin (b)	114.14 ± 0.48	−24.0
Conservation wax (a)	114.76 ± 0.58	−26.9
Conservation wax (b)	115.22 ± 0.65	−26.7

contains less ^{14}C than sample (b) without. In addition we measured the ^{14}C content of the yellow conservation wax from the skin obtained either from the Soxhlet tetrahydrofuran fraction (a) or from a treatment with hot HCI only (b). Again, both samples are modern and agree within their errors.

The results for the ^{14}C content of skin, muscular tissue, and bone are displayed in *Fig. 7* compared to the calibration curve of the last 70 years, where the values for the bomb peak period are annual averages of the radiocarbon content of tropospheric CO_2 measured at Vermunt and Schauinsland by Levin et al. (1995), Levin and Kromer (1997). Since bone contains more ^{14}C than skin and muscular tissue, it is evident from the discussion above that the time of death is in the declining part of the bomb peak. The calibration was performed with a program similar to OxCal, which was modified to adapt the step size on the calibrated age axis to the steepness of the calibration curve. Since the calibrated age of the skin is AD 1994–1996 (1σ), it can be concluded that the woman has died in this period or shortly afterward. It should be mentioned that the time lag between the apparent age of the bone and the skin is only about 5 years, which is considerably smaller than the time lag of about 20 years which has been found by the Vienna group for a 30-year-old man (Wild et al. 2000). One explanation could be that the woman was only 20 years old.

The absolute AMS dating of the mummy and related material demonstrate that it is a modern fake. This is further supported by a detailed examination of the Persian writing and by the investigation of the CT scans of the mummy. There were several mistakes in the text and grammar of the Old Persian inscription and it contained words which did not exist in the time of Xerxes. A study of craftsmanship at the wooden coffin showed that the chiseling of the motifs was too rude and seemed to be done with modern equipment – a microscopic inspection showed tracing marks of graphite pencil. CT scans of the body showed that the age of the woman was between 20 and 40 years, that all teeth have been removed, and that the embalming procedure showed features which have been never observed in Egyptian mummies. After these detailed studies it is evident that this object is a very elaborate forgery, produced with much labour and skill and successful in creating a stir in the international media as well as amongst scientists and archaeologists.

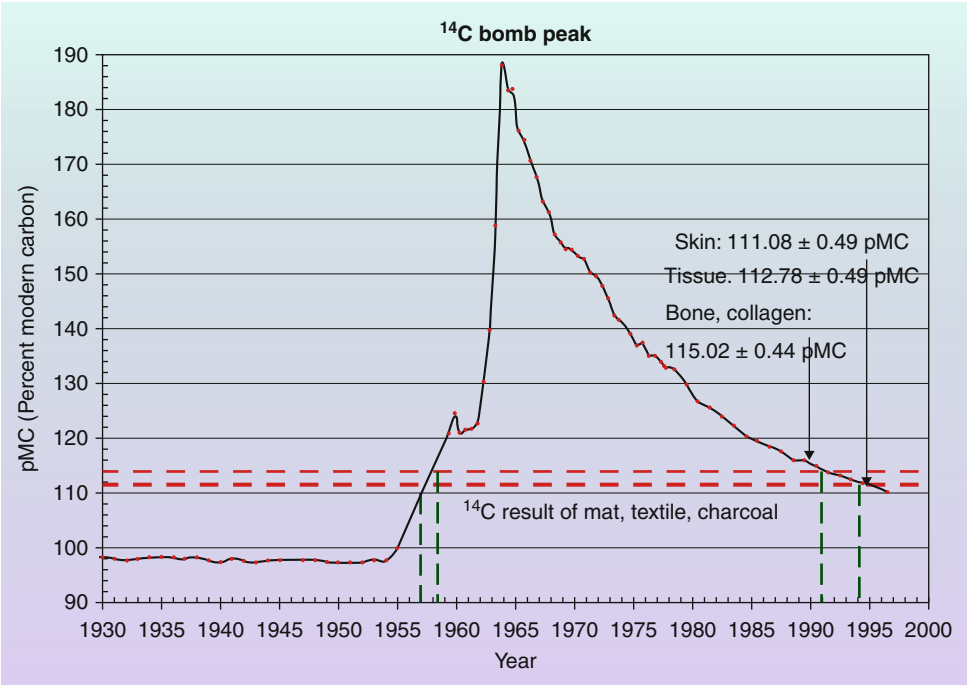


Fig. 7
^{14}C results of different samples from the Persian mummy compared to the calibration curve since 1930 AD (Levin et al. 1995; Levin and Kromer 1997), the *dashed lines* correspond to the results for textile, straw, and charcoal samples

4 Concluding Remarks

Since the first demonstration in 1977 that ^{14}C could be detected at natural level using a tandem accelerator as part of a mass spectrometer, the field of AMS has expanded into many areas of science. As shown in this contribution radiocarbon measurements with AMS can be used in archaeology, geology, and environmental research even for samples containing carbon in the 50 μg range which makes AMS superior to decay counting. Whereas the time range of ^{14}C measurements is limited to about 50 ka due to its relatively short half-life, other long-lived isotopes such as ^{10}Be, ^{26}Al, ^{36}Cl, and ^{129}I can be used by AMS with emerging applications (Fifield 1999).

References

Aerts-Bijma A-TH, Meijer HAJ, van der Pflicht J (1997) AMS sample handling in Groningen. Nucl Instr Methods B123:221–225

Bennett CL, Beukens RP, Clover MR, Gove HE, Liebert RB, Litherland AE, Purser KH, Sondheim WE (1977) Radiocarbon dating using electrostatic accelerators: negative ions provide the key. Science 198:508–510

Bruhn F, Duhr A, Grootes PM, Mintrop A, Nadeau MJ (2001) Chemical removal of conservation substances by 'soxhlet'-type extraction. Radiocarbon 43(2):229–237

Fifield LK (1999) Accelerator mass spectrometry and its applications. Rep Prog Phys 62:1223

Ibrahim A (2001) In: Ibrahim A, Lashari K (eds) The archaeological review. Karachi, p 17

Kato Y, Shinohara N, Yoshinaga J, Uchida M, Matsuda A, Yoneda M, Shibata Y (2008) Determination of $^{14}C/^{12}C$ of acetaldehyde in indoor air by compound specific radiocarbon analysis. Atmos Environ 42(5):1049–1056

Kretschmer W (1999) Accelerator mass spectrometry and its applications in archaeology, geology and environmental research. Acta Physica Polonia B31:123–133

Kretschmer W, Anton G, Benz M, Blasche S, Erler G, Finckh E, Fischer L, Kerscher H, Kotva A, Klein M, Leigart M, Morgenroth G, Küster H (1998) The Erlangen AMS facility and its applications in ^{14}C sediment and bone dating. Radiocarbon 40(1):231–238

Kretschmer W, Anton G, Bergmann M, Finckh E, Kowalzik B, Klein M, Leigart M, Merz S, Morgenroth G, Piringer I (1997a) The Erlangen AMS facility: status report and research program. Nucl Instr Meth B 123:93–96

Kretschmer W, Anton G, Bergmann M, Finckh E, Kowalzik B, Klein M, Leigart M, Merz S, Morgenroth G, Piringer I, Küster H, Low RD, Nakamura T (1997b) ^{14}C Dating of sediment samples. Nucl Instr Meth B 123:455–459

Kretschmer W, von Grundherr K, Kritzler K, Morgenroth G, Scharf A, Uhl T (2004) The mystery of the Persian mummy. Nucl Instr Meth B 223:672–675

Levin L, Kromer B (1997) Twenty years of atmospheric $^{14}CO_2$ observations at Schauinsland station, Germany. Radiocarbon 39(2):205–218

Levin L, Kromer B, Schoch-Fischer H, Bruns M, Münnich M, Berdau D, Vogel JC, Münnich KO (1995) 25 years of tropospheric ^{14}C observations in central Europe. Radiocarbon 27(1):1–19

Libby LW (1946) Atmospheric helium three and radiocarbon from cosmic radiation. Phys Rev 69(11–12):671–672

Longin R (1971) New method of collagen extraction for radiocarbon dating. Nature 230: 241–242

Morgenroth G, Kerscher H, Kretschmer W, Klein M, Reichel M, Tully T, Wrzosok I (2000) Improved sample preparation techniques at the Erlangen AMS-facility. Nucl Instr Meth B 172: 416–432

Uhl T, Luppold W, Rottenbach A, Scharf A, Kritzler K, Kretschmer W (2007) Development of an automatic gas handling system for microscale AMS ^{14}C measurements. Nucl Instr Meth B259:303–307

Wild EM, Arlamovsky KA, Goiser R, Kutschera W, Priller A, Puchegger S, Rom W, Steier P, Vycudilik W (2000) ^{14}C dating with the bomb peak: An application to forensic medicine. Nucl Instr Meth B 172:944–950

27 Geoscientific Applications of Particle Detection and Imaging Techniques with Special Focus on the Monitoring Clay Mineral Reactions

Laurence N. Warr · Georg H. Grathoff
Institut für Geographie und Geologie,
Ernst-Moritz-Arndt-Universität, Greifswald, Germany

1 Introduction 669

2 Characterization of Clay Minerals Reactions 670

3 Application of Electron- and Focused-Ion-Beam Microscopy 672

4 Applications to the Disposal of Nuclear Waste: Reactions in Bentonites 673
4.1 X-ray Diffraction Study of Bentonite Hydration under Conditions of Varying Humidity 673
4.2 Environmental Scanning Electron Microscopy 674
4.3 Wet-Cell X-ray Diffractometry 674

5 Applications to the Storage of CO_2: Reactions in Shales 677
5.1 Three-Dimensional Reconstruction Using Combined Ion- and Electron-Beam Microscopy 678

6 Conclusions and Outlook 680

7 Cross-References 681

8 Analytical Equipment Used in this Study 681

C. Grupen, I. Buvat (eds.), *Handbook of Particle Detection and Imaging*,
DOI 10.1007/978-3-642-13271-1_27, © Springer-Verlag Berlin Heidelberg 2012

Acknowledgments 681

References 681

Abstract: The combined use of focused X-ray, electron, and ion beams offers a diverse range of analytical capabilities for characterizing nanoscale mineral reactions that occur in hydrous environments. Improved image and microanalytical techniques (e.g., electron diffraction and energy-dispersive X-ray spectroscopy), in combination with controlled sample environments, are currently leading to new advances in the understanding of fluid–mineral reactions in the Earth Sciences. One group of minerals playing a key role in the containment of radioactive waste and the underground storage of CO_2 is the clay minerals: these small, expandable, and highly adsorbent hydrous phyllosilicates form important low-permeable geological barriers by which waste can be safely deposited. In this article we summarize some of the state-of-the-art particle and imaging techniques employed to predict the behavior of both engineered and natural clay mineral seals in proposed storage sites. Particular attention is given to two types of low-permeability geomaterials: engineered bentonite backfill and natural shale in the subsurface. These materials have contrasting swelling properties and degrees of chemical stability that require detailed analytical study for developing suitable disposal or storage solutions.

1 Introduction

The interaction of subatomic particles with natural geomaterials forms the backbone of the analytical sciences by which the age, structure, and chemistry of the Earth has been resolved. Today the research geoscientist relies on using a broad spectrum of advanced particle detection tools in order to (1) date rocks and minerals (Wagner 1968; Segl et al. 1984; Dickin 2008), (2) to determine geophysical properties while drilling the Earth's crust (Timur and Toksoz 1985), and (3) to study the physical and chemical properties of geomaterials and geofluids both in the laboratory and the field (Farges et al. 1993; Langford 2006). Such tools require knowledge of the behavior of radionuclides, X-rays, electrons, and ions as they interact in both solids and fluids and allow full material characterization ranging from the imaging of crustal-scale features down to the scale of the crystal lattice.

One of the major challenges facing both mineralogists and geochemists is to develop laboratory techniques for determining the rates of structural and chemical changes in minerals that occur both in natural and experimental systems and thus determine reaction kinetics (Lasaga 1981; Nagy 1985). Ideally these characteristics are also determined on an individual particle basis so that the mechanisms can be fully understood and the results compared with changes in bulk material properties. Effective determination of a mineral's crystal chemistry and structure is therefore an essential part of understanding fine-grained mineral systems and interactions with the fluid phase.

In this chapter we provide some examples of how subatomic particles are used to study the crystal chemistry of some of the smallest and most reactive minerals: the clay minerals. In the current climate of global warming, with the possibility of underground CO_2 storage and the controversial geological storage of radioactive waste, these minerals are of central importance in providing adsorbent and impermeable layers by which waste can be confined. We address these two topics to highlight how mineral reactions can be characterized by using a combination of X-ray diffraction monitoring techniques and ion- and electron-beam microscopy.

2 Characterization of Clay Minerals Reactions

Clay minerals are the most common type of minerals that form in low-temperature environments of the Earth's crust. They play a key role in surface and subsurface processes as they represent the major reactive interface between the lithosphere, hydrosphere, atmosphere, and biosphere. They also represent an economic resource of growing importance, particularly in the developing fields of industrial application, environmental protection, and nanotechnology. As a result of their shape, small grain size, large surface areas, and charged-particle surfaces, this type of mineral material displays a number of rather important properties. They can sorb cations, water, as well as other polar ions and organic complexes onto and into their particle structure, and thus play a critical role in water–rock interaction at surface and shallow crustal levels, acting as catalysts for both chemical and biochemical reactions. This role is emphasized by the intimate link between crystal chemistry and particle size, their surface properties, and their physicochemical behavior (Nadeau et al. 1984). With the additional ability for crystal replication in surface environments, it has been suggested that clay minerals formed an important precursor leading to the evolution of DNA, and hence life (Cairns-Smith 1985). At deeper levels in the Earth's crust, the hydrous clay minerals release their absorbed and adsorbed water through prograde metamorphic dehydration reactions, providing an important source of fluids for driving mass transport within the upper crust and concentrating economic deposits such as hydrocarbons and ore mineral (Bethke et al. 1991; Grathoff et al. 2001).

Clay minerals are mostly fine-grained phyllosilicates that are typically <2 μm in size and include the naturally occurring nano-sized particles (<100 nm). Their importance in geotechnical applications are related to their large specific surfaces areas (10–136 m^2/g determined by N_2 adsorption), cation exchange capacities (ca. 3–210 milli-equivalent/100 g), and the unique ability to expand during hydration, inducing swelling pressures up to 100 MPa (Pusch 2004), which can lead to extremely low permeabilities ($<10^{-10}$ m/s). The strong adsorbent nature of the clay minerals is governed by the high degree of cation substitutions that may occur in both tetrahedral and octahedral layers, which leads to a diverse range of negative layer charges that are compensated by complexed inorganic and organic cationic substances including a large range of toxic elements. In addition to these properties, clay particles are notoriously thin (typically <10 nm thick) and show various degrees of crystalline ordering and polytypism (Warr and Rice 1994; Grathoff and Moore 1996; Grathoff et al. 2001). Another notable feature is the common occurrence of mixed-layered phases, for example, the occurrence of expandable smectitic layers interlayered within the same crystallite (i.e., optically parallel) with non-swelling layers (e.g., illite, kaolinite, vermiculite, chlorite). All these features result in a large diversity in physical and chemical properties that make them particularly suitable for the geological containment of waste under diverse chemical, pressure, and temperature conditions.

Due to the small size and phase complexity of clay mineral particles, an interdisciplinary analytical approach is essential for full characterization, which typically relies on the interaction of X-rays, electrons, and ions with the variably ordered mineral material (*Fig. 1a*). The platy shape of clay minerals and the structural diversity in the crystallographic c^*-direction make them ideal for X-ray diffraction (XRD) study when prepared as oriented films (texture preparations). Identification of X-ray reflections is aided by the one-dimensional modeling of XRD patterns using programs such as NEWMOD and CALCMIX (Reynolds 1985; Plancon and Drits 2000; Yuan and Bish 2010). Such programs allow both characterization and quantification of

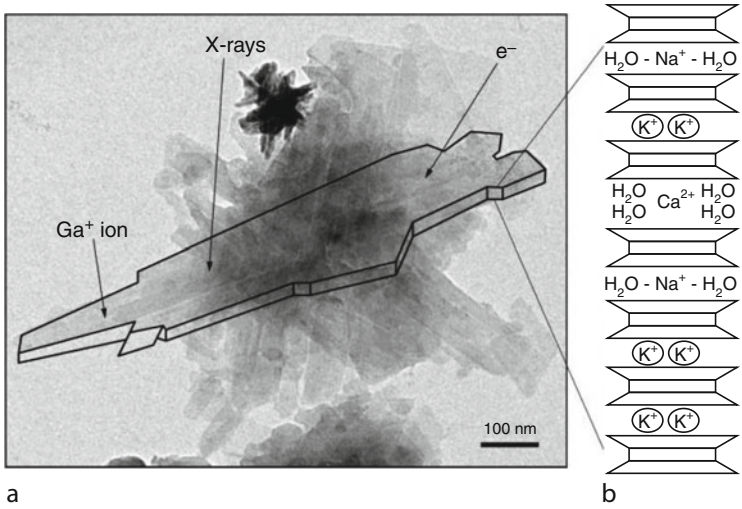


Fig. 1

(a) Illite-smectite crystals with one crystal highlighted and exaggerated in the third dimension. The crystal shows the different particles (X-rays, electrons, and Ga^+ ions) that we use for investigations. The X-rays are used for X-ray diffraction study of crystallography and the electron beam for a range of analytical and imaging applications shown in *Fig. 2*. The Ga^+ ions are used to mill the sample simultaneously with imaging. (b) A reconstruction of the illite-smectite mixed-layered crystal is shown based on all methods combined. The interlayer may be smectitic and occupied by either Ca with two water layers or Na with one water layer, or alternatively be of illitic composition with fixed K

mixed-layered clay mineral phases (*Fig. 1b*). When X-rayed as random powder preparations, the resultant patterns can be modeled in three dimensions using program methods such as WILDFIRE that calculates interstratified layer structures by the Fourier transform of individual layers and assembling the intergrowth structure in reciprocal space (Grathoff and Moore 1996). Fourier analyses of the shape of single-phase XRD reflections can also yield average crystallite thicknesses and crystallite thickness distributions useful for crystal growth studies (Eberl et al. 1998).

Whereas X-ray diffraction study is an excellent method for characterizing the crystallographic structure of fine-grained bulk materials, focused electron beams are more ideal for imaging and analyses of individual particles (*Fig. 2*). The detection of scattered electrons (secondary, backscattered, transmitted) and emitted X-rays, as they interact with the thin clay minerals particles, yields a range of crystal-chemical information that can be used to determine (1) surface topography, (2) internal microstructure, (3) crystallographic structure, and (4) microchemistry. In combination with the milling capabilities of focused ion beams, this interdisciplinary set of analytical tools is rapidly moving toward 3D imaging and microanalysis of materials.

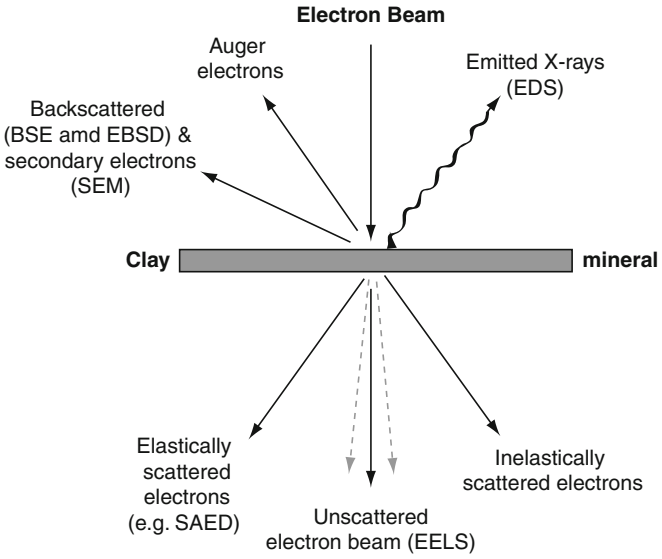


Fig. 2

Interaction of the electron beam with a thin clay mineral particle showing the signals produced and the measurement techniques. *EDS* Energy-Dispersive X-ray Spectroscopy, *BSE* Backscattered Electrons, *EBSD* Electron Backscatter Diffraction, *SAED* Selected Area Electron Diffraction, *EELS* Electron Energy Loss Spectroscopy. The electrons giving rise to an EELS signal actually deviate at a very small angle from the direct, unscattered electron beam (here shown schematically; modified after Buseck (1992))

3 Application of Electron- and Focused-Ion-Beam Microscopy

Electron microscopy studies, in particular scanning electron microscopy (SEM) and high-resolution transmission microscopy (HRTEM), have played a key role in developing our understanding of fine-grained mineral reactions that occur typically at scales <2 μm in size. Since the invention of the SEM in the 1930s, and the first pioneering studies on clays (Ardenne et al. 1940; Eitel and Radczewski 1940), this technique has become a standard instrument for the study of fine-grained rocks, sediments, and soils. The more recent development of the ultra-low vacuum "environmental" scanning electron microscope (ESEM) also allows for the study of bacterial–mineral interactions and bentonite (smectite-rich) clay swelling under varying conditions of air humidity (Montes 2005). Higher-resolution studies achieved by transmission electron techniques have equally played an important role in understanding small mineral phases, such as the nature of mixed-layered clays (Page and Wenk 1979; Lee and Peacor 1983). HRTEM has the advantage of being able to resolve lattice scale features and combined with electron diffraction and energy-dispersive spectroscopy (EDS) studies at sub-nanoscales have led to a more complete crystal-chemical description. One major limitation of HRTEM studies has been the difficulty in preparing samples, typically achieved either by ultramicrotoming or

dual ion milling. Both methods have their advantages and disadvantages. However, since the recent introduction of the focused ion beam (FIB), samples can now be thinned and coated with nanometer precision, opening up a new range of analytical possibilities. Today, the latest advanced tools combine the best of both FIB and SEM technology into multi-beam microscopes. Using both beams (ion and electron) simultaneously, these microscopes allow for ultra precise TEM sample preparation, 3D image, and microchemical and crystallographic reconstructions. The benefits of these new instruments in the field of fine particle mineral sciences are only just beginning (Obst et al. 2005; Mee et al. 2008).

In the following sections we present results from ongoing investigations on nuclear waste disposal and CO_2 sequestration in underground geological sites that illustrate these analytical methods of particle detection and imaging. In these studies some of the key properties of both engineered and natural barriers are addressed and their implications discussed.

4 Applications to the Disposal of Nuclear Waste: Reactions in Bentonites

Bentonites are employed as liners for water storage and municipal waste, and are currently proposed as low-permeability geotechnical barriers in high-level radioactive waste repositories (Pusch 1992). Their effectiveness as a barrier is dependent on their swelling capacity, which self seals during expansion in the presence of water, and the necessity to remain in this state over long periods of time despite changes in temperature, humidity, wetting–drying cycles, cation migration, and the infiltration of salt solutions (Pusch 2004; Kaufhold and Dohrmann 2010). Under ideal conditions, pressed bentonite blocks can attain extremely low hydraulic conductivities of 10^{-14} m/s (Pusch 1992). In situ monitoring of the mechanisms and rates of hydration of montmorillonite, the main-constituent clay mineral, can be achieved using reaction chambers that are mounted onto XRD diffractometers (Kühnel and van der Gaast 1993; Hanchar et al. 2000). First we address the initial hydration reactions that will be expected when ground water initially enters the repository site. Such conditions can be simulated in the laboratory using a temperature-controlled humidity chamber.

4.1 X-ray Diffraction Study of Bentonite Hydration under Conditions of Varying Humidity

Many experiments have been done to determine the hydration behavior of bentonite under conditions of varying humidity in unconfined volumes (Mooney et al. 1952; Collins et al. 1992; Kühnel and van der Gaast 1993; Chipera et al. 1997; Ferrage et al. 2005). Our experiments concentrate on how the swelling behavior of montmorillonite changes with varying temperature, humidity, and types of cation exchange. In ❯ *Fig. 3* an example of the swelling behavior of a Ca-montmorillonite is given. Although most waste repository sites currently favor the Na-montmorillonite bentonite, these types of clays readily alter to Ca or Mg varieties by cation exchange with altered cements or saline solutions (Hofmann et al. 1994; Herbert et al. 2004). During this experiment, clay slurry was placed on a flat sample holder and mounted within an

enclosed chamber in the center of the X-ray diffraction goniometer. The chamber allows control of temperature and humidity using an external controller. The initial XRD run of the clay slurry at room temperature shows a hydrated montmorillonite reflection indicating a dominate lattice thickness between 18 and 20 Å that corresponds to a mixture of three and four water layer (WL) thicknesses. Very soon after the humidity was decreased at 40 °C to <90%, the thickness of the montmorillonite collapsed to ca. 15 Å (two WLs). After initial drying and shrinkage the thickness did not decrease below the thickness of two WLs until a threshold of 20% humidity when the montmorillonite collapsed to one WL. A final collapse toward 0 water layers only occurred if the temperature was increased to 100 °C in our experiments but all water was not removable under these conditions. This dehydration process described is, however, reversible. As the humidity is increased at 40 °C to 15%, the montmorillonite re-expanded to one WL, and at 35% to two WLs, which was maintained up to 93% humidity. Any further hydration of the interlayers would require the influx of liquid water, which cannot be simulated in this type of reaction chamber.

The above results highlight some key features of the hydration of montmorillonite interlayers with changing humidity conditions. These are (1) the general reversibility of the reactions and (2) the stepwise nature of the transformations, although some gradual changes in the WL structures do occur between main hydration steps. Modifications in hydration behavior have been reported during cyclic changes in air humidity due to hysteresis related to the rigidity of the clay–water system and dynamic changes in the thickness of quasicrystals (Laird et al. 1995). More serious is the non-reversible loss of montmorillonite hydration after been subjected to hot steam >150 °C (Couture 1985), which can lead to complete barrier breakdown.

4.2 Environmental Scanning Electron Microscopy

The results of a simple hydration–dehydration cycle with Na-bentonite are given in *Fig. 4*. During the initial increase in humidity starting at 63% humidity and 430 Pa only minor changes in particle structure are observed until saturated conditions (100%) were attained (*Fig. 4a*). However, dramatic changes occurred while maintaining the humidity at 100% for ca. 2 min and increasing the chamber pressure from 692 to 730 Pa. Here, massive expansion of the clay was seen together with the loss of particle structure by gel formation (*Fig. 4b*). At the end of the cycle at 38% humidity and 260 Pa, the overall texture at the beginning of hydration returns; however, the particles do have noticeably smoother edges than at the start of the experiment, indicating some possibly nonreversible features (*Fig. 4c*). This type of experiment helps to reveal the dynamics of hydration–dehydration cycles and represents an excellent accompanying technique for combining with the X-ray diffraction humidity experiments (*Fig. 3*). Such experiments are not only important for assessing the hydration behavior of bentonites as backfill material in waste repositories, but are critical for furthering our understanding of the basic mechanisms behind water adsorption/desorption on layer-charged mineral surfaces.

4.3 Wet-Cell X-ray Diffractometry

The hydration behavior of montmorillonite in aqueous solution can be studied in the laboratory using a small flow-through reaction chamber (wet-cell device) mounted onto the X-ray diffractometer (Warr and Hofmann 2003). This type of reactor provides a good analogue for

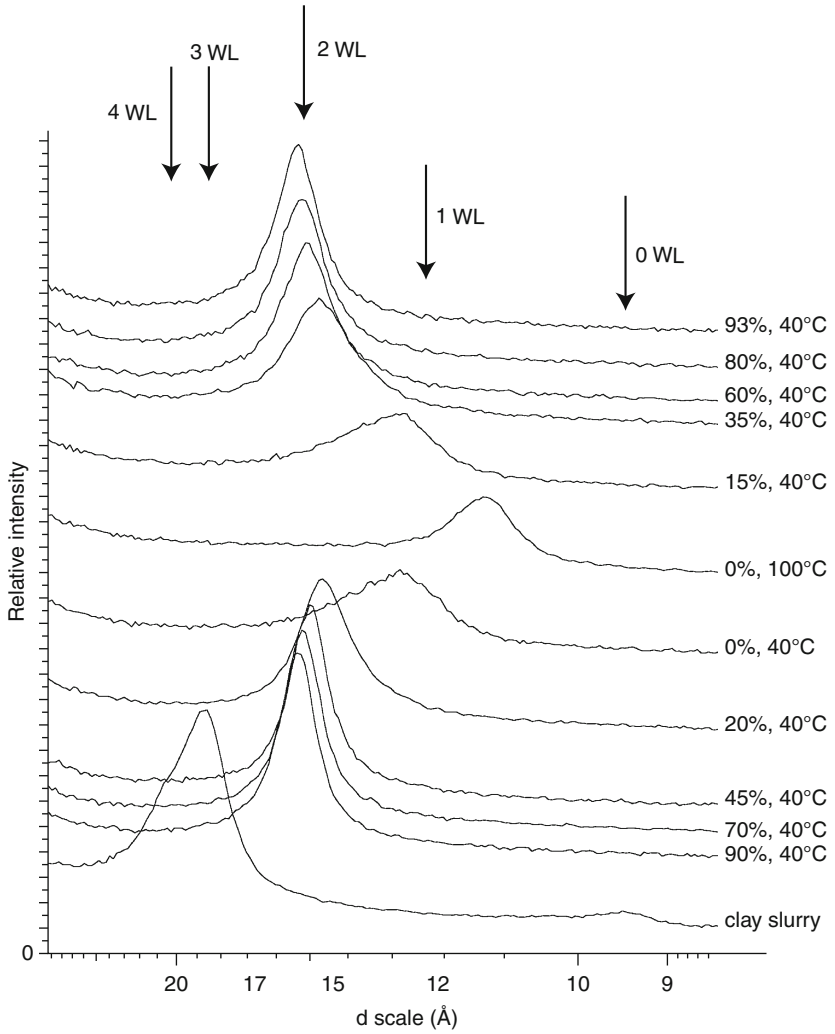


Fig. 3

X-ray diffraction (XRD) humidity experiment on bentonite clay showing the change in the Ca-montmorillonite 001 reflections as a function of decreasing and increasing air humidity. The sample was first measured as a clay slurry (bottom XRD pattern) and then dehydrated at 40 °C until 0% humidity. After complete drying at 100 °C the humidity was increased in steps up to 93%. These results highlight the reversibility of montmorillonite hydration in pure water systems at temperatures <100 °C. *WL*: water layer(s)

quantifying the mechanisms and rates of hydration in confined-volume systems equivalent to underground storage sites where a ground water infiltrates the bentonite layer (Warr and Berger 2007; Perdrial et al. 2009). In contrast to the stepwise hydration behavior of montmorillonite clays observed in air humidity experiments, the build-up of structured interlayer water in percolating water is more continuous in nature. In general, WLs do not appear spontaneously

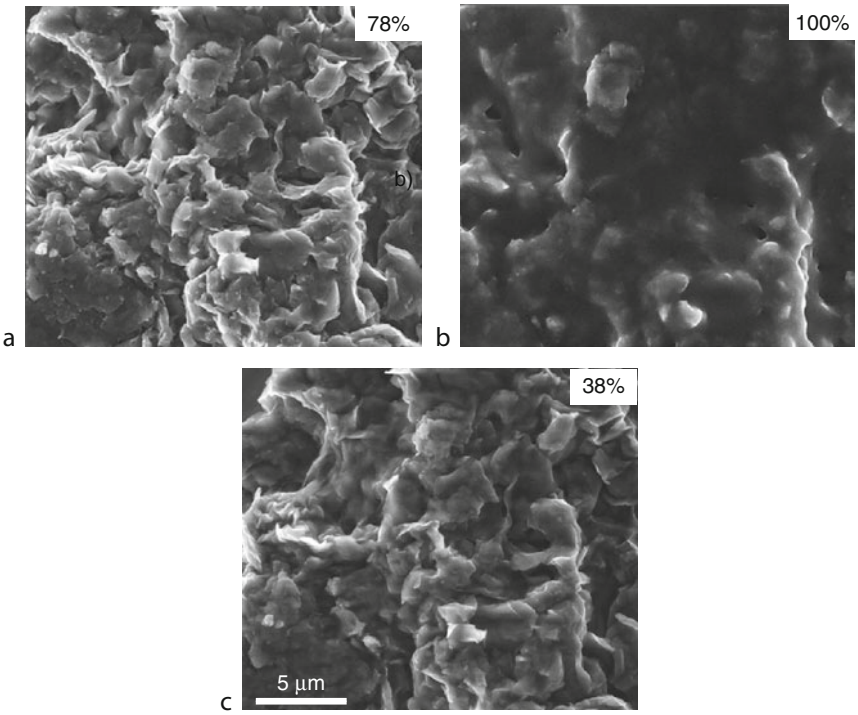


Fig. 4

Secondary-electron images of the hydration and drying of a bentonite sample in the environmental SEM mode cooled to 2 °C using a Peltier stage (7 kV). (a) The state of the bentonite clay during increasing humidity (78%) and a chamber pressure of 510 Pa, (b) fully hydrated clay with 100% humidity and 720 Pa pressure, (c) dehydrated clay at 38% humidity and 280 Pa pressure

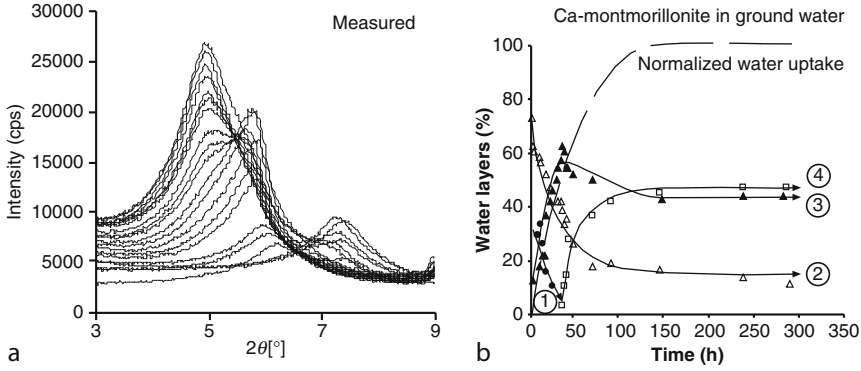


Fig. 5

Hydration of Ca-montmorillonite in ground water. (a) XRD profiles of the 001 reflection of montmorillonite during progressive hydration (peaks progressively shift toward the left). (b) Changes in the water layer thickness (1–4) over a period of ca. 300 h, determined by matching calculated XRD patterns (Modified after Warr and Berger 2006)

in high abundance, as seen with increasing humidity, but form steadily as part of a progressive layer expansion process. In *Fig. 5*, the hydration path of a pressed bentonite (containing >80% Ca-montmorillonite) powder is shown with an initial packing density of 0.94 g/cm^3. The initial hydration state of the pre-dried and laboratory-equilibrated montmorillonite consisted of 25–30% one WL and 70–75% two WL (*Fig. 5b*). No 0-WL structure was detected in this sample material. During the first phase of ground-water inflow both one and two WLs rapidly decreased during the first 30 h of hydration, with the formation of the three-WL state, reaching its maximum (ca. 60% abundance) after 40 h. A four-WL state appeared at this stage of the experiment, and corresponds to the disappearance of the one-WL, the continued decline of two-WL, and reduction of the three-WL structure. A steady state system was achieved after 150 h of hydration and comprised 15% two WL, 40% three WL, and 45% four WL. The gradient of the normalized water-uptake curve during the first 40 h of hydration matches well the rate of both three-WL and four-WL formation. The larger number of WLs developed compared to those developed under the high-humidity conditions (*Fig. 3*) relates to the total saturation with water that can be achieved in a flow-through reactor. A fully hydrated Ca-montmorillonite does develop more WLs than recorded for a typical Na-montmorillonite, which can be explained by the lack of swelling pressures related to the advanced hydration state at the beginning of the experimental run and the lower osmotic gradients associated with the hydration of bivalent interlayer Ca ions than that for monovalent Na ions (Jasmund and Lagaly 1993; Warr and Berger 2007). With a packing density of 0.94 g/cm^3, a 1 m thick barrier of Ca-bentonite would reach a saturation state in ca. 8 months, which is ca. ten times faster than a Na-bentonite of equivalent packing density (Warr and Berger 2007).

5 Applications to the Storage of CO_2: Reactions in Shales

CO_2 capture and sequestration (CCS) in deep geological formations has become an important option for reducing greenhouse gas emissions. For a significant impact on reducing the global gas emissions, huge volumes of CO_2 must be sequestered (Haszeldine 2009). Large sedimentary basins do provide geological formations with sufficient pore space to accommodate these large volumes of supercritical CO_2 (Benson and Cole 2008). Suitable formations should be deeper than 800 m and have sufficient porosity and a high permeability to allow injection of CO_2 at high flow rates without pressure build-up. For effective trapping a thick extensive seal is also required, typically composed of a low-permeability and low-porosity mudstone or shale.

Shales comprise the majority of cap rocks for hydrocarbon and hydrologic reservoirs with pore sizes typically <100 nm. As a result they are currently a central focus of study for the underground storage of CO_2. An important aspect of the shale seal is that it remains closed when coming into contact with supercritically charged CO_2 fluids, which tend to be both acidic and corrosive. When CO_2-charged fluids enter a reservoir, dissolution and crystallization reactions are expected to occur between the fluid and the wall rock of fracture surfaces and open pores. Although the temperatures of sequestration in many reservoirs will be low (<60 °C), published laboratory experiments typically use significantly higher temperatures (>150 °C) to enhance reaction rates so that alterations are more extensive and easier to study (Kaszuba et al. 2005). However, this approach assumes the same reaction mechanisms operate over a large thermal range that may well not be the case. The sluggish nature of reactions relevant to shallow crustal

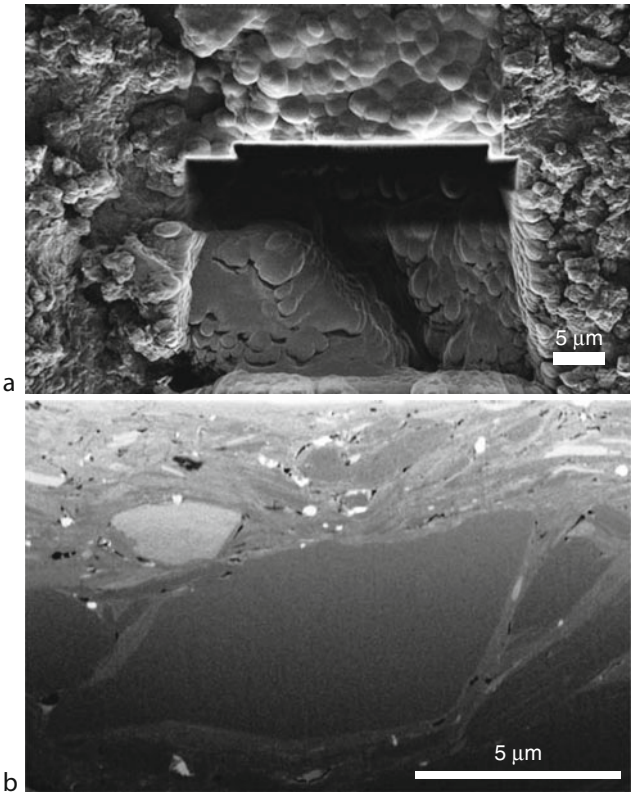


Fig. 6
(a) Focused ion beam thinning of a thin shale layer. (b) Backscattered electron image of the milled section

depth (e.g., ca. 1 km) can now be more successfully studied using high-resolution microscopic tools that allow 2D and 3D imaging of phase reactions and microchemical changes.

5.1 Three-Dimensional Reconstruction Using Combined Ion- and Electron-Beam Microscopy

Low-temperature reactions between CO_2-charged fluids and the minerals in cap shales are expected to commence as atomic scale coatings on surfaces exposed in micropores or small scale fractures (Cole et al. 2010). With time, these can develop into nanoscale coatings as altered mineral or neocrystallized material. A new type of analytical instrument to image and analyse such small alteration reactions is the dual (or cross) beam instruments that combine the imaging capabilities of field emission electron microscopy (resolution now <1 Å) with the milling power of a focused gallium ion beam.

We are currently studying the suitability of Mesozoic reservoirs in the state of Mecklenburg-Vorpommern, NE Germany, for CO_2 sequestration. In *Fig. 6*, the upper surface of a naturally

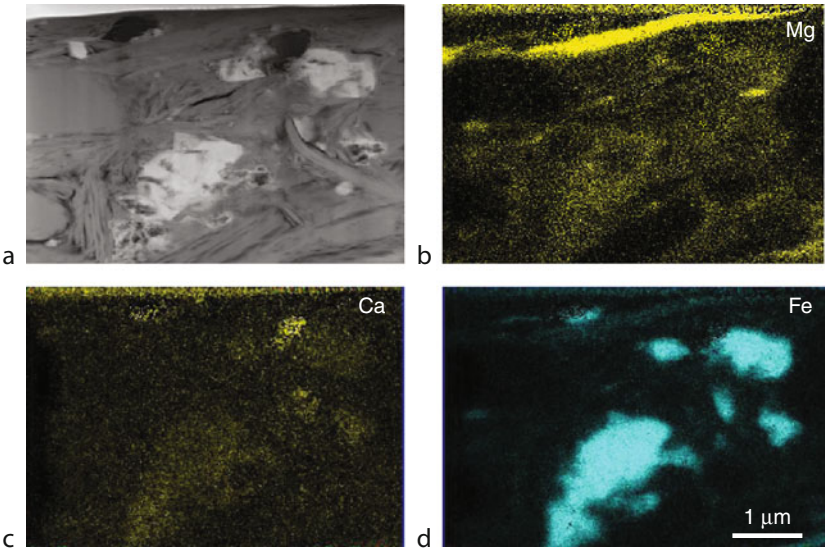


Fig. 7

Distribution of CO_2-sequestrating cations in a small-scale shale cap. (a) High-angle annular dark field STEM image of the shale fracture contact (upper boundary), (b) energy-dispersive X-ray intensity maps (K lines) of Mg, (c) K, and (d) Fe. All mapping was done in STEM mode on a focused-ion-beam thinned wafer

polished fracture surface in a shale laminate located within a sandstone bed is shown from the Löcknitz reservoir, situated near the German–Polish border. This sample represents a micro-reservoir analogue of a cap of shale overlying a sandstone reservoir composed mostly of quartz and feldspar. Gallium ion milling into the fracture surface, with a section depth of ca. 10 µm, can be used to successively slice the shale layer. The freshly cut surfaces can be simultaneously imaged by secondary or backscattered electrons (*Fig. 6b*) and used to map elemental distributions (*Fig. 7*). The location of bivalent cations, such as Ca, Mg, and Fe that are of particular importance for CO_2 sequestration, can be determined. When in contact with CO_2 fluids they can react to form carbonate minerals such as calcite ($CaCO_3$), magnesite ($MgCO_3$), and siderite ($FeCO_3$). In the Löcknitz reservoir, notable concentrations of Mg and ferrous Fe are available for reaction, and are located in minerals such as pyrite and chlorite.

An example of a 3D reconstruction attained from numerous slices into the micro-cap and reservoir analogue is shown in *Fig. 8*. This section shows the quartz and feldspar (darker grey shades) and the phyllosilicate cap of shale (lighter shades). The areas marked in blue represent the 3D microporosity, which is poorly connected, and the red-marked areas present very-low-permeability Fe–Mg-chlorite minerals. These highly oriented clay minerals in the polished thin film developed along fracture surfaces in the shale probably represent neocrystallization during localized fluid flow along small-scale faults. Their distribution should govern the dispersion paths of CO_2 once injected into the reservoir. Similar 3D microchemical reconstructions can also be mapped and used to track the diffusion of elements caused by dissolution and crystallization reactions after contact with reactive CO_2 fluids. Such analytical capabilities are ideal for investigating the dynamic changes in microporosity after fluid-rock interaction and allow the

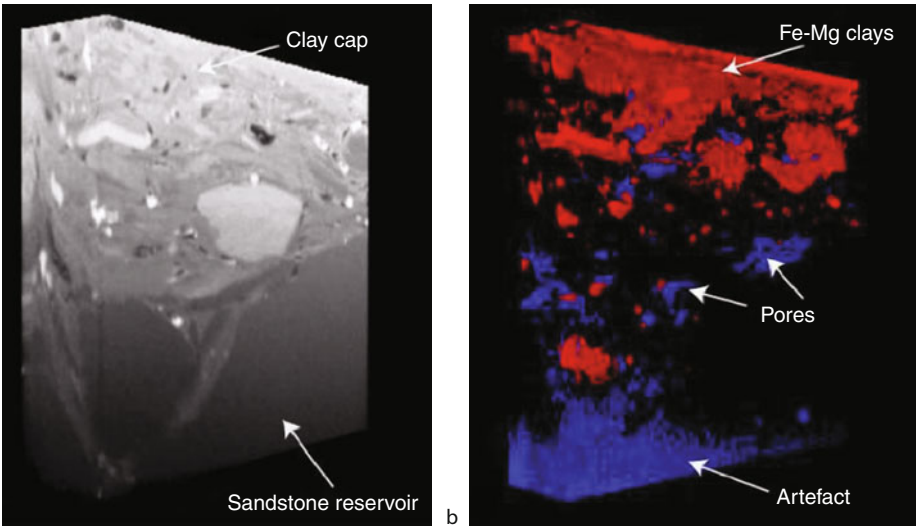


Fig. 8

3D nano-reconstruction of a polished fracture surface in sandstone from the Löcknitz reservoir based on cutting and imaging. (a) Backscatter electron image of the microstructure (varying grey scales represent different minerals by density). (b) 3D mapping of two components: red presents sealant Fe–Mg clays and blue the microporosity. Block reconstruction is 10 μm wide

effectiveness of the cap to be tested. Nanoscale characterization of reaction heterogeneities on altered surfaces and the quantification of reaction volume are much needed to assess the mechanisms and rates of CO_2 sequestration and help reveal the type of fluid-driven alteration that will occur within cap rock and reservoir. Further laboratory investigation of these samples after treatment with supercritically pressured CO_2 and brine fluids at reservoir-relevant conditions is currently in progress.

6 Conclusions and Outlook

Continuing improvements in particle detection and imaging techniques is having an important impact on researching current geoscientific issues such as the disposal of nuclear waste and the sequestration of atmospheric CO_2 in underground geological sites. The combined use of focused X-ray and electron and ion beams is offering improved characterization tools by which the mechanisms and rates of nanoscale mineral reactions can be better studied. Such analytical techniques in the laboratory combined with samples measured under controlled analytical environment (e.g., temperature, pressure, humidity, fluid circulation) enables close to real time, in situ monitoring of fluid–mineral reactions that provide more accurate analogues to shallow crustal environments of the Earth. Stability of clay minerals in subcrustal conditions is a topic of central importance as they form the most common type of geotechnical and geological

low-permeability seal in waste disposal or storage solutions. Monitoring clay mineral reactions in bentonites and shales as they interact with ground waters or supercritically CO_2-charged fluids in reservoir brines provides insights into the nature of dissolution and crystallization reactions that occur at the nanoscale interfaces of fractures and pores. New applications of combined ion and electron microscope are moving toward a more complete characterization of reaction products in three dimensions, particularly by combining serial slicing with electron-based imaging, dispersive X-ray spectroscopy, and electron backscatter diffraction techniques. Such achievements are leading to a more fundamental understanding of mineral–fluid reactions that should inevitably lead to safer containment solutions for unwanted waste in underground geological sites.

7 Cross-References

- Chapter 8, "Synchrotron Radiation and FEL Instrumentation"
- Chapter 15, "Scintillation Counters"
- Chapter 28, "Particle Detectors Used in Isotope Ratio Mass Spectrometry, with Applications in Geology, Environmental Science and Nuclear Forensics"
- Chapter 29, "Particle Detectors in Materials Science"

8 Analytical Equipment Used in this Study

Bruker D8 Advance theta-theta diffractometer with a LynxEye detector http://www.bruker-axs.com/d8_advance.html

MRI Humidity controlled reaction chamber http://www.mri-gmbh.de/

JEOL JEM 1210 High resolution transmission electron microscope and a JEOL JXA 840A Scanning electron microscope http://www.jeol.com/

Zeiss Auriga CrossBeam® (FIB-SEM) http://www.smt.zeiss.com/auriga

Acknowledgments

We would like to thank the "Deutsche Forschungsgemeinschaft" for their financial support in the form of large equipment grants.

References

Ardenne M, von Endell L, Hofmann U (1940) Investigation of the finest fraction of bentonite and clay soil with the universal electron microscope, Bericht der Deutschen Keramischen. Gesellschaft 21:209–227

Benson SM, Cole DR (2008) CO_2 sequestration in deep sedimentary formations. Elements 4: 325–331

Bethke CM, Reed JD, Oltz DF (1991) Long-range petroleum migration in the Illinois basin. AAPG Bull 75:925–945

Buseck P (1992) Principles of transmission electron microscopy. In: Buseck P (ed) Minerals and reactions at the atomic scale: transmission electron microscopy. Reviews in mineralogy, vol 27, pp 1–36

Cairns-Smith AG (1985) The first organisms. Sci Am 252:90–100

Chipera SJ, Carey JW, Bish DL (1997) Controlled-humidity XRD analyses: application to the study of smectite expansion/contraction. In: Gilfrich J et al (eds) Advances in X-ray analysis, vol 39. Plenum, New York, pp 713–722

Cole DR, Chialvoa AA, Rothera G, Vlcekbc L, Cummings PT (2010) Supercritical fluid behavior at nanoscale interfaces: implications for CO_2 sequestration in geologic formations. Phil Mag 90:2339–2363

Collins DR, Fitch AN, Catlow CRA (1992) Dehydration of vermiculites and montmorillonites: a time-resolved powder neutron diffraction study. J Mater Chem 2(8):865–873

Couture RA (1985) Steam rapidly reduced the swelling capacity of bentonite. Nature 318:50–52

Dickin (2005) Radiogenic isotope geology, 2nd edn. Cambridge University Press, Cambridge, p 472

Dickin AP (2008) Radiogenic isotope geology, 2nd edn. Cambridge University Press, Cambridge, p 510

Eberl DD, Drits VA, Srodon J (1998) Deducing growth mechanisms for minerals from the shapes of crystal size distributions. Am J Sci 298:499–533

Eitel W, Radczewski OE (1940) On recognition of montmorillonite clay minerals in supennicroscope pictures. Naturwissenschaften 28:397–398

Farges F, Sharpsa JA, Brown GE (1993) Local environment around gold (III) in aqueous chloride solutions: an EXAFS spectroscopy study. Geochim Cosmochim Acta 57:1243–1252

Ferrage E, Lanson B, Sakharov BA, Drits VA (2005) Investigation of smectite hydration properties by modeling of X-ray diffraction profiles. Part 1. Montmorillonite hydration properties. Am Min 90:1358–1374

Grathoff DH, Moore DM (1996) Illite polytype quantification using WILDFIRE calculated X-ray diffraction patterns. Clay Clay Miner 44: 835–842

Grathoff DH, Moore DM, Hay RL, Wemmer K (2001) Origin of illite in the lower Paleozoic of the Illinois basin: evidence for brine migrations. GSA Bull 113:1092–1104

Hanchar JM, Nagy KL, Fenter P, Finch RJ, Beno DJ, Sturchio NC (2000) Quantification of minor phases in growth kinetics experiments with powder X-ray diffraction. Am Miner 85:1217–1222

Haszeldine RS (2009) Carbon capture and storage: how green can black be? Science 325:1647–1652

Herbert HJ, Kasbohm J, Moog HC, Henning KH (2004) Long-term behaviour of the Wyoming bentonite MX-80 in high saline solutions. Appl Clay Sci 26:275–291

Hofmann H, Bauer A, Warr LN (2004) Behaviour of smectite in strong salt brines under conditions relevant to the disposal of low- to medium-grade nuclear waste. Clay Clay Miner 52:14–24

Jasmund K, Lagaly G (1993) Tonminerale und Tone – Struktur. Anwendungen und Einsatz in Industrie und Umwelt. Steinkopff-Verlag Darmstadt, Eigenschaften

Kaufhold S, Dohrmann R (2010) Effect of extensive drying on the cation exchange capacity of bentonites. Clay Miner 45:441–448

Kaszuba JP, Janecky DR, Snow MG (2005) Experimental evaluation of mixed fluid reactions between supercritical carbon dioxide and NaCl brine: relevance to the integrity of a geologic carbon repository. Chem Geol 217:277–293

Kühnel RA, van der Gaast SJ (1993) Humidity controlled diffractometry and its applications. Adv X Ray Anal 36:439–449

Laird DA, Shang C, Thompson ML (1995) Hysteresis in crystalline swelling of smectities. J Colloid Interface Sci 171:240–243

Langford RM (2006) Focused ion beams techniques for nanomaterials characterization. Microsc Res Tech 69:538–549

Lasaga AC (1981) Rate laws of chemical reactions. In: Lasaga AC, Kirkpatrick J (eds) Kinetics of geochemical processes, vol 8. Mineralogical Society of America, Blacksburg, pp 1–67

Lee JH, Peacor DR (1983) Intralayer transitions in phyllosilicates of the Martinsburg shale. Nature 303:608–609

Mee SJ, Hart JR, Singh M, Rowson NA, Greenword RW, Allen GC, Heard PJ, Skuse DR (2008) The use of focused ion beam for the characterisation of industrial mineral microparticles. Appl Clay Sci 39:72–77

Montes GH (2005) Swelling-shrinkage measurements of bentonite using coupled environmental scanning electron microscopy and digital image analyses. J Colloid Interface Sci 284: 271–277

Mooney RW, Keenan AG, Wood LA (1952) Adsorption of water vapor by montmorillonite. II. Effect of exchangeable ions and lattice swelling as measured by X-ray diffraction. J Am Chem Soc 74(6):1371–1374

Moore DM, Reynolds RC Jr (1997) X-ray diffraction and the identification and analysis of clay minerals, 2nd edn. Oxford University Press, New York, p 378

Nadeau PH, Wilson MJ, McHardy WJ, Tait JM (1984) Interstratified clays as fundamental particles. Science 225:923–925

Nagy KL (1995) Dissolution and precipitation kinetics of sheet silicates. In: Chemical weathering rates of silicate minerals. Reviews in mineralogy, vol 31. Mineralogical Society of America, Washington, DC, 173, pp 173–225

Obst M, Gasser P, Marrocordatos D, Dittrich M (2005) TEM-specimen preparation of cell/mineral interfaces by focused ion beam milling. Am Miner 90:1270–1277

Page R, Wenk HR (1979) Phyllosilicate alteration of plagioclase studied by transmission electron microscopy. Geology 7:393–397

Perdrial JN, Warr LN, Perdrial N, Lett MC, Elsass F (2009) Interaction between smectite and bacteria: implications for bentonite as backfill material in the disposal of nuclear waste. Chem Geol 264:281–294

Plancon I, Drits VA (2000) Phase analysis of clays using an expert system and calculation programs for X-ray diffraction by two- and three-component mixed-layer minerals. Clay Clay Miner 48(1):57–62

Pusch R (1992) Use of bentonite for isolation of radioactive waste products. Clay Miner 27: 353–361

Pusch R (2004) Mechanical properties of clays and clay minerals. In: Bergaya F, Theng BKG, Lagaly G (eds) Handbook of clay science. Elsevier, Amestrdam, pp 247–260

Reynolds RCJ (1985) NEWMOD a computer program for the calculation of one-dimensional X-Ray diffraction patterns of mixed-layered clays. Reynolds RCJr, 8 Brook Dr., Hanover, New Hampshire

Segl M, Mangini A, Bonani G, Hofmann HJ, Nessi M, Suter M, Wölfli W, Friedrich G, Plüger WL, Wiechowski A, Beer J (1984) ^{10}Be-dating of a manganese crust from Central North Pacific and implications for ocean palaeocirculation. Nature 309:54–543

Timur A, Toksoz MN (1985) Downhole geophysical logging. Annu Rev Earth Pl Sci 13:315–344

Wagner GA (1968) Fission-track dating. Earth Planet Sc Lett 4:411–415

Warr LN, Berger J (2004) Hydration of bentonite in natural waters: application of "confined volume" wet-cell X-ray diffractometry. Phys Chem Earth 32:247–258

Warr LN, Hofmann H (2003) In situ monitoring of powder reactions in percolating solution by wet-cell X-ray diffraction techniques. J Appl Crystallogr 36:948–949

Warr LN, Rice AHN (1994) Interlaboratory standardization and calibration of clay mineral crystallinity and crystallite size data. J Metamorph Geol 12:141–152

Yuan H, Bish DL (2010) NEWMOD+, a new version of the NEWMOD program for interpreting X-ray powder diffraction patterns from interstratified clay minerals. Clay Clay Miner 58: 318–326

28 Particle Detectors Used in Isotope Ratio Mass Spectrometry, with Applications in Geology, Environmental Science and Nuclear Forensics

Nicholas S. Lloyd[1] · *Johannes Schwieters*[1] · *Matthew S. A. Horstwood*[2] · *Randall R. Parrish*[2,3]

[1]Thermo Fisher Scientific, Bremen, Germany

[2]NERC Isotope Geosciences Laboratory, British Geological Survey, Keyworth Nottingham, UK

[3]University of Leicester, Leicester, UK

1 Introduction 686

2 Isotope Ratio Mass Spectrometry 687

2.1 Ion Sources 687

2.2 Mass Analyzers 690

3 Detectors 690

3.1 Faraday Cups and Amplifiers 691

3.2 Secondary Electron Multipliers 692

3.3 The Daly Detector 694

3.4 Energy Filters and Abundance Sensitivity 694

4 Applications 694

4.1 U–Pb Dating of Zircons by LA-MC-ICP-MS and ID-TIMS 695

4.2 Depleted Uranium in Urine 697

4.3 Nuclear Forensic Science 698

5 Conclusions 699

6 Cross-References 700

References 700

C. Grupen, I. Buvat (eds.), *Handbook of Particle Detection and Imaging*,
DOI 10.1007/978-3-642-13271-1_28, © Springer-Verlag Berlin Heidelberg 2012

Abstract: This chapter introduces the reader to mass spectrometry and the instruments used to determine high-precision isotope ratios. These instruments separate ion beams, of charged atomic particles with kinetic energies of several keV, by mass-to-charge ratio. Quantitative detection of these energetic charged particles is a key technology in mass spectrometry. For isotope ratio determination the main detector types are Faraday cups, the Daly detector, and discrete dynode secondary electron multiplier (SEM) ion counters. For high-precision applications, arrays of these detectors are arranged to collect several ion beams simultaneously. Examples are given for the application of these detectors in geology, environmental sciences, and nuclear safeguards.

1 Introduction

The fundamental principle of mass spectrometry is the quantitative separation of analyte ions with respect to their mass-to-charge ratio. This is achieved by creating positive or negative, atomic or molecular ions, accelerating and shaping these into an ion beam, separating the individual ions by mass-to-charge ratio (m/z), and then quantitatively detecting these.

A wide variety of instrument types exist, with one possible division being made between organic and inorganic mass spectrometry. Organic mass spectrometers are designed to preserve or partially fragment compounds, and then quantify molecular ions at a high mass resolution. Inorganic mass spectrometers are designed to separate the analyte into atomic ions, or in some cases to consistently create a simple molecular ion from an analyte which may exist in the sample in many different molecular compounds. Applications include the determination of elemental concentrations and determination of ratios between isotopes of the same element.

The focus of this chapter is isotope ratio mass spectrometry. The particles of interest for detection in these instruments are typically, but by no means exclusively, singly charged positive atomic ions of masses from across the periodic table, accelerated by a potential of several thousand volts. Ion beam currents are usually between 10^{-8} and 10^{-19} amps, but no single detector type covers this range adequately. The difference in intensity between two isotopes being measured may be greater than 10^6. Depending on the application, the required uncertainty of the measurement for isotope ratio applications is usually better than one part per thousand (‰), but for high-precision applications this can be better than ten parts per million (10^{-5}).

The desirable characteristics of particle detectors for these applications are therefore a high signal-to-noise ratio, and accurate calibration across a wide dynamic range. To achieve the highest analytical precisions, separated ion beams are detected simultaneously on an array of several detectors. The design of these 'multi-collector' mass spectrometers requires relatively compact detector types. In order to achieve high-vacuum conditions ($<10^{-8}$ mbar) to optimize ion transmission and minimize scattered ions, the detector must not degas, and may have to survive thermal conditioning of the instrument at temperatures greater than 180 °C ('bake-out'). This chapter will describe the common detector types used in isotope ratio mass spectrometry, and then illustrate their use with example applications.

2 Isotope Ratio Mass Spectrometry

Comprehensive coverage of the wide variety of mass spectrometers is beyond the scope of this chapter. The interested reader is referred to texts on inorganic mass spectrometry by Becker (2007), or Hoffmann and Stroobant (2007) for coverage of organic mass spectrometry. A useful glossary of mass spectrometry terms can be found in Mallet and Down (2009). Recommended books on geological applications of isotope ratios include Hoefs (2009) and Dickin (2005). The following section describes some of the principles of more commonly used mass spectrometers.

Traditional divisions have been made between instruments and geosciences laboratories researching stable isotope and radiogenic isotope systems. The former isotope systems are traditionally light in mass (H, C, O, N, S) and are introduced to the mass spectrometer as a gas. Stable isotope research investigates biological or physicochemical fractionation processes which affect ratios between isotopes of the same element in the natural environment.

The traditional radiogenic isotope systems are heavy in mass (e.g., Sr, Nd, Hf, Pb) and usually cannot be introduced to the mass spectrometer as a gas. These systems include an isotope that is the daughter of a parent radionuclide, effecting change to the ratio of two isotopes, for example, in a closed mineral system, as the radionuclide decays. A smaller number of laboratories research radiogenic isotope systems involving the noble gases (e.g., He and Ar) using specialized gas isotope instruments.

There is increasing interest in "non-traditional" stable isotope systems such as calcium and iron (e.g., Johnson et al. 2004), which cannot be introduced as gases. Mass spectrometers typically used for radiogenic isotope systems are therefore used for the applications.

The generalized mass spectrometer is comprised of a sample introduction system, ion source, a potential to accelerate ions, ion lenses to shape and focus an ion beam, a mass analyzer, and an ion detection system. This chapter briefly describes the common ionization sources, mass analyzers, and instruments that are used for the determination of high-precision isotope ratios.

2.1 Ion Sources

The aim of the ionization sources used for isotope ratio mass spectrometry is to produce either atomic ions with a common charge, or to consistently produce a simple molecular ion from an analyte that may be present in the sample in a variety of molecular compounds. The type of ionization source used is dictated by the original or processed sample type, and the type of information required.

Electron ionization (EI) is used for analytes that can be readily introduced as a gas, such as the traditional stable isotope systems (H, C, O, N, S), noble gases, and also for UF_6 for nuclear industry applications. A filament and accelerating voltage are used to bombard the gas phase sample with energetic electrons, creating analyte ions for the mass spectrometer.

Analyte solutions can be deposited and dried onto the surface of a rhenium ribbon or 'filament' for thermal ionization mass spectrometry (TIMS). In some instances solid particulate samples can be loaded directly. Electrical heating is used to evaporate and ionize the analyte; in some configurations these processes are separated between 2 or 3 filaments. Ion emission can be highly sensitive to the presence of other elements; therefore, careful chemical separation of the analyte is required for best precision and accuracy. Uranium and lead ionize at different temperatures; the two analyte fractions are typically loaded onto one filament and analyzed in two

separate runs using different filament currents. Activator compounds may be added to improve analyte ion emission, although for some elements with high first ionization potentials, including Hf and Th, thermal ionization is very difficult. Thermal ionization can produce a very stable ion beam with minimal energy spread, and is therefore capable of some of the most accurate and precise measurements.

A highly versatile ion source is used in inductively coupled plasma mass spectrometers (ICP-MS), where an argon plasma at ca. 10,000 K efficiently ionizes liquid or solid aerosols. A radio frequency (RF) electromagnetic field is used to excite argon atoms and produce a plasma. Laser ablation (LA) is a sample introduction technique that can be coupled to ICP-MS, in order to micro-sample directly from solid materials. Further information can be found in ICP-MS texts by Jarvis et al. (2003) and Nelms (2005). *Figure 1* shows a schematic of the Thermo Scientific Neptune multi-collector (MC) ICP-MS.

Spectral interferences are common for ICP-MS, therefore isotope ratios may need to be corrected by monitoring other masses that relate to the interfering element. Alternatively higher mass resolution may be used to filter polyatomic interference species from the analyte, e.g., $^{40}Ar^{16}O^{+}$ from $^{56}Fe^{+}$ (Weyer and Schwieters 2003). If the ionization energy of the interference is higher than that of the analyte, the interference species can be minimized by reducing the plasma temperature or using a collision cell. Vanhaecke et al. (2009) provide an informative tutorial review of the accurate use of ICP-MS for isotope ratio analysis.

Secondary ion mass spectrometry (SIMS) uses a focused ion beam from a separate source (usually O^{-} or Cs^{+}) to sputter, dissociate, and ionize material from a sample surface. Sub-μm spatial sampling resolution can be achieved; however, the technique is sensitive to the surface conditions of the sample. SIMS produces a large number of different molecular ions and different charge states in addition to atomic ions. Therefore high mass resolution is required to resolve spectroscopic interferences, in order to obtain accurate isotope ratio data.

A feature of all ion sources is 'mass fractionation' or bias. Where possible the isotope ratio of interest is corrected by a mathematical function to an isotope ratio that is known. For example, isotope ratios of radiogenic systems can be corrected using stable isotope ratio pairs within these systems (e.g., $^{88}Sr/^{86}Sr$, $^{146}Nd/^{144}Nd$, and $^{179}Hf/^{177}Hf$), which are considered invariant in nature. Double spikes of artificial isotopes added to a sample are also sometimes used to make independent corrections for fractionation, where the assumption of invariance may not hold.

In many isotope systems this is not possible, because all ratios are variable and therefore unknown. This is especially true for lighter stable isotope systems where natural fractionation processes affect all of the isotope ratios. In isotope dilution TIMS (ID-TIMS, see Sect. 4.1) an accurately measured quantity of 'spike' material with a known isotopic composition can be added to the sample, and the unknown isotope ratio can be accurately calculated. A 'double spike' with gravimetrically weighed artificial isotopes can also be added, e.g., a $^{236}U/^{233}U$ double spike has been used to reveal variability in $^{235}U/^{238}U$, an isotope ratio previously considered to be invariant in natural samples (Weyer et al. 2008).

Alternatively, the unknown ratio of the sample can be normalized to the same ratio measured in a known standard reference material. It is common for stable isotope systems to report isotope ratios that are relative to that of a standard reference material (δ or ε notation), rather than as absolute values. Mass bias cannot be assumed to be constant, or to be the same for both the sample and the reference material. Therefore great care must be taken when external normalization is used; the composition of the reference material and matrix should be similar to the sample, and it must be measured frequently.

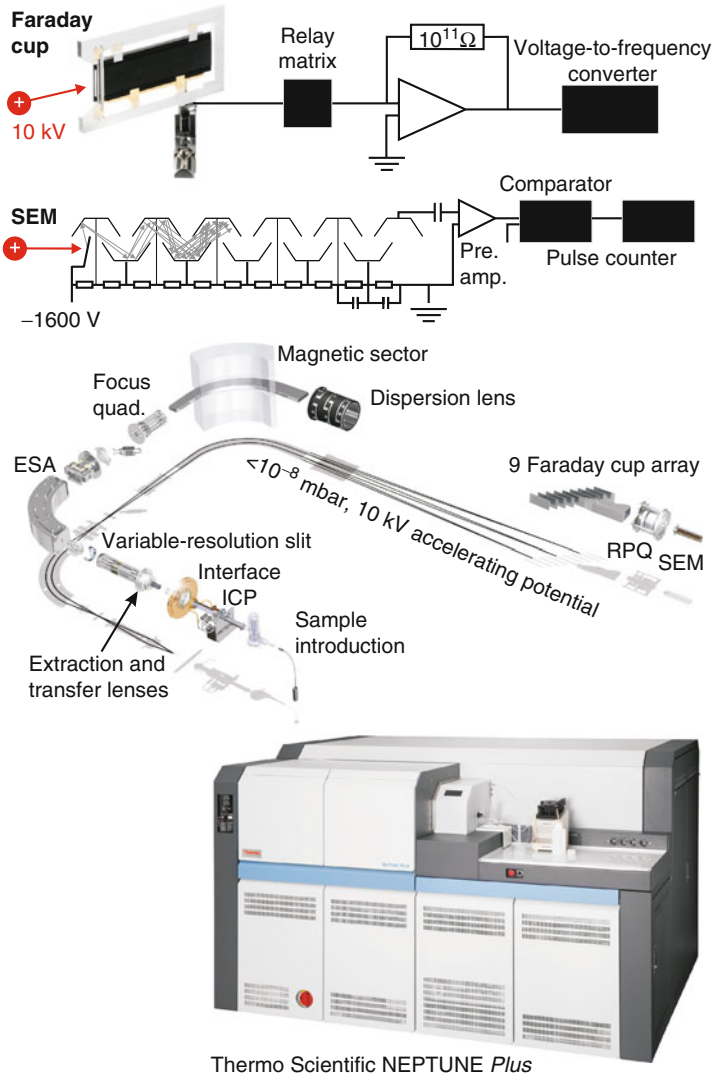


Fig. 1

Photograph of the Thermo Scientific NEPTUNE *Plus* MC-ICP-MS (Thermo Scientific, Bremen, Germany), a high-precision isotope ratio mass spectrometer, with schematic of the ion path from ICP ion source to multi-collector detector array. High sensitivity is achieved with a 10 kV acceleration potential, and a variable-resolution slit allows polyatomic interferences to be filtered from the analyte ions. Ion beams can be simultaneously measured on an array of 9 Faraday cups, with 10^{10}, 10^{11}, or 10^{12} Ω amplifiers connected via a relay matrix. Up to eight ion counting detectors can be installed, with or without RPQ abundance sensitivity filters on two channels

2.2 Mass Analyzers

The most common mass analyzer used for isotope ratio determination is the magnetic sector field (SF), which spatially separates ions by their mass-to-charge ratio (m/z), as well as by their energy. In ICP-MS, the spread of ion energies from the ion source is large compared to TIMS and gas source instruments, and an electrostatic analyzer (ESA) is typically used in a double focussing configuration to refocus ions of differing energy, either before or after the magnetic sector field mass analyzer. Alternatively some instruments have used a collision cell to thermally equalize the ions prior to mass separation (e.g. the GV Instruments Isoprobe-P). The stability of the magnetic sector field analyzer allows neighboring masses to be detected simultaneously on a multi-collector array. The magnetic sector field is also capable of high mass resolution, which may be required to separate isobars of polyatomic ions with mass-to-charge ratios that are close to that of the isotope of interest.

Where mass resolution is not sufficient for resolving an analyte from spectral interferences, accelerator mass spectrometry (AMS) can be used to 'strip' these isobaric interferences from the isotope of interest, enabling the accurate measurement of ratios significantly lower than 10^{-7}. Acceleration potentials of several MV are used, see, e.g., Steier et al. (2004).

The time-of-flight (TOF) mass analyzer spatially separates analyte ions along the path of the ion beam, allowing ions to be sequentially detected, which is potentially attractive for isotope ratio mass spectrometry. The quadrupole mass analyzer uses a combination of alternating radiofrequency electromagnetic and electrostatic fields to filter by mass-to-charge ratio. It is commonly used in low-resolution ICP-MS instruments to jump rapidly between different masses; these 'workhorse' single-collector instruments are used to quantify elemental concentrations across the whole mass spectrum and for lower-precision isotope ratios.

3 Detectors

The highest-precision isotope ratio measurements are made possible by multi-collector arrays. These allow each of the measured ion beams to be measured simultaneously on separate detectors, and therefore temporal differences in intensity from the ion source are measured on both sides of the ratio. A multi-collector array is illustrated in ❯ *Fig. 1*, with schematics of the two most commonly used detector types.

The detection efficiency and 'duty cycle' (ratio of time spent measuring the ion signal to idle time) of these instruments are also higher than for single-collector instruments, because all of the ion beams are measured continuously, rather than alternately. Thus, counting statistics can be improved for a limited sample size. This chapter will therefore focus on multi-collector mass spectrometry, although for many applications single-collector mass spectrometry has adequate precision.

A desirable feature for high-precision isotope ratio mass spectrometry is the 'flat-topped' peak. This is represented by a detector measuring a constant signal as a narrow ion beam is scanned across a comparatively wide entrance slit to the detector. It enables precise isotope ratios to be measured even if the centering of the ion beams into the detectors via the mass analyzer is not perfectly stable.

For many isotope systems, the isotope ratio of interest may be corrected by measuring mass bias from a separate known isotope ratio. This might be from an artificial spike added to the sample, or an invariant stable isotope pair. In many systems other masses are measured that relate to spectral interferences, allowing a correction to be made. This is especially true for ICP-MS, for example, in order to accurately determine the ratio $^{87}Sr/^{86}Sr$ it is necessary to measure 6 ion beams simultaneously ($^{82}Kr^+$, $^{83}Kr^+$, $^{85}Rb^+$, $^{86}Sr^+$, $^{87}Sr^+$, $^{88}Sr^+$). Precision and accuracy is therefore dependent on intercalibration of six separate detectors.

The mass of the isotopes of interest also dictates the field strength of the magnet, and therefore the distance between the ion beams on the focal plane of the detector array (i.e., dispersion). Instruments that are designed to measure several isotope systems therefore require detectors aligned for each of the systems, moveable detectors, and/or ion optics that allow the mass dispersion to be varied. Multi-collector arrays are therefore complex assemblies of high-specification ion detectors.

The following sections describe the principle detector types used in isotope ratio mass spectrometry. A brief mention should be made of array detectors, which incorporate several detector elements into one device, a review of these can be found in Barnes and Hieftje (2004). Although these are of potential interest, they have not so far been proven in commercial isotope ratio mass spectrometers.

3.1 Faraday Cups and Amplifiers

The Faraday cup is a device that ideally captures all ions entering it, and converts these to a measurable flow of electrons. It is an open-ended box, constructed from a conducting material such as graphite. An ion beam entering the Faraday cup produces a potential and thus current flow from ground, which can be amplified and measured.

The amplifier and measuring circuit is typically constructed from an operational amplifier connected with a feedback loop through a high-ohmic (10^9–10^{12} Ω) resistor. The intensity of the ion beam is measured as the potential across the amplifier, using a voltage-to-frequency converter. The circuit results in a high gain with low electronic noise. The detector and amplifier circuit are illustrated in *Fig. 1*.

The signal-to-noise ratio is improved for higher resistor values, for a 10^{12} Ω resistor the gain is a factor of 10 larger than for a 10^{11} Ω resistor, but the noise increases by only a factor of $\sqrt{10}$. Thus, in theory, the signal-to-noise ratio is improved by a factor of $10/\sqrt{10}$; in practice a two-fold improvement is achieved for workable 10^{12} Ω amplifiers.

The electronic noise of the amplifier dictates the precision of the electronic baseline measurement, which becomes increasingly significant for small ion beams. For example, for a typical 10^{11} Ω amplifier, a 0.2 fA (20 μV) uncertainty on the electronic baseline contributes a one permil uncertainty for a 0.2 pA ion beam. For a 0.2 nA ion beam this uncertainty is only one part per million. It is common for modern isotope ratio mass spectrometry instruments to be equipped with 10^{11} Ω amplifiers with a dynamic range extending to ca. 0.5 nA (50 V).

Resistance, and thus gain, are sensitive to temperature, which must therefore be kept constant. This is usually achieved by electrically heating the electronic circuits in an isolated housing. Another parameter of an amplifier is the exponential decay of a current through the circuit; for a 10^{11} Ω resistor it takes a couple of seconds for a current to reach 10^{-5} of its

initial value. For a 10^{12} Ω resistor the response is substantially slower, they are therefore not recommended for transient signals from laser ablation or chromatography sources.

On a multi-collector array, the amplifier gains can be readily and precisely cross calibrated using a stable reference voltage connected to each circuit. A relay matrix on the Thermo Scientific Triton (TIMS) and Neptune (MC-ICP-MS) instruments allows each amplifier to be connected to each Faraday cup in turn during an analysis, which improves external precision further by cancelling out relative gain differences.

Faraday cups can be designed to give near identical responses to each other, thus avoiding 'cup-factor' corrections or the need for 'multi-dynamic' analytical routines that cancel out channel differences by jumping isotopes across cups. Response is also similar across the width of each entrance slit, giving flat-topped peaks, thus the measured ratios are insensitive to the precise centering of the ion beams. A potential applied to an isolated entrance-defining slit can be used to repel secondary electrons and prevent escape of charge from the Faraday cup.

Faraday cups offer excellent linearity over a large dynamic range; they are also robust with a long lifetime. Multi-collector arrays of these detectors are capable of measuring $^{142}Nd/^{144}Nd$ isotope ratios with 2σ external relative precisions of 5×10^{-6}, see, e.g., Caro et al. (2003).

3.2 Secondary Electron Multipliers

Energetic ions impacting onto a surface (usually a metal oxide, e.g., Cu-Be-O) release secondary electrons, which, if attracted by a high field potential to hit another surface or part of the same surface, release further secondary electrons. If this process is repeated many times, the cascade multiplies the incidence of one or several ions into pulses of electrons that can be detected by electronic circuitry. This is the principle of the secondary electron multiplier (SEM). The gain across such a device is usually in excess of 10^5. These devices can either be of the continuous dynode (channel electron multiplier) or discrete dynode types. The discrete dynode SEM is illustrated in ❯ *Fig. 1*.

The electronic detection circuitry can either count individual pulses (cps, counts per second), or measure a current from the SEM. The first 'ion counting' detection mode is useful for precise and accurate measurement of small ion beams, typically limited to less than 1.2 Mcps (0.19 pA, equivalent to 19 mV on a Faraday cup with a 10^{11} Ω amplifier). Ideally precision is limited only by counting statistics. The second, analogue mode, is often used in single-detector mass spectrometers to increase the dynamic range of the detector, but having poorer linearity and baseline characteristics, it is not generally used in multi-collector mass spectrometry.

Discrete dynode electron multipliers have better linearity and yield stability (yield = registered counts/incident ions) when compared to continuous dynode electron multipliers. The latter can be readily reduced in size, which is useful for multiple ion counting (MIC) arrays. However, it is possible to accommodate several of the larger discrete dynode SEMs by using beam deflectors to guide ion beams into detectors positioned away from the main array. Beam deflectors can also be used to switch between detector types. This mechanism is used in the Element 2 XR (Thermo Scientific, Bremen, Germany), where the dynamic range of this single-collector high-resolution SF-ICP-MS is improved by seamlessly switching from dual mode SEM to Faraday cup (isotopes are jumped sequentially across a single detector slit).

The perfect ion counting detector would be linear with no electronic noise and constant yield. In practice, careful attention is required to use these detectors, starting with selection of

an operating voltage from a plateau region where there is a minimum of variation in yield with respect to change in operating voltage. If the operation voltage is set too high, it is possible to double count ions (yield >100%) and the lifetime of the detector is reduced. After the SEM is installed, or after the system is vented and the device is exposed to atmosphere, the plateau position and gain rapidly changes with use. A period of 'burn-in' using an ion beam at a high count rate is required to stabilize the device, conditioning the dynode surfaces. Subsequently, the position of the plateau with respect to operating voltage increases with use at a more moderate rate. The ion yield continues to vary with use and has to be frequently determined, and the operating voltage adjusted when the rate of change increases. A disadvantage of the SEM is yield drift and limited lifetime, which might typically be of the order of 1–2 years with moderate use. With care, some new designs of SEM have lasted more than 4 years of heavy use ($>7 \times 10^{12}$ registered counts).

The yield of an SEM can be determined from a reference material with known minor isotope ratio abundance, where the signal is corrected for mass bias using two major isotopes measured on Faraday cups. A lower-precision (10^{-3} range) single-collector measurement can be made by peak jumping a small ion beam between a Faraday cup and SEM(s). Alternatively the unknown sample data can be normalized to known standard ratios during a sequence of samples bracketed by matrix matched standards. This method is especially convenient for arrays of several SEMs, and combines the correction for mass bias.

SEMs are also sensitive to where and from which angle an ion strikes. The latter is especially a problem for continuous dynode SEMs, where the complex geometry of the dynode leads to a variety of electron cascade paths with different gains. The stability of the mass analyzer and precision of peak centering is therefore more significant than for other detector types.

SEM linearity was investigated by Richter et al. (2009). The linearity characteristics of SEMs manufactured by MasCom (Bremen, Germany) were compared with those from ETP (New South Wales, Australia). A correction is required for dead time, which is a finite time during which the pulse counting circuitry registers a count and is unable to register coincident pulses. This time is typically between 20 and 80 ns, and is significant for higher count rates when the probability of coincident ions increases.

The dead-time parameter may be determined experimentally, from a series of solutions with varying abundances of a minor isotope, such that the dead-time correction gives the correct value(s) for the ratio of the minor isotope. The Institute of Reference Materials and Measurements produces the IRMM-074 series for this purpose; for each of the solutions the ^{235}U /^{238}U ratio is approximately equal to one, so that mass-bias can be precisely measured using Faraday detectors, and varying abundances of ^{233}U can be used to test the linearity of an ion counter at both high and low count rates. Alternatively a series of dilutions of the same solution can be used, and the dead time chosen to give a constant ratio, see, e.g., Vanhaecke et al. (2009).

Clearly the Faraday cup offers advantages for linearity, stability, and gain calibration compared to the SEM. The advantage of the SEM in pulse counting mode is the extremely low electronic baseline, or so-called 'dark noise.' This is typically less than 1 cps, allowing quantification of ion beams smaller than 10^{-18} A (aA). In contrast, the variability of the electronic baseline for a Faraday cup with low-noise 10^{12} Ω amplifier is of the order of 10^{-16} A (0.1 fA), equivalent to 624 cps. The response and decay times of the SEM are also well suited to transient signals, such as those from laser ablation, whereas response and decay times are significant for low-noise high-gain Faraday cup amplifiers.

3.3 The Daly Detector

The Daly detector is a scintillation-type detector (Daly 1960). Ions impacting a conversion dynode release secondary electrons, which are accelerated to a scintillator target, resulting in photon emissions. A photomultiplier tube sitting behind the scintillator detects these events.

The photomultiplier is sealed independently of the instrument and can be replaced without venting the instrument. They can be less susceptible to yield drift, and typically have a larger dynamic range and longer lifetime than an SEM. Disadvantages of the Daly detector include size, need for a large accelerating voltage, and thermal stability of fast scintillator materials.

3.4 Energy Filters and Abundance Sensitivity

Electronic noise is not the only limitation for quantification of low ion beam intensities. Although abundance sensitivity is an issue relating to the mass spectrometer, it makes sense to discuss this in relation to high-sensitivity ion counting detectors. Abundance sensitivity is the term given for ions of a given mass 'tailing' onto neighboring masses. This may be significant where a low-intensity ion beam is measured that is close in mass to a high-intensity ion beam, see, e.g., Thirlwall (2001). Ball et al. (2008) extrapolated an exponential function for the analytical baseline measured between peaks, to accurately correct $^{234}U/^{238}U$ and $^{230}Th/^{232}Th$ ratios measured by MC-ICP-MS.

Abundance sensitivity, and therefore limits of quantification, can be improved using an energy filter. This can be placed in front of the ion detector and consists of ion lenses which decelerate the ion beam, and suppress ions with a low kinetic energy. The improvement in abundance sensitivity may be a factor of 10, with only minor loss of analyte detection efficiency. Various names are used for these devices, including Retardation Potential Quadrupole (RPQ) and Wide Aperture Retardation Potential (WARP) energy filter.

4 Applications

Mass spectrometry is used for a huge variety of scientific applications, from determining trace metal concentrations in drinking water to elucidating the components of complex molecules in genetic materials. Isotope ratios from mass spectrometry have been used to date the oldest rocks on earth, reveal details about sedimentary processes on the continental scale, measure climatic changes through earth history, and trace human migration patterns. They have been used to trace the provenance of food and drink, pharmaceuticals, and forensic samples. In the nuclear industry isotope ratios quantify the enrichment grade of uranium fuels, which may be independently verified so that these materials are not misappropriated for clandestine weapons programs. All of these applications, and many more, require detectors to accurately and precisely measure the intensity of ion beams. This section includes three examples of the use of particle detectors for geology, environmental science, and nuclear forensic science applications.

4.1 U–Pb Dating of Zircons by LA-MC-ICP-MS and ID-TIMS

Arguably the fastest growing area of isotope ratio science is in the use of laser ablation sampling. Here, a laser beam usually of wavelength 266, 213, or 193 nm, focused to a spot size of 10–100 μm, is used to ablate the surface of a material, which may be anything from a Roman coin to a Neanderthal tooth, a fish ear bone, a meteorite, or the oldest terrestrial material in the form of a single crystal of zircon. This latter application has seen an explosion in recent years as a tool for determining the provenance of sedimentary rock successions and dating individual volcanic, intrusive, and metamorphic rocks.

Zircon is an extremely tough and refractory mineral which can survive multiple erosive recycling events within sedimentary sequences, resist reincorporation within magmas, as well as high pressures and temperatures during metamorphism, without losing all of its original geochemical information. By incorporating uranium during its crystallization without lead, it provides an ideal material for U–Pb geochronology where the radiogenic in-growth of ^{206}Pb and ^{207}Pb from their parents, ^{238}U and ^{235}U respectively, allows precise and accurate determination of the age of the crystal.

In ablating minerals such as zircon using ultraviolet lasers, a plume of nanometer-sized particles is generated and swept to the mass spectrometer using a carrier gas flowing over the sample surface. The shorter the wavelength of the laser beam used, the smaller the absolute size and range of particles generated. Laser ablation systems are typically coupled to ICP-MS instruments which provide an ideal ionization source with sufficient energy to ionize virtually completely, most materials which pass into it. The particles pass through the argon plasma in around 1 ms and into the mass spectrometer. A single 3–25 ns pulse of a laser beam produces a plasma plume of material within which condensation and agglomeration processes result in a multitude of particle forms creating a stream of particles entering the argon plasma.

For a typical zircon analysis, a single laser pulse ablates 0.2–1 ng of material. The structure and shape of this particle stream is strongly influenced by the volume of the laser ablation cell housing the sample, the carrier gas density, and the transport tube volume. The greater the total volume of the transport system and the higher the density of the carrier gas, the more time is taken for ablated material to reach the mass spectrometer. In addition, packages of ablated material from successive laser pulses mix and become attenuated, obscuring spatial information, and reducing signal intensities. By improving transport dynamics, pulses on the order of 50–100 ms (Asogan et al. 2009) can be resolved in time as distinct from each other. Therefore spatial sampling information is retained and higher signal-to-noise ratios are achieved, helping to improve analytical precision and overcome background signal contributions. Spatially resolved sampling can be used to analyze separate growth zones within samples or individual crystals, recording separate intervals in time, which would otherwise be averaged and interpreted as one event.

Highly time-resolved signal pulses do, however, present a challenge to some detectors used in isotope ratio mass spectrometry. SEM detectors typically have dead times on the order of 20 ns, therefore signal pulses of 100 ms do not represent a problem. Typical laser ablation analysis using quadrupole ICP-MS or single-collector SF-ICP-MS, utilizing dual ion counting and analogue mode SEM detectors, are the instruments of choice for routine elemental concentration determination, where multiple elements of widely varying concentrations are quantified to percent level precision.

For isotope ratio work, however, where relative precisions of 50–200 ppm are desired (e.g., Sr isotope ratios), simultaneous acquisition using MC-ICP-MS or TIMS is required. Here, Faraday detectors are used with amplifier boards typically utilizing resistors which have a signal decay

time (tau) of 2 s to reach 10^{-5} of the peak signal. These slow-response detectors are therefore not appropriate to resolve highly spatially zoned materials using laser ablation. Laser ablation analysis using these detectors usually involves integrating many (ca. 30) seconds worth of data to achieve the precisions required, masking and sacrificing much of the time-resolved information which could otherwise be gleaned.

An improvement on this is to use multiple ion counting (MIC) arrays. These typically combine three or more ion counting channels allowing much greater time-resolved interpretation of the data. When combined with Faradays in a multi-detector array, the rapid response time of the ion counters can clearly be seen to differentiate them from the slow-response Faradays, apparently shifting in time the defined signal peak and apparently making the measurement non-simultaneous. This can be overcome by using single-pulse ablation and integrating the whole of the peak signal as one, using a selection window for the integration which includes the full peak width from each detector. An example of this was shown by Cottle et al. (2009) as applied to U–Pb geochronology, where data from individual laser pulses accurately quantified the age of zircons and combination of 10–30 pulses provided equivalent statistics to a normal 30-s, 150-pulse analysis.

U–Pb geochronology is a critically important methodology for accurately and precisely determining the age of geological materials and rates of change within the geological record. These observations and measurements underpin modern interpretations of earth processes and dynamics that are relevant to understanding climate change. ID-TIMS is the method of choice for this work where spike materials can be gravimetrically calibrated and uncertainties accurately quantified and minimized to a level which enables age precision <0.1% to be achieved. The detectors used in this work are the same as those already described, namely Faraday detectors and discrete dynode SEM ion counters or Daly detectors.

They are, however, used in a different way than during laser ablation analysis. The solid sample material, dried onto a filament and heated to 1,200–1,600 °C to ionize Pb or U sample fractions, emits a constant very stable ion beam over perhaps 2 h or more depending on the size of the sample load. This allows long integration times and high counting statistic precisions to be achieved with inter-calibration of the different detectors performed within the measurement routine. This also allows the differential response times of the detectors to be overcome. The requirement for chemical separation of the material is a slight hindrance, increasing the blank contribution to the sample, but modern chemical separation techniques can keep the Pb blank contribution well below 1 pg, and in the low fg range for U. Using a silica-gel emission enhancer, Pb detection efficiencies up to 10 percent can be achieved for ID-TIMS, compared to ≤0.8% typical for MC-ICP-MS.

Sample Pb loads of only a few pg can therefore be accurately quantified using ID-TIMS, although this method typically precludes spatially resolved information. LA-MC-ICP-MS is a rapid technique that can be used to analyze large quantities of detrital zircons, for example, from which isotopically disparate populations can be defined, revealing changes in sedimentary source materials. The right methodology therefore needs to be used to address the problem in question. ID-TIMS and LA-MC-ICP-MS can be used as complementary techniques, where the rapid spatially resolved analysis capability of LA-MC-ICP-MS can be used to screen for the most appropriate materials for high-precision ID-TIMS work.

U and Pb isotope analysis by ID-TIMS and MC-ICP-MS can suffer from polyatomic interferences. Usually assigned to inefficient chemical separation procedures, these polyatomic ions are usually thought to result from organic materials within the samples, and/or residues from the separation resin, forming long-chain polyatomic species interfering on the Pb and U

mass spectrum. This can be a particular problem when dealing with complex matrices, such as urine, containing low concentrations (ca. 1 ppt by weight) of analyte. Such interferences on the very small ^{234}U, ^{235}U, and ^{236}U peaks were a significant concern for Parrish et al. (2006).

Doubly charged ions can also be an issue, these appear at half the mass of the ion (i.e., at their mass-to-charge ratio), where they may interfere with the isotope of interest. SIMS analysis commonly requires correction for polyatomic and doubly charged interferences generated by the ion sputtering process. LA-MC-ICP-MS can also contribute these species. Horstwood et al. (2008), and references therein, detailed the effect of doubly charged REE ions, Ca dimers (Ca_2^+), and a $CaPO^+$ polyatomic on the Sr mass region and in particular ^{87}Sr. Although very small (only a few thousand counts per second on a Faraday), this latter polyatomic interference was enough to bias the $^{87}Sr/^{86}Sr$ ratio by >1% in low Sr concentration (ca. 50 ppm by weight) samples and typically 0.1–0.3% in samples with 100–200 ppm Sr. Analysis protocols to mitigate and/or correct for these effects can be employed to reduce the inaccuracy to ca. 0.3‰, but usually at the sacrifice of detection efficiency, which often then renders precise determination impossible.

Poor sample utilization (sample through to ion detection) is a serious problem in mass spectrometry and in particular for TIMS and ICP-MS techniques. TIMS methodologies have been honed over decades to optimize procedures for particular elements and utilize emission enhancers to improve the number of atoms detected from the total number of atoms sampled. In this respect ICP-MS is more restricted, and most of this efficiency is lost in the sample transport to the plasma and also in extracting the ions generated in the plasma into the mass spectrometer. Recent improvements in this latter aspect have dramatically improved sample utilization to 2–4% (compared with a more typical 0.6–0.8%) which directly impacts on the ultimate precisions which can be achieved and/or the size of sample which can be analyzed. For example, in nuclear forensic determinations it should now be possible to achieve <5% precision on the $^{235}U/^{238}U$ ratio from 500 femtograms of uranium (ca. 0.5 μm diameter particle). This higher 'sensitivity' also means that better resolution of discrete spatial changes can be achieved using laser ablation techniques when combined with approaches to improve the signal-to-noise ratio of the analysis (e.g., low-volume ablation cell technology and single-pulse methodologies).

4.2 Depleted Uranium in Urine

Exposure to aerosols containing depleted uranium (DU) has been cited as a potential cause of 'Gulf War Illnesses' that have afflicted many veterans of the 1991 Gulf War. The impact of high-velocity DU projectiles on heavy armor results in the creation of fine aerosols of uranium oxide that can be inhaled deep into the lungs. Exposure to these potentially toxic inhaled particles can be monitored by measurement of the uranium isotope ratio of urine passed by the veterans. Parrish et al. (2006) report the results of an inter-laboratory comparison study of veterans urine by multi-collector and single-collector ICP-MS instruments, attempting to quantify the presence of a depleted uranium isotope ratio signature which could indicate likely exposure to DU particulates.

Enrichment of the fissile isotope ^{235}U from natural uranium ($^{238}U/^{235}U$ ca. 137.88) for use in nuclear fuels results in by-product 'depleted uranium' with $^{238}U/^{235}U$ ratios up to ca. 500. A small addition of this to the natural uranium present in the body changes the isotope ratios of biomonitoring samples (e.g., urine or blood), which can be detected by high-precision analysis.

Isotope ratio deviations of ca. 1% can be 'detected' as significant outside of total analytical uncertainty. ^{236}U, essentially absent in natural uranium, is produced in nuclear reactors and typically 'contaminates' DU, providing a fingerprint to the production cycle of this material. Typical ^{236}U/^{238}U ratios in urine samples with a DU contribution are of the order of 10^{-6}. Typical quantities of total uranium in urine are only 1–5 ng/l.

Detecting these small shifts in ratio in such small amounts of material requires careful chemical concentration and purification procedures. Even after this, 1 ng/ml of processed sample still represents a significant challenge to accurately quantify a distinct shift in the natural composition of U and determine whether ^{236}U signals are significant above backgrounds.

Abundance sensitivity (tailing of ^{238}U onto m/z 236) and hydride ($^{235}U^{1}H^{+}$) corrections are the most significant, contributing potentially 100% of the total m/z 236 signal. Stable, highly linear ion counters are therefore required so that ^{235}U signals of up to several hundred thousand counts per second can be measured accurately alongside 236 signals of only a few tens of counts per second. Careful characterization of the ion counter and calibration of the dead time, linearity, and plateau voltage is required. Faraday signals for ^{238}U may also be small (<0.5 pA) and amplifier noise can become a significant problem for precision with very-low-concentration samples. Combined with the potential for polyatomic interferences resulting from variable chemical separation efficiencies, the accurate determination of U isotopes in urine and the assignment of an appropriate uncertainty is a tough analytical challenge.

Parrish et al. (2006) noted all these problems and reported in a later paper (Parrish et al. 2008) that none of the 466 veterans who volunteered for testing showed DU in their urine. Although this didn't rule out the possibility of undetectable exposure, this later study of a different population demonstrated that the technique was capable of detecting small levels of DU contamination in urine >20 years after exposure.

4.3 Nuclear Forensic Science

A significant impetus for early ion counting development was for nuclear mass spectrometry applications, where minor uranium isotopes ^{234}U and ^{236}U are often measured with ratios to ^{238}U in units of parts per million. For example, during the 1960s, the Daly detector was developed at the UK's Atomic Weapons Research Establishment (Daly 1960), and work at Knolls Atomic Power Laboratory saw improvements in SEM design (Dietz 1965). More recently a significant driver for progress in commercial ion counting technologies has been for geochemistry applications, including high-precision U–Pb dating of zircons by ID-TIMS (see Sect. 4.1).

An area of international political concern is nuclear proliferation, and the safeguards systems which ensure that nuclear materials are not misappropriated. Nuclear forensics is the process that allows nuclear materials to be traced to origin, and interpretations to be made about processing technologies employed and the intended use of such materials. *Nuclear Forensics Analysis* by Moody et al. (2005) is a highly readable book on the subject, and a paper by Mayer et al. (2007) provides a short summary of the subject.

Of special interest are μm diameter particles, with pg masses, which would be especially hard to clean-up and conceal. They can be formed by the hydrolysis of UF_6 escaping from enrichment processes, see, e.g., Kips et al. (2009). A paper by Donohue (1998) describes the International Atomic Energy Authorities (IAEA) strategies for safeguards monitoring, where

swipe samples are taken from suspected and declared nuclear facilities. These could be contaminated by uranium and plutonium bearing particulates with isotope ratios indicative of origin and intended use. Safeguards samples are investigated by a variety of techniques, including SIMS, fission track, and TIMS (FT-TIMS). Sub-picogram detection levels were possible for uranium using TIMS (Finnigan-MAT 262) with an ion counting system and energy filter. The traditional scheme for locating and analyzing uranium and plutonium particles is fission track analysis followed by direct loading of particles of interest onto rhenium ribbons for analysis by TIMS.

For uranium and plutonium samples, none of the isotope ratios can be assumed to be invariable, and therefore it is not possible to make an internal correction for mass bias. For TIMS it is possible to run a total evaporation method, where the filament is heated until ion emission effectively ceases. The entire signals for each isotope are integrated and the ratios calculated with minimal bias. The method is used in ASTM C1672 (ASTM International, West Conshohocken, USA). With ion counting, it is possible to measure femtogram samples.

Secondary ion mass spectrometry (SIMS) is a useful and commonly employed tool for nuclear forensics and safeguards. Ranebo et al. (2009) discusses the advantages of large-geometry SIMS. Polyatomic ion interferences that commonly cause inaccurate isotope ratio data for SIMS can be resolved, whilst maintaining sensitivity that is not possible on standard-geometry SIMS operated in high-resolution mode. A sample utilization of 1.2% allows 1 μm diameter (sub-pg) particles of interest to be analyzed on arrays of SEM detectors. As for ICP-MS, $^{235}U^{1}H$ forms on ^{236}U and cannot be resolved spectroscopically, $^{238}U^{1}H$ must therefore be monitored and a correction made for m/z 236.

LA-MC-ICP-MS is a technique that could potentially be used for safeguards particle analysis. It offers rapid sampling directly from samples, with significantly lower instrument costs compared to large-geometry SIMS.

In Lloyd et al. (2009) we demonstrated the technique for the determination of precise and accurate uranium isotope ratios from more than one hundred larger uranium oxide particles (20–64 μm diameter). The MC-ICP-MS used was a VG Elemental Axiom, a first-generation instrument with a single continuous dynode electron multiplier, and eight moveable Faraday cups. Data were externally corrected for mass bias, ion counter yield, $^{235}U^{1}H^{+}$ on $^{236}U^{+}$, and $^{238}U^{+}$ tailing onto $^{236}U^{+}$. Relative 2σ uncertainties for $^{235}U/^{238}U$ and $^{236}U/^{238}U$ were typically 0.4 and 2.7%, respectively. The isotopic signatures of these particles, at this level of precision, could reveal which gaseous diffusion plant produced the depleted uranium.

Challenges for analyzing fine particles by LA-MC-ICP-MS from safeguards samples include LA targeting and sample utilization. Modern MC-ICP-MS instruments offer significantly improved sample utilization (analyte sensitivity), options for counting all of the uranium isotopes on discrete dynode SEMs, as well as ion energy filters for improved abundance sensitivity.

5 Conclusions

High-precision isotope ratio mass spectrometry is made possible by arrays of high-specification ion detectors. Faraday cups with high-gain low-noise amplifiers are used for the highest-precision measurements. These detectors are robust, stable, linear, and readily cross-calibrated. For small ion beam intensities (<0.1 pA) low-electronic-baseline detectors, such as SEM or Daly ion counting detectors, are required.

6 Cross-References

- Chapter 2, "Electronics Part I"
- Chapter 3, "Electronics Part II"
- Chapter 4, "Data Analysis"
- Chapter 5, "Statistics"
- Chapter 15, "Scintillation Counters"
- Chapter 26, "Accelerator Mass Spectrometry and its Applications in Archaeology, Geology and Environmental Research"
- Chapter 29, "Particle Detectors in Materials Science"

References

Asogan D, Sharp BL, O' Connor CJP, Green DA, Hutchinson RW (2009) An open, non-contact cell for laser ablation-inductively coupled plasma-mass spectrometry. J Anal At Spectrom 24:917–923

Ball L, Sims KWW, Schwieters J (2008) Measurement of $^{234}U/^{238}U$ and $^{230}Th/^{232}Th$ in volcanic rocks using the Neptune MC-ICP-MS. J Anal At Spectrom 23:173–180

Barnes JH, Hieftje GM (2004) Recent advances in detector-array technology for mass spectrometry. Int J Mass Spectrom 238:33–46

Becker JS (2007) Inorganic mass spectrometry: principles and applications. Wiley, Chichester

Caro G, Bourdon B, Birck J-L, Moorbath S (2003) ^{146}Sm-^{142}Nd evidence from Isua metamorphosed sediments for early differentiation of the Earth's mantle. Nature 423:428–432

Cottle JM, Horstwood MSA, Parrish RR (2009) A new approach to single shot laser ablation analysis and its application to in situ Pb/U geochronology. J Anal At Spectrom 24:1355–1363

Daly NR (1960) Scintillation type mass spectrometer ion detector. Rev Sci Instrum 31:264–267

Dickin AP (2005) Radiogenic isotope geology, 2nd edn. Cambridge University Press, Cambridge

Dietz LA (1965) Basic properties of electron multiplier ion detection and pulse counting methods in mass spectrometry. Rev Sci Instrum 36:1763–1770

Donohue DL (1998) Strengthening IAEA safeguards through environmental sampling and analysis. J Alloys Compd 271–273:11–18

Hoefs J (2009) Stable isotope geochemistry, 6th edn. Springer-Verlag, Berlin

Hoffmann E, Stroobant V (2007) Mass spectrometry: principles and applications, 3rd edn. Wiley, Chichester

Horstwood MSA, Evans JA, Montgomery J (2008) Determination of Sr isotopes in calcium phosphates using laser ablation inductively coupled plasma mass spectrometry and their application to archaeological tooth enamel. Geochim Cosmochim Acta 72:5659–5674

Jarvis KE, Gray AL, Houk RS (2003) Handbook of iductively coupled plasma mass spectrometry. Viridian Publishing, Woking

Johnson CM, Beard BL, Albarede F (2004) Geochemistry of non-traditional stable isotopes, In: Rosso JJ (ed) Reviews in mineralogy and geochemistry, vol. 55. Mineralogical Society of America and The Geochemical Society, Washington, DC

Kips R, Pidduck AJ, Houlton MR, Leenaers A, Mace JD, Marie O, Pointurier F, Stefaniak EA, Taylor PDP, Van den Berghe S, Van Espen P, Van Grieken R, Wellum R (2009) Determination of fluorine in uranium oxyfluoride particles as an indicator of particle age. Spectrochim Acta B 64:199—207

Lloyd NS, Parrish RR, Horstwood MSA, Chenery SRN (2009) Precise and accurate isotopic analysis of microscopic uranium-oxide grains using LA-MC-ICP-MS. J Anal At Spectrom 24:752–758

Mallet AI, Down S (2009) Dictionary of mass spectrometry. Wiley, Chichester

Mayer K, Wallenius M, Fanghänel T (2007) Nuclear forensic science: from cradle to maturity. J Alloys Compd 444–445:50–56

Moody KJ, Hutcheon ID, Grant PM (2005) Nuclear forensic analysis, 1st edn. CRC Press, Boca Raton, FL

Nelms SN (2005) Inductively coupled plasma mass spectrometry handbook. Blackwell Publishing, Oxford

Parrish RR, Thirlwall MF, Pickford C, Horstwood M, Gerdes A, Anderson J, Coggon D (2006)

Determination of $^{238}U/^{235}U$, $^{236}U/^{238}U$ and uranium concentration in urine using SF-ICP-MS and MC-ICP-MS: an interlaboratory comparison. Health Phys 90:127—138

Parrish RR, Horstwood M, Arnason JG, Chenery S, Brewer T, Lloyd NS, Carpenter DO (2008) Depleted uranium contamination by inhalation exposure and its detection after 20 years: implications for human health assessment. Sci Total Environ 390:58–68

Ranebo Y, Hedberg PML, Whitehouse MJ, Ingeneri K, Littmann S (2009) Improved isotopic SIMS measurements of uranium particles for nuclear safeguard purposes. J Anal At Spectrom 24:277–287

Richter S, Alonso A, Aregbe Y, Eykens R, Kehoe F, Kühn H, Kivel N, Verbruggen A, Wellum R, Taylor PDP (2009) A new series of uranium isotope reference materials for investigating the linearity of secondary electron multipliers in isotope mass spectrometry. Int J Mass Spectrom 281:115—125

Steier P, Golser R, Kutschera W, Priller A, Vockenhuber C, Winkler S (2004) VERA, an AMS facility for "all" isotopes. Nucl Instrum Methods Phys Res, Sect B 223–224:67–71

Thirlwall M (2001) Inappropriate tail corrections can cause large inaccuracy in isotope ratio determination by MC-ICP-MS. J Anal At Spectrom 16:1121–1125

Vanhaecke F, Balcaen L, Malinovsky D (2009) Use of single-collector and multi-collector ICP-mass spectrometry for isotopic analysis. J Anal At Spectrom 24:863–886

Weyer S, Schwieters JB (2003) High precision Fe isotope measurements with high mass resolution MC-ICPMS. Int J Mass Spectrom 226:355–368

Weyer S, Anbar AD, Gerdes A, Gordon GW, Algeo TJ, Boyle EA (2008) Natural fractionation of $^{238}U/^{235}U$. Geochim Cosmochim Acta 72:345–359

29 Particle Detectors in Materials Science

Xin Jiang · Thorsten Staedler
University of Siegen, Siegen, Germany

1 Introduction *704*

2 Detector Application in Materials Science *705*

3 The Low-Temperature Synthesis of Diamond Films *706*

4 Structure Characterization: SEM, TEM, XRD, and Raman *709*

5 Properties/Applications: Electron Field Emission *714*

6 Material Development for Particle Detection *716*

7 Conclusions *716*

8 Cross-References *716*

References *716*

C. Grupen, I. Buvat (eds.), *Handbook of Particle Detection and Imaging*,
DOI 10.1007/978-3-642-13271-1_29, © Springer-Verlag Berlin Heidelberg 2012

Abstract: Particle detectors play a key role in today's materials science. Being generally based on the interaction of particles with matter they naturally form the foundation of any analytical tool to derive information on the structure of materials. Therefore, advances in particle detector technology are closely interrelated with improvements in instrumentation as well as an increased knowledge gain with respect to the corresponding interaction underlying the method. Illustrated by an example of a chemical vapor deposition (CVD)-based diamond synthesis process the correlation between particle detector technology and the different stages of process and materials characterization will be shown.

1 Introduction

From the earliest days of high energy physics in the 1930s to the latest twenty-first century initiatives, the innovative ideas and technologies of particle physics have entered the mainstream of society to transform the way we live. Selected examples illustrate a long and growing list of beneficial practical applications with contributions from particle physics.

In order to develop advanced materials to support new technologies it is essential to have detailed knowledge about the linear interrelationship between processing, structure, properties, and performance of those materials. These terms represent the core of the scientific research area known as materials science and engineering, whereby the terms "processing" and "performance" are associated with materials engineering and the terms "structure" and "properties" fall into the field of materials science. The key that allows gaining insight into individual components of this interrelationship is the interaction of the item of interest with appropriate probes such as electrons, ions, neutrons, and photons. In turn, the quantitative registration of the resulting interaction products along with a fundamental understanding of the nature of interaction forms the basis of modern materials science (see ❯ *Fig. 1*).

With this in mind the importance of particle detectors with respect to the field of materials science becomes quite obvious. The strength of each and every analytical tool with which one strives to shed light on the structure as well as properties of the materials of interest is

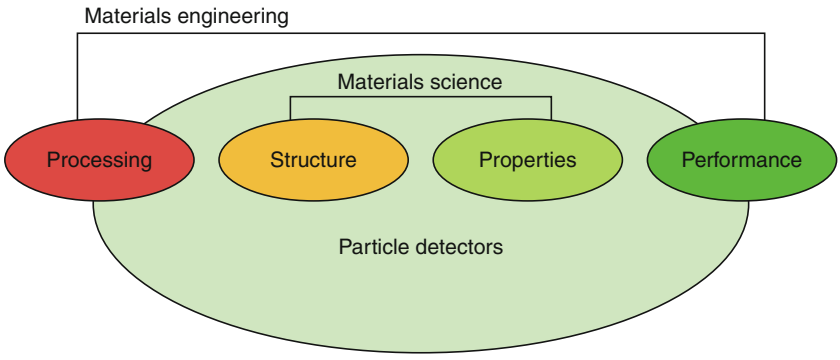


Fig. 1
Particle detector technology does have a crucial impact on all stages of the materials science and engineering chain

directly related to the performance of its detector. In the following, we will try to show the strong connection between particle detectors and materials science based on the example of microwave-assisted growth of diamond by a chemical vapor deposition process (MW-CVD). In particular, we will show fields of application of particle detectors in the area of gas-phase diagnostics, electron microscopy, x-ray diffraction, and RAMAN spectroscopy.

2 Detector Application in Materials Science

Before we go into detail about the individual analytical tools mentioned above and the particle detectors associated with them, we want to spend a few words about particle detectors in materials science in general.

Particle detectors themselves are based on the interaction of the particles of interest with detector material. Here the physical processes involved differ for particles that feature a charge and those that do not. In the case of charged particles, it is almost exclusively their electromagnetic interaction that is exploited for detection. If a charged particle passes through matter three possible phenomena can be observed: Ionization, emission of Cherenkov light (Cherenkov and Pavel 1934), and transition radiation (Allison et al. 1991). For particles that travel faster than the phase velocity of light in a material where Cherenkov light is emitted, slower particles will cause ionization. However, in case of inhomogeneous materials even slower particles might result in the emission of Cherenkov light. This phenomenon is called transition radiation.

As for neutral particles we will restrict ourselves to neutrons and photons as these are the ones that play an important practical role in materials science. Neutrons will typically interact with the nuclei of the matter they permeate and hereby generate charged secondary particles. Photons on the other hand feature three fundamental mechanisms with which they can interact with matter depending on their energy, photo effect, Compton effect (Compton 1922), and pair generation, respectively. For photons with energy below 100 keV the photo effect is the dominating mechanism, around 1 MeV the Compton effect is the prime one, and above 2 MeV pair generation becomes the most important interaction process.

In the following few paragraphs, some specific analytical tools of materials science are introduced along with their associated particle detector technology. As already mentioned we will illustrate this using the example of CVD-based diamond synthesis (see *Fig. 2*).

During the last decades, the research on diamond films has attracted the strongest attention of materials scientists. By virtue of its very strong chemical bonding, diamond features extreme, unique, and diverse properties such as very high mechanical hardness, dielectric strength, energy band gap, thermal conductivity, radiation resistance, corrosive resistance and electrical resistivity as well as a very low thermal coefficient of expansion, and others. These properties endow diamond with a large application potential (Field 1979; Angus and Hayman 1988; Yarbrough and Messier 1990; Wilks and Wilks 1994; Davis 1992; Pan and Kania 1995). A typical application is the utilization of diamond in the context of cutting tools. With a thermal conductivity of 20 $W/(cm^{\circ}C)$, diamond is unparalleled as a thermal conductor. In addition its high thermal conductivity suggests diamond as the ideal heat exchange material (heat sink and heat spreader). Diamond has been applied as an electrically insulating thermal conductor for various electronics applications. Recently, high-power laser diodes have also been mounted on diamond in order to improve the performance and increase the output power of the diodes. Very large integrated circuit (VLSI) and multiple chip module (MCM) compacts also the use of

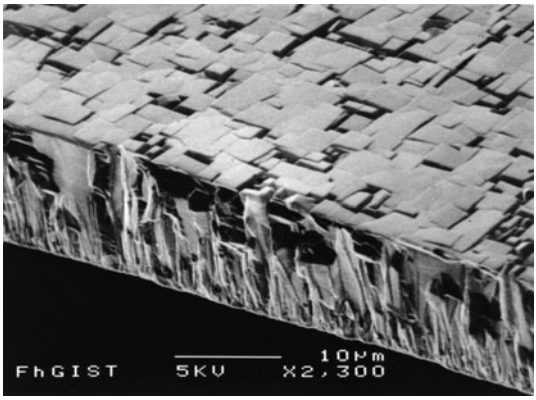


Fig. 2
Secondary-electron-based SEM image of a diamond film grown by means of a CVD-based process

thick diamond films as heat spreaders to increase the packaging density (Eden 1993). Diamond has the potential for both passive and active optical applications. Recently, it has been proposed that CVD polycrystalline diamond film can be used as a very fast optical switch (60 ps), due to its low dielectric constant and high breakdown voltage (Baba et al. 1995; Yoneda et al. 1995). Because of its high carrier mobility, breakdown field, saturation velocity, thermal conductivity, and wide band gap, diamond is considered an ideal material for electronic devices that function at high temperatures, voltages, power-levels, frequencies, and radiation environments.

3 The Low-Temperature Synthesis of Diamond Films

Diamond is the crystalline form of carbon and it occurs in nature. The availability of natural diamond is limited and the purity of such diamond is not controllable, requiring efforts to synthesize it in the laboratory.

The hot filament CVD (HF-CVD) method is the earliest method used for the growth of diamond under low pressures, and is also the most popular method (see *Fig. 3*). Matsumoto et al. (1982) exploited a refractory metal filament (such as W) and heated it to a temperature above 2,000 °C, at which atomic hydrogen could be easily produced as H_2 passed over the hot filament. The simultaneous production of atomic hydrogen during hydrocarbon pyrolysis could enhance the deposition of diamond. Diamond was deposited preferentially as graphite formation was suppressed. As a result, the deposition rate of diamond increased to about 1 mm/h, which proved valuable for industrial manufacturing (Davis 1992). The simplicity and comparatively low capital and operating cost of hot-filament-assisted CVD have made the method popular in industry where it is imperative to minimize the price of synthetic diamond. A wide variety of refractory materials have been used as filaments including W, Ta, and Re. Carbide-forming refractory metals must be first carburized before starting the deposition of diamond films. HF-CVD possesses the ability to adjust to a wide variety of carbon sources such as methane, propane, ethane, and other hydrocarbons. Even oxygen-containing hydrocarbons including acetone, ethanol, and

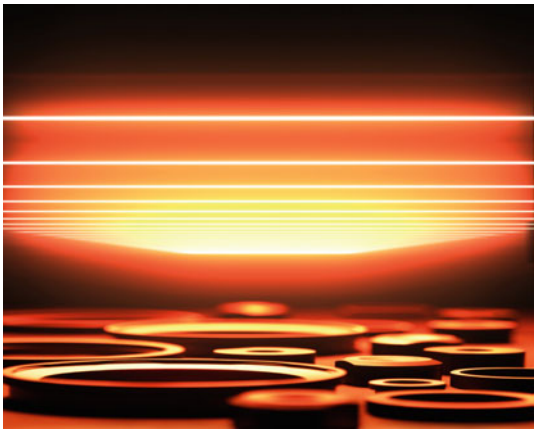


Fig. 3
Impression of a HF-CVD process (© Fraunhofer IST)

methanol can be applied. The addition of oxygen-containing species may widen the temperature range within which diamond deposition can take place. In addition to the typical design of HF-CVD, some modifications have been developed.

Besides HF-CVD, microwave plasma-assisted CVD (MW-CVD) is most frequently used. The excitation frequency for microwave plasma CVD is typically 2.45 GHz. Microwave plasma is unique in that microwave frequency can oscillate electrons. High ionization fractions are generated as electrons collide with gas atoms and molecules. Microwave plasma is often said to have "hot" electrons and "cool" ions and neutrals. A typical microwave reactor is referred to in Davis (1992). Microwaves enter into the reaction chamber from a proprietary antenna that converts a rectangular WR284 microwave signal into a circular mode. The microwave proceeds through a silica window into the plasma-enhanced CVD process chamber. The size of the luminous plasma ball will increase with increasing microwave power. Diamond films have been grown with the edge of the luminous plasma located about 2 cm higher than the substrate. The substrate does not have to be in immediate contact with the luminous glow for diamond to grow via microwave plasma. Uniform diamond films with diameters of up to 4 in. can be deposited using this system.

To control the film growth processes many experimental conditions, most importantly the process temperature, concentration, and pressure of reactive gases must be detected.

Even though one might argue that the research area of gas-phase diagnostics falls in the materials engineering regime as it deals with tools to mainly study the synthesis process, we feel that the techniques developed here have such a strong impact on materials science in general that an inclusion is justified.

In order to identify and understand the gas-phase and gas–solid reaction mechanisms in the CVD process, quantitative measurement of the concentrations of both free radical and stable species in the gas phase is required. Various in situ diagnostic techniques are available for such studies. Optical spectroscopy is a widely used technique, but is typically only specific with respect to a particular target species. Gas chromatography and mass spectrometry feature the distinct advantages of being generic and being able to analyze several stable species simultaneously. Especially the so-called molecular beam mass spectrometry (MBMS)

(Hsu et al. 1992) that features minimal disturbance of the process environment is a very potential tool in the context of analyzing diamond synthesis processes and the one we want to focus on in this chapter.

In general, there are two basic forms of electron multipliers that are commonly used in mass spectrometry as well as MBMS: the discrete-dynode electron multiplier and the continuous-dynode electron multiplier (often referred to as a channel electron multiplier or CEM) (Heroux and Hinteregger 1960). The working principle of both is analogical. The impact of a particle on a dynode results in the emission of secondary electrons. Naturally the dynode material is chosen in such a way that it features a high secondary electron emission coefficient such as BeO or Mg-O-Cs, respectively. The emission of three to five secondary electrons for an impacting electron with an energy of about 100–200 eV is possible. Therefore, a series of dynodes in a discrete-dynode electron multiplier with appropriate potential differences will lead to significant signal amplification. A similar result can be achieved utilizing a continuous dynode. Hsu and Tung (1992) applied such a continuous-dynode electron multiplier to an MBMS in the context of diamond synthesis and were able to detect radicals with sensitivity better than 10 ppm.

As final part of the gas-phase diagnostic section we want to introduce a rather unique technique, which is tailored to detect hydrogen in the gas phase. The importance of the presence of hydrogen for selecting diamond formation is widely accepted. Several possible roles within the mechanism of diamond formation have been attributed to hydrogen. Hydrogen molecules are supposed to prevent the formation of polycyclic aromatic hydrocarbons in the gas phase, which otherwise lead to the deposition of nondiamond carbon phases. On the other hand, there are several ways for atomic hydrogen, created by the activation of the gas atmosphere, to influence the deposition process. H atoms may react with hydrocarbons, which together with molecular hydrogen usually constitute the source gas, to form hydrocarbon species like methyl, which is key species in the process of diamond formation. By its ability to etch nondiamond carbon phases at the growth front much faster than diamond and to create new growth sites, atomic hydrogen may lead to the formation of thermodynamically metastable diamond as the only phase. Furthermore, H atoms bonded to carbon on the surface may stabilize the sp^3 bonds necessary for diamond formation. The dependence of the atomic hydrogen concentration (c_H) on methane addition and filament temperature in a filament-assisted diamond deposition system has been investigated using resonance-enhanced multiphoton ionization (REMPI) (Celii and Butler 1989). Although this technique offers a good spatial resolution, a quantitative determination of H concentrations was not possible because of the lack of a calibration for the REMPI signals. One can determine absolute H concentration profiles at hot filaments under typical diamond CVD conditions by a two-photon laser-induced fluorescence (LIF) method used earlier for the detection of H atoms in low-pressure flames (Meier et al. 1986) (see also ❯ *Fig. 4*). The applicability of LIF for different pressures, methane fractions, and varying filament temperature, diameter, and material used in filament-assisted CVD of diamond is demonstrated by Meier et al. (1990) (❯ *Fig. 4*). A detailed description of the experimental setup and the calibration of the LIF signals is provided in Meier et al. (1990). The signal itself is detected by a photomultiplier, which will be introduced in more detail below. Radial temperature profiles, necessary for the determination of calibration factors accounting for quenching and Doppler broadening of the LIF signal, were estimated using a model introduced by Langmuir and coworkers for the description of heat losses from hot filaments to the surrounding gas phase (Blodgett et al. 1932). The temperature estimation includes a temperature drop at the filament over the first mean free path of hydrogen molecules. This temperature drop depends on gas pressure and filament

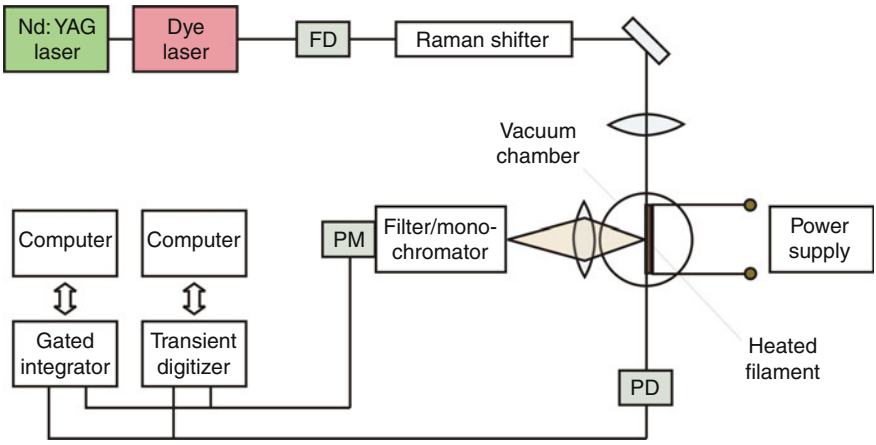


Fig. 4
Schematic of the operation principle of laser-induced fluorescence

radius. In conclusion, the quantitative determination of c_H using two-photon laser-induced fluorescence yields kinetic limitations for hydrogen dissociation under conditions of HF-CVD of diamond. The deviation from thermal equilibrium is confirmed by subequilibrium values of c_H near the filament and the dependence of c_H on filament geometry and material. One has to note that the presence of a substrate will change the concentration profiles. The presented results are, however, in good agreement with values published by Harris and Weiner (1988). The effect of methane addition on the measured H profiles can be attributed to the consumption of atomic hydrogen by the hydrocarbon species in the gas phase, possibly in addition to surface effects due to the reaction of methane with the filament.

4 Structure Characterization: SEM, TEM, XRD, and Raman

In the following we will present the most common tools for materials analysis in the context of diamond film characterization such as scanning electron microscopy (SEM), transmission electron microscopy (TEM), x-ray diffraction (XRD), and Raman spectroscopy, respectively. Special emphasis is given on the corresponding detector technology used.

The surface morphology of the deposited diamond films was analyzed by scanning electron microscopy (SEM). SEM is probably one of the most common tools in materials science. A beam of high-energy electrons is scanned in a raster pattern over the sample of interest. The various interaction products of this sample with the primary electron beam are in turn used to image and/or characterize the sample surface. Among those interaction products that are widely used are secondary electrons, backscattered electrons, and x-rays.

Secondary electrons are typically detected by a so-called Everhart–Thornley (ET) setup (Everhart et al. 1960). This kind of detector usually combines a Faraday cage, a scintillator element, a light guide, and a photomultiplier. Due to an appropriate charge on the Faraday cage covering the scintillator element it is possible to collect the relatively slow (below 50 eV of kinetic energy) secondary electrons. Light is emitted once the scintillator element is struck by

the electrons. This, in turn, is channeled via a light guide to the photomultiplier for signal gain (up to ~10^6 times). The photomultiplier itself consists of a photocathode followed by an electron multiplier (described above). The photocathode, which is activated by the impinging photons, emits, based on the photoelectric effect, electrons, which are multiplied. The advantage of the ET setup is that only the Faraday cage and the scintillator element are needed in the vicinity of the sample, whereas, the bulky photomultiplier can conveniently be placed outside of the sample chamber. However, the increasing demand for high resolution in SEM often results in very small working distances. Here, the angular spectrum from which the ET setup would be able to collect its signal is significantly blocked by the pole shoe of the final electron lens system. Alternative detection strategies for secondary electrons are necessary. The current path to solve this issue is an in-lens solution. In this context, solid state annular detectors are utilized. This device is essentially a disk-shaped silicon-based diode positioned inside the column of the SEM. For an overview on semiconductor-based detectors see Spieler (2005).

In order to detect backscattered electrons three standard approaches can be found: An ET setup with small shielding, a solid state (quadrant) annular detector, or a scintillator type of backscattered electron detector. In the ET-based strategy, the Faraday cage is simply charged in such a way as to deflect all incoming secondary electrons only allowing the more energetic backscattered electrons to pass. This solution, however, turns out to be relatively inefficient as typically only a very small dihedral angle is sampled. The annular detector is in case of backscattered electron detection usually positioned above the sample in a doughnut-type arrangement. Commonly quadrant annular detectors are used in this context as they offer aside from the elementary contrast imaging the option of topographical images if the signals are processed in an asymmetric way. In-lens systems similar to those used in secondary electron detection exist also for backscattered electron imaging. The design of a scintillator type of detector can be recognized as very similar to the ET detector, except that it lacks the collection and accelerating fields required to detect secondary electrons. Some scintillator types of backscattered electron detectors have been constructed using single-crystal scintillators. The most common scintillator type is, however, the Robinson detector (Robinson 1973). This design uses a plastic scintillator material to both produce the light and channel it to the external photomultiplier. The advantage of this design is that the collection surface can be made quite large, fairly economically.

The characteristic x-ray radiation emitted by the sample surface during SEM operation can be exploited to deduce the local elemental composition. The radiation can be tracked either in an energy-dispersive (energy-dispersive x-ray spectroscopy (EDX)) or in a wavelength-dispersive (wavelength-dispersive x-ray spectroscopy (WDX)) manner. In principle, two main solid state layouts are used to detect the photons in this context: Si(Li) as well as silicon drift detectors (SDD) (Lechner et al. 1996), respectively. A Si(Li) detector is a cylindrical Si crystal (3–5 mm thick) that has been drifted with Lithium. The x-ray photons are absorbed in the central volume of the crystal and result in the generation of electron–hole pairs and an according current. To reduce the noise level the Si(Li) detectors need significant cooling that is typically realized by liquid nitrogen. For this reason appropriate windows have to be used to separate the cryostat with detector unit from the SEM chamber. Earlier such windows have been produced from beryllium. Current solutions utilize polymer foils that feature a thickness of about 300 nm (so-called super ultra thin windows (SUTW)). SDDs, on the other hand, are based on Si wafers and are therefore much thinner compared to Si(Li) detectors. The main characteristic of a SDD is its transversal field generated by a series of ring electrodes. This setup, which is imported from particle physics, forces the generated charges to drift to a small central collection electrode and allows for significantly higher count rates. Additionally, SDDs, based on their smaller volume,

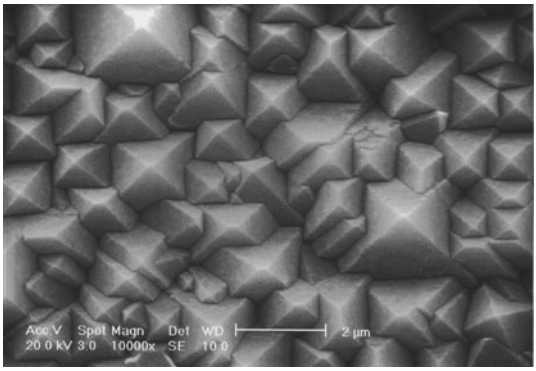


Fig. 5
Scanning electron image obtained from a heteroepitaxial diamond film with a thickness of about 15 µm

naturally feature smaller stray currents and, in turn, smaller noise. Therefore, it is sufficient to cool these detectors only by Peltier elements. Their size along with faster data acquisition, a cost-efficient production, and inexpensive cooling are strong arguments for this technology that starts to replace Si(Li) on many occasions.

Figure 5 shows the ET-based secondary electron SEM surface image of a 15 µm thick film. Crystallites with lateral sizes of about $2 \times 2\,\mu m^2$ with well-developed (111) surfaces are demonstrated. The substrate orientation is indicated in the figure. Nearly all crystal edges are aligned parallel to each other. Some of the crystals have grown together without perceptible grain boundaries. Most of the crystallites show (001) surfaces parallel to the substrate, and a few of them are slightly tilted toward the substrate surface.

Another electron-based microscopy technology heavily used in materials science is transmission electron microscopy (TEM). Owing to its superior performance of high-resolution imaging as well as nanoprobe analysis TEM has attracted much attention. Imaging strategies in TEM are entirely different compared to SEM. In TEM photo-films, television cameras, imaging plates as well as slow-scan charge-coupled device cameras are used to record the information. In this article, we will focus on the latter two as these have a relatively close relationship to particle detectors.

The imaging plate was originally developed as recording device in the context of x-ray imaging. The working principle of plates designed for x-rays and those for electrons in a TEM are very similar. It is based on the trapping of electrons and holes, which are generated by impinging electrons, at defect positions and the Eu^{2+} doped in the halide phosphor [$BaF(Br, I):Eu^{2+}$] powder, respectively. Later this information is read out by a laser that releases the electrons which upon recombination with the holes generates light, which in turn is amplified by a photomultiplier. The operation principle of a slow-scan charge-coupled device (CCD) camera is based on the conversion of incident electrons into light by a YAG (yttrium–aluminum garnet: $3Y_2O_3 \cdot 5Al_2O_3$) scintillator. The light is transferred through a fiber optics plate to the CCD where it is converted at a photoactive region (epitaxial layer of silicon) to an electric charge that is temporarily stored in each channel. The charge is then sequentially transferred to the neighboring pixel and finally read out.

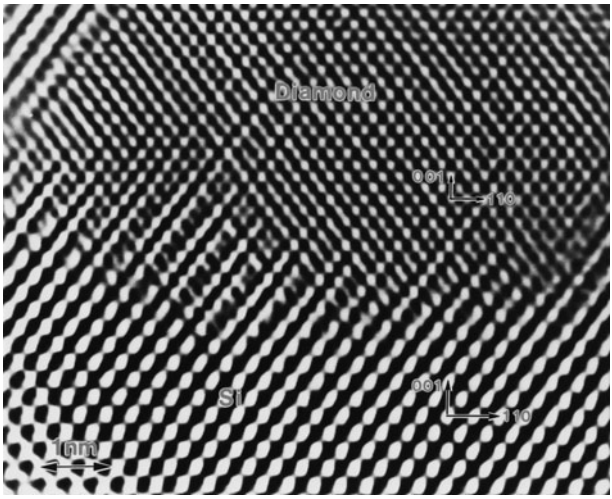


Fig. 6
High-resolution TEM image of a diamond/Si interface

The HREM observation shown in *Fig. 6* was carried out at 400 keV with a point-to-point resolution of 1.7 Å. The HREM images of interface areas were typically taken along the Si [110] direction, showing the interface parallel to the silicon (001) plane. For nearly every diamond crystal observed, planar defects (which are predominantly microtwins and stacking faults on {111} planes) were found extending from the diamond–silicon interface. The figure shown here is an image of an epitaxial diamond–silicon interface, obtained in the usual [110] projection. Orientation of the lattice fringes exhibits the [001] parallel epitaxial relationship between diamond and silicon. Because the ratio of the lattice constants of diamond and silicon is close to 1.5, a nearly perfect 3-to-2 correspondence (1.5% mismatch) of lattice spacings is seen at the interface. Every three Si {111} fringes are matched with four diamond {111} fringes (marked by arrows). Individual 60° interface misfit dislocations for every third (111) atomic plane, with a spacing of two such dislocations of about 7 Å, can be clearly identified. The presented results are reproducible for all of the epitaxially oriented diamond grains observed, and they give strong evidence that diamond crystals can be epitaxially grown directly on silicon. Therefore, a thin epitaxial SiC intermediate layer is found to be unnecessary for the epitaxy.

Besides the electron-based techniques, photons are heavily used to get information on the structure of films. In this context x-ray diffraction (XRD) is most likely the most common technique to analyze the atomic structure of a (crystalline) material. Aside from the already mentioned Si(Li) detectors, SDDs, and photo plates, the diffracted beam of x-ray photons in most of the commercial diffractometers is analyzed by Geiger counters. These detectors are based on the ionization of a gas. An x-ray photon that enters a cylinder (representing the cathode) containing a gas (usually helium, argon, or neon with halogens added) ionizes it forming ions and electrons. The latter are accelerated toward the anode (a wire in the center of the cylinder) causing further ionization along their trajectories. This phenomenon known under the term of Townsend avalanche is electrically detected and counted.

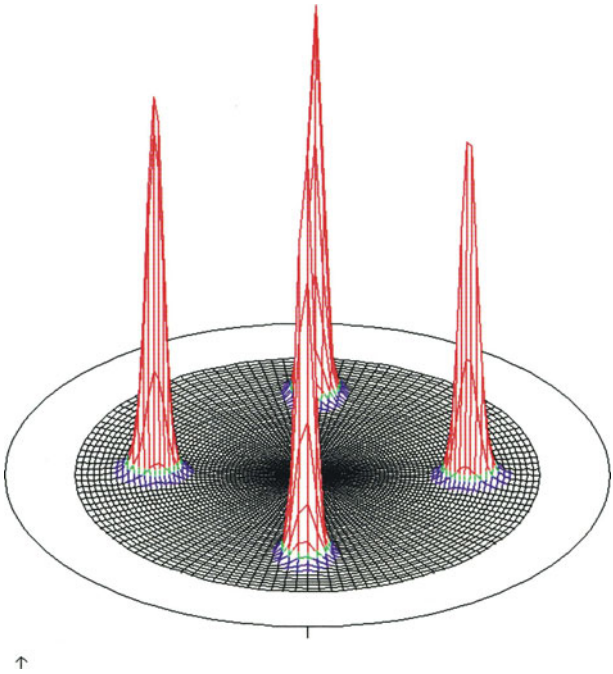


Fig. 7
{111} pole figure of an epitaxial diamond film

Figure 7 shows a {111} diamond pole figure, generated by a standard x-ray diffraction measurement under reflection geometry, of an epitaxial diamond film. The four {111} peaks clearly dominate the diagram.

Another photon-based tool that allows getting an insight into the bonding structure of the diamond films is Raman spectroscopy. As carbon bonds are typically very Raman active, Raman spectroscopy is an ideal tool to study diamond. The basic idea is to excite bond or molecular vibrations via electromagnetic radiation. When irradiating material with electromagnetic radiation of a single frequency, the light will be scattered elastically as well as inelastically. In Raman spectroscopy, one is interested in the inelastically scattered portion as it contains information about the bond structure of the material under study. In dispersive Raman spectroscopy, the scattered light is usually analyzed with a CCD detector (see above) after being wavelength separated by a diffraction grating.

Micro-Raman spectroscopy was applied to obtain information on the phase purity and to determine the in-plane crystallographic orientation of the diamond crystallites relative to the Si substrate. The laser beam was focused on the samples with a spot size of approximately 10 μm in diameter. By varying the pole angle ($0^\circ < \chi < 80^\circ$) and the azimuthal angle φ ($0^\circ < \varphi < 360^\circ$), the sample can be rotated relative to the scattering wave vector. The Raman intensities (peak areas) of both the silicon and the diamond line were measured with parallel polarization of incident and scattered light while rotating the sample under the microscope. The results are shown in *Fig. 8*. The Si intensity shows a behavior expected from the selection rules of Raman backscattering from a (001) surface of a system with diamond crystal structure (Loudon 1964).

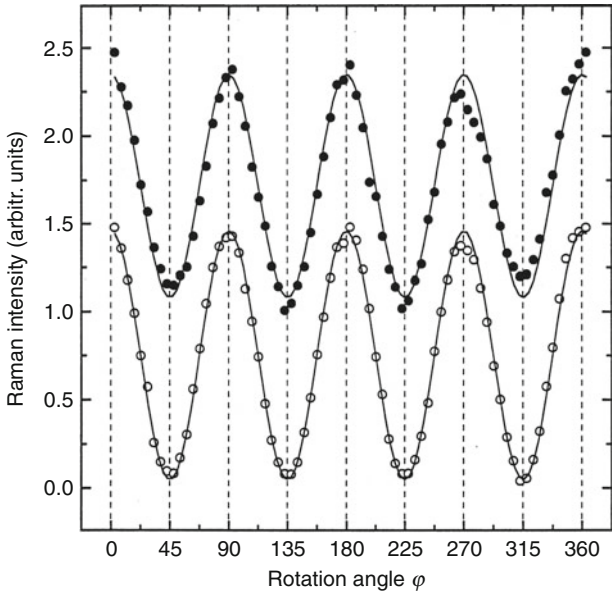


Fig. 8

Linear plot of the Raman intensities of diamond (*closed dots*) and the underlying silicon (*open dots*) as a function of the azimuth angle φ. $\varphi = 0$ denotes scattering geometry 001(110,110)001. The solid curves give the calculated $I \approx I_0 \cdot \cos^2 2\varphi$

5 Properties/Applications: Electron Field Emission

Diamond can possess a negative electron affinity (NEA) surface that allows its surface to emit electrons under low electric field. Comparing with the complicated and costly process of fabricating metal or semiconductor sharp microtips to provide geometric field enhancement, diamond films can show a planar electron emission property at low electric field and be manufactured cheaply (Geis et al. 1996; Zhirnov et al. 1997; Zhirnov and Hern 1998). Here we describe that strong electron emission at low applied fields is obtained from nanocrystalline diamond films.

Nanocrystalline diamond films were prepared by MWPACVD method assisted by a continuous ion bombardment. During this process, ion bombardment of different energies was induced by applying a negative bias voltage from 0 V to −140 V on the substrate relative to the grounded vacuum chamber. Scanning electron microscopy and Raman spectroscopy indicated the nanometer structure of the films. The grain size of nanocrystalline diamond films can be modified by means of changing the energy of the bombarding ions. The results suggested that low-field electron emission and high emission current can be obtained from the films consisting of nano-sized diamond grains.

Field emission experiments were performed at a pressure of 10^{-8} Torr. Sputtered indium–tin-oxide (ITO) glass was used as the anode. The anode–cathode spacing was approximately 120 μm.

Figure 9 shows the typical current–voltage (I–V) curve from the emissions of diamond films deposited at various biases. To the film deposited at −140 V (grain size: 40 nm), the

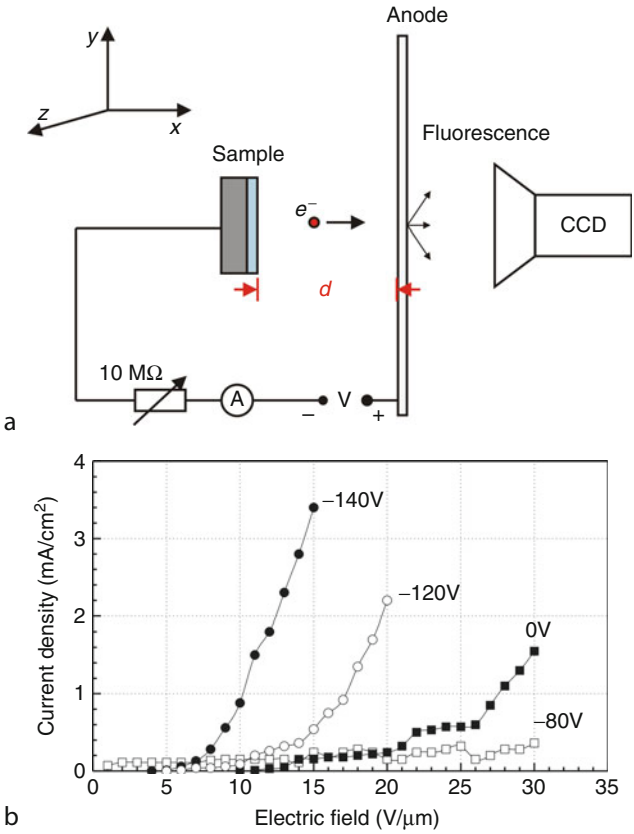


Fig. 9

(a) Schematics of an experimental setup to acquire field-emission properties of a film material (anode: ITO- and ZnS-coated quartz glass; base and working pressure of 4×10^{-7} Pa and 5×10^{-6} Pa, respectively). (b) The corresponding I–V curves of diamond films deposited at different bias voltages

emission current density increased rapidly at an applied voltage of about 4 V/μm and reached 3.4 mA/cm^2 at 15 V/μm. To the samples deposited at 0 V (grain size: 0.3 μm), the emission characteristics shifted to the high-voltage region, the current increased clearly at 15 V/μm, and reached 2.2 mA/cm^2 at 20 V/μm. To the (001)-textured film deposited at −80 V, there is no clear increase in emission current when increasing the applied voltage to 30 V/μm. The results show that nanocrystalline diamond films are advantageous for obtaining high emission current and low emission threshold.

The nanocrystalline diamond films with high grain boundary density can provide more efficient emission sites and electron sources, which present the properties of high emission current and low-field electron emission.

The results indicate that the nanocrystalline film deposition at −140 V bias voltage has characteristics of low-field electron emission and high emission current, which can be attributed to the high grain boundary density of nanocrystalline diamond films.

6 Material Development for Particle Detection

As already mentioned, due to its outstanding properties, diamond can be used for developing many electronic devices for application in very stringent conditions. Actually, one of such applications even falls into the field of radiation and high-energy particle detection. As an example, radiation-hard diamond-based neutron detectors are being tested in fusion experiments where conventional solid state detectors cannot be employed due to extremely high radiation fluxes causing irreversible detector damage. Based on its high reliability, the most widely used technique to produce such devices probably is chemical vapor deposition (CVD). Great effort is now being devoted in many laboratories to the growth of both synthetic single-crystal diamond and homoepitaxial diamond films showing very promising performance of single-crystal diamond-based devices as α-particle and neutron detectors. Resolutions as low as 0.4% and 2.9% were reported as the best results obtained for α-particle and for neutron detection, respectively (Balducci et al. 2005).

7 Conclusions

In this chapter, we tried to show the striking affinity of particle detector science to materials science and engineering. Illustrated by the example of a chemical-vapor-deposition-based diamond deposition process, some of the accompanying particle detector technology was presented for all of the four aspects of materials science and engineering such as process, structure, properties, and performance. In retrospective, most of the crucial developments in materials science have been strongly correlated with appropriate advances in the corresponding particle detector technology being utilized. In the last section, we also hinted at the fact, that, of course, progress in materials science and engineering can also lead to improved designs and performance of particle detectors. This is not surprising as both research fields are based on the understanding of interaction mechanisms between particles and matter. Advances in one area typically will lead to progress in the other and vice versa.

8 Cross-References

- Chapter 1, "Interactions of Particles and Radiation with Matter"
- Chapter 2, "Electronics Part I"
- Chapter 11, "Gaseous Detectors"
- Chapter 13, "Photon Detectors"
- Chapter 15, "Scintillation Counters"
- Chapter 21, "New Solid State Detectors"
- Chapter 31, "Neutron Detection"

References

Allison WWM, Wright PRS (1991) The physics of charged particle identification. In: Ferbel T (ed) Experimental techniques in nuclear and particle physics. World Scientific, Singapore

(reprinted from: Bock RK (ed) (1984) Formulae and methods in experimental data evaluation. European Physical Society, Geneva)

Angus JC, Hayman CC (1988) Low-pressure, metastable growth of diamond and "diamondlike" phases. Science 214:913

Baba K, Aikawa Y, Shohata N, Yoneda H, Ueda K-I (1995) Photoconductive switch with CVD diamond films by ultraviolet light pulse. NEC Res Dev 36(3):369

Balducci A, Marinelli M, Milani E, Morgada ME, Pucella G, Tucciarone A, Verona-Rinati G, Angelone M, Pillon M (2005) Synthesis and characterization of a single-crystal chemical-vapor-deposition diamond particle detector. Appl Phys Lett 86:213507

Blodgett KB, Langmuir I (1932) Accommodation coefficient of hydrogen: a sensitive detector of surface films. Phys Rev 40:78, and references therein

Celii FG, Butler JE (1989) Hydrogen atom detection in the filament-assisted diamond deposition environment. Appl Phys Lett 54:1031

Cherenkov PA (1934) Visible emission of clean liquids by action of γ radiation. Dokl Akad Nauk SSSR 2:451. Reprinted in Selected Papers of Soviet Physicists (1967) Usp Fiz Nauk 93:385. V sbornike: Pavel Alekseyevich Čerenkov: Chelovek i Otkrytie pod redaktsiej A. N. Gorbunova i E. P. Čerenkovoj, M., Nauka, 1999, s. 149–153

Compton AH (1922) Secondary Radiations produced by X-rays and some of their applications to physical problems. In: Bulletin of the National Research Council 20:10; Nachdruck in: Compton AH, Shankland RS (1973) Scientific papers of Arthur Holly Compton. University of Chicago Press, Chicago

Davis RF (ed) (1992) Diamond films and coatings. Noyes, New Jersey

Eden RC (1993) Application of diamond substrates for advanced high density packaging. Diam Relat Mater 2(5–7):1051

Everhart TE, Thornley RFM (1960) Wide-band detector for micro-microampere low-energy electron currents. J Sci Instrum 37(7):246–248

Field JE (1979) The properties of diamond. Academic Press, Oxford

Geis MW, Twichell JC, Efremow NN, Krohn K, Lyszczarz TM (1996) Comparison of electric field emission from nitrogen-doped, type lb diamond, and boron-doped diamond. Appl Phys Lett 68:2294

Harris SJ, Weiner AM (1988) Measurement of stable species present during filament-assisted diamond growth. Appl Phys Lett 53:1605

Heroux L, Hinteregger HE (1960) Resistance strip magnetic photomultiplier for the extreme ultraviolet. Rev Sci Instrum 31:280

Hsu WL, Tung DM (Sept 1992) Application of molecular-beam mass-spetrometry to chemical vapor-deposition studies. Rev Sci Instrum 63(9):4138

Lechner P et al (1996) Silicon drift detectors for high resolution room temperature X-ray spectroscopy. Nucl Instrum Methods A377:346–351

Loudon R (1964) Raman effect in crystals. Adv Phys 13:423

Matsumoto S, Sato Y, Tsutsimi M, Setaka N (1982) Growth of diamond particles from methane hydrogen gas. J Mat Sci 17:3106

Meier U, Kohse-Hoinghaus K, Just Th (1986) H and O atom detection for combustion applications: study of quenching and laser photolysis effects. Chem Phys Lett 126:567

Meier U, Kohse-Hoinghaus K, Schafer L, Klages C-P (1990) Two-photon excited LIF determination of H-atom concentrations near a heated filament in a low pressure H_2 environment. Appl Opt 29:4993

Pan LS, Kania DR (eds) (1995) Diamond: electronic properties and applications. Kluwer, Boston

Robinson VNE (1973) A reappraisal of the complete electron emission spectrum in scanning electron microscopy. J Phys D Appl Phys 6:L105–L106

Spieler H (2005) Semiconductor detector systems. Oxford Science, Oxford

Wilks J, Wilks E (eds) (1994) Properties and application of diamond. Butterworth Heinemann, Oxford

Yarbrough W, Messier R (1990) Current issues and problems in the chemical vapor deposition of diamond. Science 247:688

Yoneda H, Ueda K-I, Aikwaa Y, Baba K, Shohata N (1995) Appl Phys Lett 66(4):460

Zhirnov VV, Hern JJ (1998) Diamond films: recent developments – electron emission from diamond films. MRS Bull 9:42

Zhirnov VV, Wojak GJ, Choi WB, Cuomo JJ, Hern JJ (1997) Wide band gap materials for field emission devices. J Vac Sci Technol A 15:1733

30 Spallation – Neutrons Beyond Nuclear Fission

Harald Conrad
Forschungszentrum Jülich GmbH, Jülich, Germany

1 Introduction: Fission Versus Spallation 721
1.1 The Fission Reactor 721
1.2 Source Strength *S* 722
1.3 Early Reactor Development 722
1.4 Technical Limitations 723

2 Neutron Sources: Why Not Pulsed? 723
2.1 The Two Essentials in Neutron Scattering: Single-Crystal and Time-of-Flight Techniques 724
2.2 Pulsed Reactors or What Else? 724

3 Spallation: The Future for Rapidly Pulsed Neutron Sources 726
3.1 The Spallation Reaction 726
3.1.1 Total Neutron Yield 727
3.1.2 Spectral Distribution 728
3.1.3 Source Distribution, Proton Mean Free Path, and Range 728
3.1.4 Heat Deposition 730
3.2 Technical Details of a Pulsed Spallation Source 730
3.2.1 The Accelerator 732
3.2.2 The Target: Solid or Liquid? 733
3.2.3 The Moderators 735
3.3 Examples of Spallation Sources 738
3.3.1 The US Spallation Neutron Source SNS 738
3.3.2 The European Spallation Source (ESS) 741

4 Experimental Methods at Spallation Neutron Sources 741
4.1 *Epithermal* Neutrons: An Important Reason for Ultra-Short Proton Pulses 743
4.2 Spectroscopy at High Energy Transfers 743
4.3 Powder Diffractometry at Pulsed Sources 744
4.4 Neutron Powder Diffractometry in the History of Arts 745
4.5 Neutron Radiography 746

5 Spallation: Accelerator-Driven Nuclear Energy 748

C. Grupen, I. Buvat (eds.), *Handbook of Particle Detection and Imaging*,
DOI 10.1007/978-3-642-13271-1_30, © Springer-Verlag Berlin Heidelberg 2012

5.1 ADS Research and Development: The Belgian MYRRHA Project 750
5.1.1 The Accelerator 751
5.1.2 The Target 752
5.1.3 The Subcritical Core 752
5.2 Energy Amplifier 753
5.3 Nuclear Waste Incineration 754

6 Conclusions ***755***

7 Cross-References ***756***

References ***756***

Further Reading ***757***

Abstract: The classical research neutron sources are fission reactors. They have reached their technical limits as far as neutron flux is concerned. But there is an alternative way with many advantages: spallation. The emphasis in this context is on pulsed operation, which is easily achieved with spallation as being accelerator-driven. The extension of neutron scattering to fields not covered with reactors is discussed as well as the utilization of spallation neutrons for other fields such as nuclear waste transmutation and future power reactors.

1 Introduction: Fission Versus Spallation

Induced by high-energetic particles, spallation denotes a nuclear reaction, which results in a more or less complete disintegration of the involved nuclei. The reaction has long been known by cosmic-ray physicists, who observed it in nuclear emulsions exposed in high altitudes. In 1947, E.O. Lawrence observed the reaction in the laboratory by bombarding uranium targets with neutrons of 90 MeV (quoted in Crandall and Millburn (1958)). The most practical means of triggering the spallation reaction is by using high-energy protons from accelerators. The salient feature of the reaction is the release of – among other reaction products – a large number of neutrons, which makes it more than competitive with fission reactors.

In both fission and spallation the primary *fast* neutrons have energies of at least a few MeV, which are useless for neutron scattering. Neutrons useful for the latter are called *slow*, in particular *thermal*. In reactor physics, the term *thermal* is used to distinguish these neutrons, which sustain the nuclear chain reaction, from the *fast* fission neutrons. Thermal neutrons are in thermal equilibrium with an adequate slowing-down medium (moderator) like graphite, H_2O, or D_2O at ambient temperature ($E = k_B T \cong 0.025\,\mathrm{eV}$). In neutron scattering studies of condensed matter and biologically relevant substances, *thermal* neutrons are a nearly unique tool, because their energies are comparable to excitations in these materials, and their wavelengths are comparable to interatomic distances according to the de Broglie formula $\lambda = h/p = h/\sqrt{2mE}$, which relates a particle property (momentum p) to a wave property (wavelength λ) via Planck's constant h. For practical work it is convenient to use the numerical expression $\lambda[\mathrm{nm}] = 0.0286/\sqrt{E[\mathrm{eV}]}$, where the neutron energy has to be inserted in units of electron volts giving the neutron wavelength in units of nanometers. For example, for $E \cong 0.025\,\mathrm{eV}$, we obtain a wavelength of 0.18 nm, which is of the order of interatomic distances. Needless to say that these wavelengths render *thermal* neutrons at least as useful for structure determination as X-rays, because neutrons easily penetrate high-Z materials or large bulk samples. Moreover, due to the perfect match of the energy of thermal neutrons to the dynamics of condensed matter, energy transfers during a scattering event can easily be resolved.

1.1 The Fission Reactor

In a fission reactor a *thermal* neutron is captured by a uranium (^{235}U) nucleus, which entails the splitting into two in general unequal nuclei and the release of 2–3 (fast) MeV neutrons. The exact number depends on the particular fission fragments of a reaction. On average about 2.5 neutrons are emitted per fission. These fast neutrons now have in turn to

be slowed down by an appropriate moderator medium. If H_2O or D_2O is used as a moderator, the medium serves as a reactor coolant at the same time as well. Now, obviously, there are more neutrons produced than needed to sustain the controlled chain reaction. The surplus of about 1.5 neutrons is exactly what makes the fission reactor a powerful source of useful neutrons.

1.2 Source Strength *S*

The source strength S is an important quantity in the present context and can be estimated by the following consideration. An energy of about 200 MeV per fission is released by Coulomb repulsion of the fission fragments. At least one neutron out of the above mentioned 2.5 neutrons is not needed to sustain the chain reaction. Then the number of available neutrons per megawatt of reactor power can be calculated according to the expression

$$S\,(\mathrm{n/s/MW}) = \text{Reactor power/energy per fission} = \text{Fissions/second} \\ \triangleq \text{Available neutrons/s/MW}\,.$$

Using the unit conversion relation $1\,\mathrm{eV} = 1.6 \times 10^{-19}$ J we obtain for a reactor power of 1 MW

$$S\,(\mathrm{n/s/MW}) = (1 \times 10^6\ \mathrm{J/s})/(3.2 \times 10^{-11}\ \mathrm{J/fission}) \cong 3.1 \times 10^{16}\ \mathrm{fissions/s} \\ \triangleq 3.1 \times 10^{16}\ \mathrm{n/s/MW}\,. \qquad (1)$$

This result will be used in the discussion on technical limits of fission reactors and the prospects of spallation neutron sources based on their respective achievable source strengths. Enhancing the source strength had been and is the major aim in reactor as well as in advanced neutron source development (Sects. 2 and 3).

1.3 Early Reactor Development

For the purpose of elaborating on the possibilities and limits of reactor neutron source developments, it is interesting to take a look at the history of fission reactors and their use as a research tool. Only 10 years after Chadwick's identification of the neutron in 1932, the first controlled nuclear chain reaction had been put in operation by Fermi in Chicago. Immediately afterward a remarkably rapid development of research reactor performance took place. The central aim of this achievement was to increase the thermal neutron flux, mathematically defined as $\Phi(\mathbf{r}) = \int\int n(\mathbf{r}, \mathbf{\Omega}, E)\, v(E)\, \mathrm{d}\Omega\, \mathrm{d}E = n(\mathbf{r}) v_{\mathrm{th}}$, where $n(\mathbf{r})$ is the number density of neutrons at position $\mathbf{r}$ inside the moderator and $v_{\mathrm{th}} \approx 2{,}200$ m/s the (average) thermal neutron velocity.

In less than 30 years the flux had been increased by three orders of magnitude. At the top of the list rank reactors like the High Flux Beam Reactor (HFBR) in Brookhaven (New York State, 60 MW), the High Flux Isotope Reactor (HFIR) in Oak Ridge (Tennessee, 100 MW), and the High Flux Reactor of the Institut Laue–Langevin (ILL) in Grenoble, 57 MW, all with thermal neutron fluxes Φ of the order of 10^{15} neutrons/(cm^2 s) (*Table 1*). Interestingly, this development reached an end with the ILL reactor. A very ambitious project to increase the flux further by a factor of 5–6 had been started at Oak Ridge. This Advanced Neutron Source (ANS) called

Table 1
Examples of the development of research reactors

	First criticality	Average thermal power (MW)	Average power density (MW/l)	Thermal flux Φ (n cm^{-2} s^{-1})
Graphite reactor (Oak Ridge)	1942	3	0.00002	1×10^{12}
FRM[a] (Garching, "Atom-Ei")	1957	1	0.04	1×10^{13}
FRJ-2[a] (Jülich) (decommissioned in 2006)	1962	10	0.1	1×10^{14}
	1970	20	0.2	3×10^{14}
HFR (ILL Grenoble)	1971	57	1	1.2×10^{15}
ANS (Oak Ridge) (project canceled in 1996)	–	350	4	0.7×10^{16}
FRM II (Garching)	2005	20	2	0.6×10^{15}

[a]FRM = Forschungsreaktor München, FRJ-2 = Forschungsreaktor Jülich 2

350 MW project had been canceled after 10 years of development in 1996 – mostly for budgetary reasons.

1.4 Technical Limitations

It should be pointed out that the enhancement of neutron fluxes was only possible by both increasing the ^{235}U enrichment (up to 93%) and reducing the reactor core size, in other words by increasing the number of fissions per unit volume. This is equivalent with a large increase in power density and thermo-hydraulic requirements. A limit in cooling capability seemed to have been reached with the average power density of 1.1 MW/l and 3 MW/l at the hot spot of the (single) fuel element reactor core of the ILL reactor.

It is now well established that power densities in reactor cores cannot substantially be increased without unwanted and impracticable consequences, such as liquid-sodium cooling. Moreover, the service time of reactor vessel components like beam-tube noses or cold sources (vessels next to the core containing cryogenic fluids like liquid hydrogen for producing *sub-thermal* neutrons) would become intolerably short due to radiation damage. Experience with the HFR in Grenoble shows that these service times are of the order of 7 years. Ten times higher fluxes would result in impracticable service times under one year.

2 Neutron Sources: Why Not Pulsed?

Regarding these arguments, we may ask, whether high-flux reactors have already reached a fundamental limit. This were certainly the case, if we expected a flux increase by another order of magnitude like the one observed in reactor development since the 1950s (*Table 1*). An enhancement by a factor of 6 over the ILL reactor with the ANS would have been only possible by a power increase to 350 MW with a simultaneous increase of the average power density by a factor of 4 compared to the HFR in Grenoble.

2.1 The Two Essentials in Neutron Scattering: Single-Crystal and Time-of-Flight Techniques

If we accept that steady state reactors have more or less reached their technical limits, how can progress for neutron scattering be obtained otherwise? The answer is obvious from a closer look at the methods of neutron scattering (see also ❯ Chap. 33, "The Use of Neutron Technology in Archaeological and Cultural Heritage Research"). Investigation of structure and dynamics of materials with neutrons require the monochromatization of the inherently polychromatic radiation. The two standard methods, i.e., monochromatizing neutrons by Bragg reflection from single-crystal and/or the time-of-flight technique, only use a minute fraction (10^{-2} to 10^{-4}) of the source flux.

Time-of-flight spectroscopy inefficiently utilizes the continuous reactor flux for two reasons, because it requires a pulsed and monochromatic beam. Typically, the neutron beam is chopped at a rate of the order of 100 Hz into pulses of 0.1 ms duration. In other words, 99% of the available neutrons are wasted. Spectrometers and diffractometers based on monochromatization by Bragg reflection use a comparatively narrow fraction of the polychromatic radiation, too. Time-of-flight techniques with pulsed source operation at the same average power can yield gain factors of the order of the ratio of peak-to-average flux (❯ Sect. 3.2.3). With crystal techniques, even higher-order Bragg reflections can be utilized, because they become distinguishable by their time of flight. In other words, the peak flux will be usable between pulses as well. The latter is particularly true for time-of-flight diffractometry (❯ Sect. 4.3). So, the task is to put these 99% of wasted neutrons into the opening time of the chopper, in other words, to run the reactor in a pulsed mode. Thereby for equal average reactor power the flux during the pulse could surpass the average flux by a factor of 100.

2.2 Pulsed Reactors or What Else?

A condition for a pulsed source to be superior to a continuous source is that the product of the pulsed source peak flux Φ_{pk} times its repetition rate f_{rep} is large compared to the product of the steady state source Φ_{st} times the typical slot opening rate f_{ch} of a chopper, i.e.,

$$\Phi_{pk} \cdot f_{rep} \gg \Phi_{st} \cdot f_{ch} \,. \tag{2}$$

Demanding, for example, a pulsed source peak flux of 100 times the flux of the ILL reactor and a repetition rate of 50, say, the condition ❯ 2 for a generic 100 Hz chopper instrument at the ILL, reads $10^{17} \times 50 \gg 10^{15} \times 100$. In other words, this hypothetical pulsed source outperforms the reactor by a factor of 50.

Now, which types of pulsed sources of sufficient source strengths are available? Various types of pulsed thermal reactors have been built in the past, the most famous thereof are the TRIGA reactors. They reach thermal peak fluxes of up to several 10^{17} cm^{-2} s^{-1}, but do not fulfill the requirements of modern research reactors due to their inherent extremely low repetition rate of only a few pulses per hour. But there is the famous high-flux pulsed reactor IBR-2 in Dubna (Russia), a successful but unique example of that technique. It has an average thermal power of only 2 MW, is operated at 5 Hz with a pulse duration of about 230 μs, and reaches a thermal peak flux of 2×10^{16} cm^{-2} s^{-1} (❯ *Table 2* and ❯ *Fig. 19*).

The "high" repetition rate of the IBR-2 compared to TRIGA reactors is only possible due to periodically running through a prompt supercritical configuration (in IBR-2 a moving reflector). But therefore the external control mechanisms (absorbers) of the continuously operating

Table 2

Comparison of the performances of various modern neutron sources. Except for the first two, only spallation sources are quoted. Fluxes of the pulsed spallation sources are for a coupled, unpoisoned H_2O moderator (the ESS-LP parameters are as of 2010)

	Accelerator type, energy, average current	Average power (MW)	Proton pulse duration (10^{-6} s)	Pulse repetition rate ν (s^{-1})	$\overline{\Phi}$ ($cm^{-2}\ s^{-1}$)	$\hat{\Phi}$ ($cm^{-2}\ s^{-1}$)	$\hat{\Phi}\cdot\nu$ ($10^{17}\ cm^{-2}\ s^{-2}$)
High flux reactor ILL Grenoble (FR)	N/A	57	N/A	N/A	1.2×10^{15}	1.2×10^{15}	1.2[a]
Pulsed reactor IBR-2 Dubna (RU)	N/A	2	N/A	5	2×10^{13}	2×10^{16}	1
SINQ, steady st. Pb target, D_2O tank PSI (CH)	Isochronous cyclotron, 0.59 GeV, 2 mA	1.2	N/A	N/A	1×10^{14}	N/A	N/A
ISIS Ta target Chilton (UK)	Rapid cycling synchroton, 0.8 GeV, 0.2 mA	0.16	0.39	50	7×10^{12}	8×10^{14}	0.4
LANSCE W target Los Alamos (USA)	Linac + 1 compr. ring, 0.8 GeV, 0.1 mA	0.08	0.27	20	3.4×10^{12}	1×10^{15}	0.2
J-PARC Hg target Tokai (Japan)	Rapid cycling synchroton, 3 GeV, 0.333 mA	1	1.2	25	1.3×10^{14}	3×10^{16}	7.5
SNS Hg target Oak Ridge (USA)	Linac + 1 compr. ring, 1 GeV, 2 mA	2	0.69	60	1.2×10^{14}	1.2×10^{16}	7
ESS SP 2003 Hg target (abandoned)	Linac + 2 compr. rings, 1.33 GeV, 3.75 mA	5	0.96	50	3.1×10^{14}	1.3×10^{17}	65
ESS-LP Hg target (Lund, S)	Linac only, 1 GeV, 5 mA	5	2,000	16 2/3	3.6×10^{14}	1.1×10^{16}	1.8

[a] with neutron chopper at 100 s^{-1}

reactors will not work. The power excursion must be limited by inherent mechanisms, e.g., by the temperature rise of the fuel. From this point of view, it is immediately obvious that not any desired pulse structure, i.e., a much higher pulse repetition rate and shorter pulse duration will be realizable.

3 Spallation: The Future for Rapidly Pulsed Neutron Sources

Although it may be unlikely in reality (and never happened at the IBR-2), malfunctions of the necessarily mechanical insertion of excess reactivity (rotating parts of the IBR-2 reflector) may lead to substantial damage of the reactor core with the related environmental hazards. So, rapid cycling (50 Hz or more) reactors do not present the desired solution. The exploitation of the spallation reaction has meanwhile proven its potential as the neutron source of the future (Sect. 3.3). And, importantly, accelerator-driven spallation sources are inherently safe, because *no critical configuration* is needed for the neutron production. Since the proton-driven reaction can only be sustained by external energy supply, it ceases with the shutdown of the ion (proton) source. Therefore, a spallation neutron source can be brought into an *immediately safe condition* within a few milliseconds. Moreover, neutron generation by protons enables the virtually *unlimited tailoring of pulse structures*, i.e., pulse durations from milliseconds down to below 1 μs and pulse repetition rates basically unfeasible with mechanical devices. This flexibility even enables the feeding of more than one target with a single accelerator. The future extension to a second target station is the major advantage of all existing or planned spallation sources (Sect. 3.3.2).

3.1 The Spallation Reaction

For kinetic energies above about 100 MeV, protons cause a reaction in atomic nuclei, which leads to a release of a large number of neutrons, protons, mesons (if the proton energy is above 400 MeV), nuclear fragments, and γ radiation. The spallation reaction is a two-stage process, which can be distinguished by the spatial and spectral distribution of the emitted neutrons. This is depicted schematically in *Fig. 1*.

In stage 1, the primary proton knocks on a nucleon, which in turn knocks on another nucleon of the same nucleus (intra-nuclear cascade) or of a different nucleus (internuclear cascade). With increasing energy of the primary particle the nucleons kicked out of the nuclei will for kinematic reasons (transformation from center-of-mass to laboratory system) be emitted into decreasing solid angles around forward direction. The energy distribution of the cascade particles extends up to the primary proton energy (*Fig. 3*). After emission of the cascade particles the nuclei are in a highly excited state, the energy of which is released in stage 2 mainly by evaporation of neutrons, protons, deuterons, α particles, and heavier fragments as well as γ radiation. Depending on the particular evaporation reaction course, different radioactive nuclei remain. The evaporation neutrons are isotropically emitted. They are the primary source neutrons, in which we are interested in the present context. The spectrum of the evaporation neutrons is very similar to that of nuclear fission and has a maximum at about 2 MeV (*Fig. 3*). This is the very reason why we can utilize the spallation neutrons as with a fission reactor.

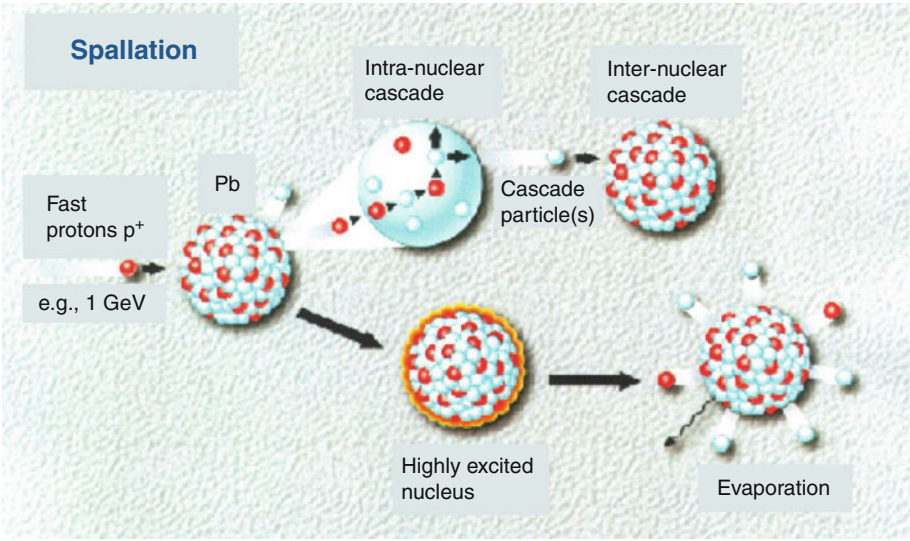


Fig. 1
The spallation reaction

3.1.1 Total Neutron Yield

The neutron yield had been calculated numerically by a Monte Carlo high-energy transport code system, which treats the processes during the intra-nuclear cascade and the entailing evaporation of neutrons (Prael and Lichtenstein 1989). Experimentally, the total yield of evaporation neutrons emitted from the target surface had been determined for several target materials, target geometries, and proton energies (Bartholomew and Tunnicliffe 1966). Examples of these data are shown in ❯ *Fig. 2*.

The data of ❯ *Fig. 2* can be described by the following approximate empirical expression (Carpenter 1977)

$$Y = 0.1 \times (A + 20) \times (E - E_0)\,[\text{neutrons/proton}]\,, \tag{3}$$

where A is the mass number of the target material ($A > 9$, except ^{238}U), E is the proton energy ($0.2\,\text{GeV} \le E \le 1.5\,\text{GeV}$), and $E_0 = 0.12\,\text{GeV}$. For uranium, which releases neutrons by fission as well, the total yield is described by $Y = 40 \times E$. Neutron yields depend on target size and are somewhat larger for larger targets. The total number N of neutrons emitted per second, e.g., for a 61 cm long, 10.2 cm diameter lead target is given by

$$N^{\text{Pb}} = 1.25 \times 10^{17} \times I\,[\text{mA}] \times (E\,[\text{GeV}] - 0.12)\,, \tag{4}$$

where I denotes the proton current in units of milliamperes. The offset E_0 in ❯ Eq. 3 expresses the fact that for $A < 210$ and proton energies below E_0 no spallation reactions take place.

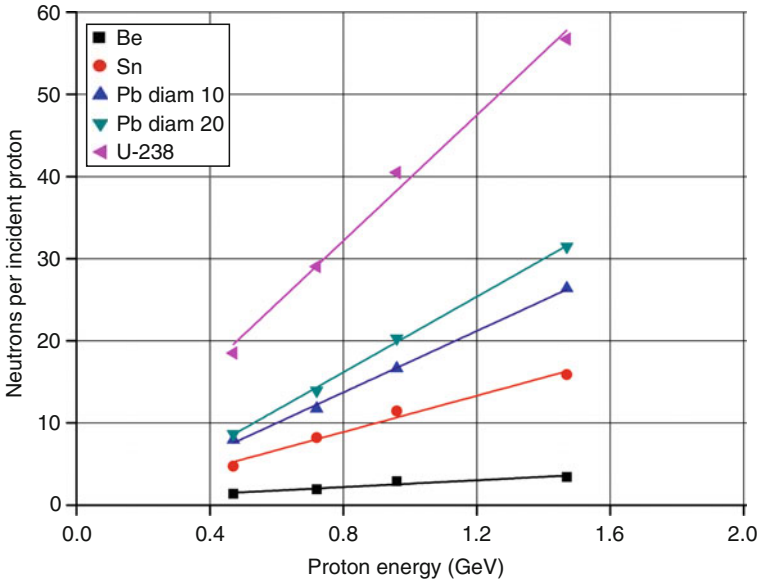


Fig. 2

Experimental fast-neutron yield *Y* for selected targets as a function of proton energy. Targets: Be (10 × 10 cm^2, 91.6 cm long); Sn, Pb, and U-238 (10 cm diameter) and Pb diam 20 (20 cm diameter), all 61 cm long (linear fits to data points from Bartholomew and Tunnicliffe (1966))

3.1.2 Spectral Distribution

The spectra of neutrons emerging from lead targets for proton energies of 590 MeV and 800 MeV have been measured and are shown in *Fig. 3* (Raupp F, Cierjacks S, Hino Y, Buth L, Howe SD 1980, unpublised). The spectra clearly consist of two components, the evaporation components with maxima around a few MeV and the cascade components extending up to the energy of the incident protons. The evaporation spectrum is very similar to the fission spectrum of a reactor, which is the very reason for making spallation competitive to reactors at all. The cascade component on the other hand, although containing only 3–5% of the total fast neutrons generated, entails particular shielding requirements for the target stations.

3.1.3 Source Distribution, Proton Mean Free Path, and Range

Neutron production inside the target is exponentially decreasing over a distance comparable to the effective range of protons in the particular target material with a build-up zone at the target front end (*Fig. 4*) (Filges et al. 1995). The exponential decay constant is the mean free path Λ for collisions of high-energy nucleons, which can be estimated from the total inelastic cross section as given by the following expression (Ashmore et al. 1960):

$$\sigma_{\text{inel}} = 15.9\,\pi\,A^{2/3}\,[\text{mbarn/nucleus}]\,. \tag{5}$$

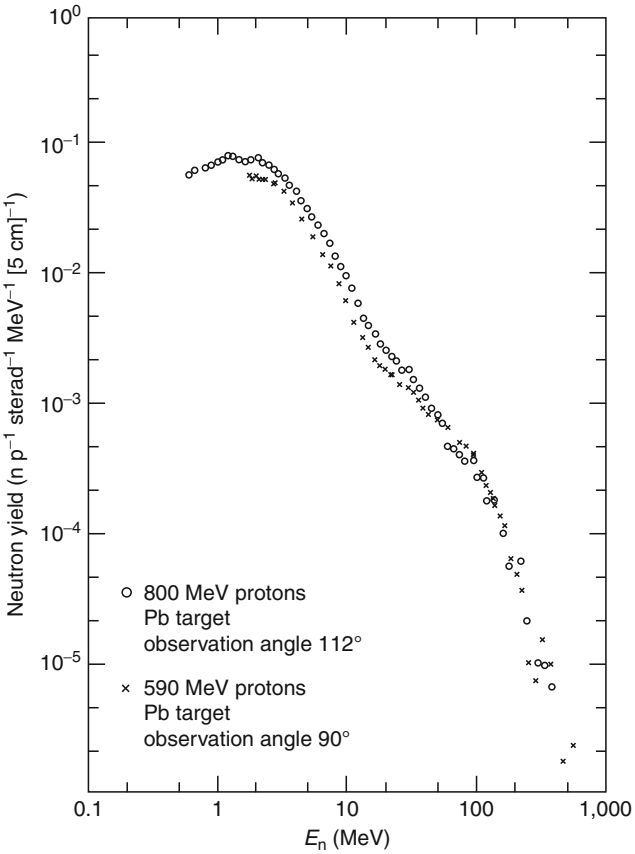


Fig. 3

Differential spectrum of neutrons emitted at 90° and 112°, respectively, from the first 5 cm of a 10 cm diameter lead target for incident proton energies of 590 MeV and 800 MeV, respectively (Raupp F, Cierjacks S, Hino Y, Buth L, Howe SD 1980, unpublised)

The proton mean free path is then given with Eq. 5 as the inverse macroscopic cross section Σ

$$\Lambda = \Sigma^{-1} = (n \cdot \sigma)^{-1} = 33.2\,[\mathrm{g/cm^2}]\,A^{1/3}\rho^{-1}, \tag{6}$$

where n is the number density and ρ the mass density of the target material. Differential neutron production $P'(z)$ along the target axis z is therefore written approximately as $P'(z) = P'(0)\exp(-\Sigma \cdot z)$. The effective proton range $R(E)$ can be expressed by the following empirical relationship (Carpenter 1977):

$$R(E) = 233\left[\mathrm{g/cm^2}\right] \times \rho^{-1} \times Z^{0.23} \times (E[\mathrm{GeV}] - 0.032)^{1.4}\,,\ \text{for } Z > 10 \text{ and } 0.1\,\mathrm{GeV} \le E \le 1\,\mathrm{GeV}\,.$$

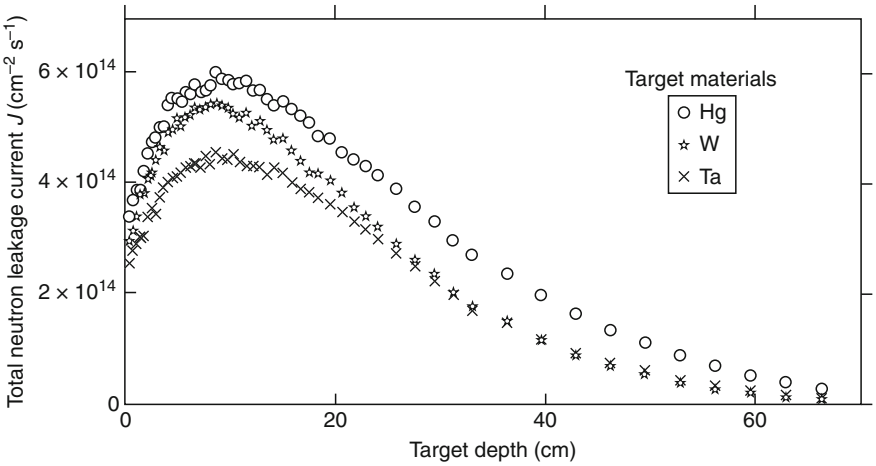


Fig. 4

Calculated axial leakage distributions of fast neutrons from a lead-reflected mercury target compared to water-cooled tantalum and tungsten targets, respectively (Filges et al. 1995)

3.1.4 Heat Deposition

Heat is deposited within the target partly directly by ionization due to primary protons and partly by spallation products. Specific ionization losses $dE/dS\ [\mathrm{MeV}/(\mathrm{g\ cm^{-2}})]$ of charged particles are steeply decreasing with increasing kinetic energy and exhibit a flat minimum at an energy of about twice the rest mass of the particle (Jackson 1963). This is in favor of using a proton energy of the order of a few GeV (*Fig. 5*). Other considerations show that an energy of 1–1.5 GeV is sufficient (Sect. 3.2.1). Examples of heat deposition in lead and depleted uranium targets are shown in *Fig. 5*.

Monte Carlo calculations (R.D. Neef, 1995, private communication) have shown that the heat deposition is, except for a build-up zone at the target front end, exponentially decreasing with the same mean free path as given above (*Fig. 6*).

Concluding, some useful physical and neutronic data of possible target materials are complied in *Table 3*.

3.2 Technical Details of a Pulsed Spallation Source

The generic spallation neutron source consists of three major components, the proton accelerator, the target, and the moderators. For reasons discussed in Sects. 1 and 2, all spallation sources but one are pulsed. Based on arguments like relation 2 one might wish to get the highest possible thermal-neutron peak flux at the highest possible repetition rate for a given time-average flux. However, the highest peak flux is not simply proportional to the highest proton pulse current, but rather determined by the physics of neutron slowing down and diffusion in moderators (Sect. 3.2.3).

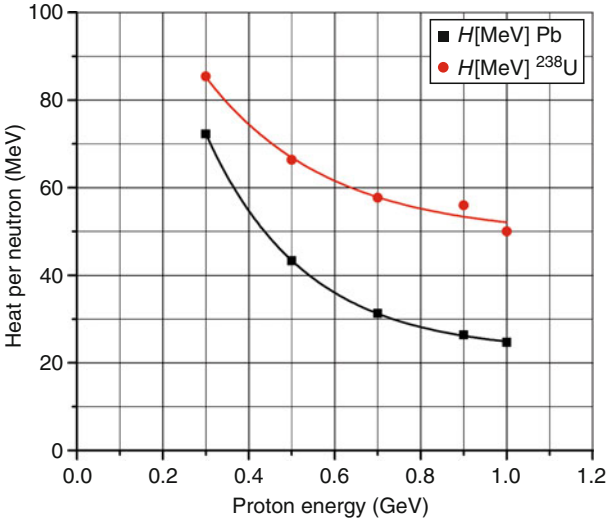


Fig. 5

Heat deposition *H* (MeV/n) in lead and depleted uranium targets of 10 cm diameter and a length of 61 cm as a function of proton energy. Lines are a guide to the eye only. (Data points from Bartholomew and Tunnicliffe (1966).) For comparison: 200 MeV/n in fission reactions

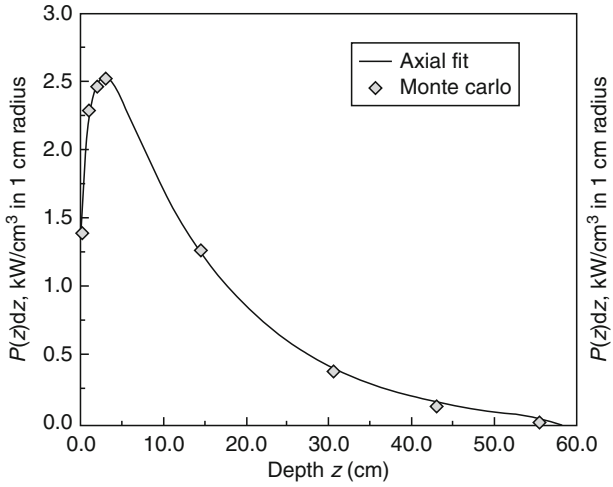


Fig. 6

Monte Carlo calculations of the power deposition density in a cylinder of 1 cm radius along the axis of a mercury target for a beam power of 5 MW (1.33 GeV, 3.75 mA) (R.D. Neef, 1995, private communication). The total deposited power is 2.8 MW. Note the drop to zero power deposition near proton range

Table 3

Some useful physical and neutronic data of possible target materials (atomic number, *Z*; atomic mass, *A*; density, ρ; melting temperature, T_M; boiling temperature, T_B; proton mean free path, Λ; proton range, *R*; total neutron yield, *Y*)

		Ta	W	Hg	Tl	Pb	Bi	Th	^{238}U
Z		73	74	80	81	82	83	90	92
A		180.95	183.85	200.59	204.37	207.19	208.98	232.04	238.03
ρ**(g/cm^3)**		16.6	19.3	13.6	11.8	11.4	9.8	11.7	19.1
T_M**(°C)**		2,996	3,410	−38.4	303	327	271	1,750	1,132
T_B**(°C)**		5,425	5,930	357	1,457	1,725	1,560	3,850	3,818
Λ **(cm)**		11.3	9.8	14.4	16.6	17.4	20.1	17.8	10.8
***R*(*E*) (cm)**	0.5 GeV	13.4	11.5	16.7	19.2	20	23.3	19.9	12.2
	1 GeV	36.4	31.4	45.4	52.5	54.5	63.6	54.3	33.8
	1.5 GeV	65	56.1	81	93.7	97.2	113.4	96.8	59.6
***Y*(*E*) (n/p)**	0.5 GeV	7.6	7.7	8.4	8.5	8.6	8.7	9.6	20.8
	1 GeV	17.7	17.9	19.4	19.7	20	20.2	22.2	40
	1.5 GeV	27.7	28.1	30.4	31	31.4	31.6	34.8	59.2

3.2.1 The Accelerator

If proton pulse durations in the microsecond range are demanded, the accelerator (Chap. 7, "Accelerators for Particle Physics") will be a complex multicomponent machine. With the present technology (i.e., proton induction accelerators excluded) the three main components of a high-power machine are (1) a high-current ion source, (2) a linear accelerator ("linac"), actually composed of three major parts, and (3) finally followed by one or more accumulator ring(s), which will compress the usually millisecond long linac pulses to the desired final duration. In addition, there are nontrivial transport lines between the different components as well as to the target. Before we will describe the machine in detail below, reasonable values for the proton energy and the current have to be selected.

Choice of Proton Energy

Since the fast-neutron yield rises linearly with the energy of the incident protons (Eq. 3 and *Fig. 2*), it seems at a first glance obvious to use the highest possible energy, whereby the lowest ion currents could be used for a given beam power. The latter is important, because high-current ion sources are still a challenge for high-power accelerators, in particular when H^--ion sources are needed for injection into compressor rings to get very short proton pulses (see below). However, to utilize the highest proton energies would be a valid argument only, if the investment and operation costs of the respective accelerator were to increase less than linear. Moreover, due to the offset E_0 = 0.12 GeV in Eq. 3, the neutron yield per MW of beam power levels off as shown in *Fig. 7*. This experimental verification of the mathematical consequences of Eq. 3 sets a natural limit to the design parameters energy and current of the accelerator. From *Fig. 7*, it is clear that choosing proton energies beyond 1–1.5 GeV does not really gain much in total neutron yield. Actually, the three major spallation sources use or plan to use energies between 0.8 and 1.4 GeV (*Table 2*).

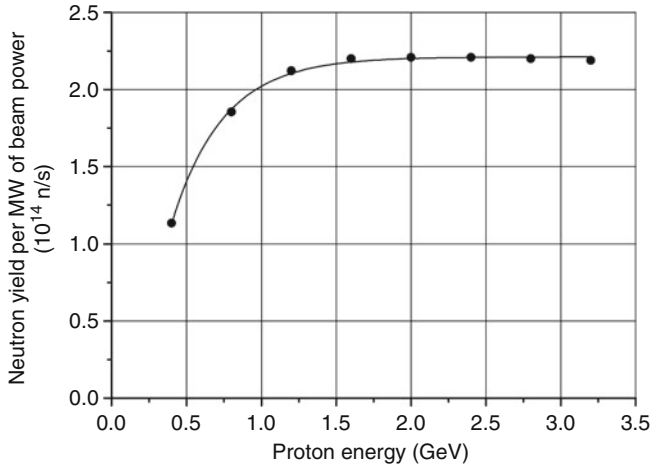


Fig. 7
Calculated total fast-neutron production rate per MW of proton beam power as a function of energy. The line is a guide to the eye. Data points from Pynn et al. (1993)

Proton-Current Requirements: A Source-Strength Estimate
Typical target dimensions are very similar to the compact core of a modern research reactor (FRM II: 20 cm outer diameter, 60 cm long annulus). The spectrum of the emitted fast-evaporation neutrons is quite similar as well. It can therefore reasonably be assumed that a comparable number of primary neutrons have to be generated in targets in order to obtain comparable thermal fluxes in a water moderator placed next to the target (*Fig. 8*). So, how many protons per unit time have to be injected into, e.g., a lead target (20 neutrons per proton of 1 GeV) in order to obtain a thermal neutron flux comparable to the steady state flux of the ILL reactor? Using relation 1 one calculates for the 57 MW ILL reactor a neutron source strength of about 2×10^{18} n/s, which yields $(2 \times 10^{18}\ \text{n/s})/(20\ \text{n/p}) = 10^{17}$ p/s. From the relation 1 ampere $\hat{=} 6.25 \times 10^{18}$ p/s follows $(10^{17}\ \text{p/s})/(6.25 \times 10^{18}\ \text{p/s/A}) = 0.016$ A. A current of 0.016 A at 1 GeV corresponds to a beam power $L = 0.016\ \text{A} \times 10^{9}\ \text{V} = 16$ MW. This estimate confirms the experimental results of the German feasibility study for a 5.5 MW spallation source (Bauer et al. 1983). According to this study, a 1.1 GeV/5 mA proton beam on lead generates a steady state thermal flux in a water moderator, which is equivalent to about a third of the flux of the ILL reactor (compare *Fig. 9* and Eq. 7).

3.2.2 The Target: Solid or Liquid?

According to Eq. 3, heavy elements are favored as target materials, in particular the refractory metals tantalum, tungsten, or rhenium, but also lead, bismuth, mercury, or even uranium. Whatever material is selected, it will be subject to heavy multiple loads. Firstly, about 60% of the average beam power is dissipated within the target as heat; the rest is transported as released radiation to the target vicinity like moderators, reflectors, and shielding or is converted into nuclear binding energy. Secondly, all materials hit by protons (and fast neutrons) will suffer

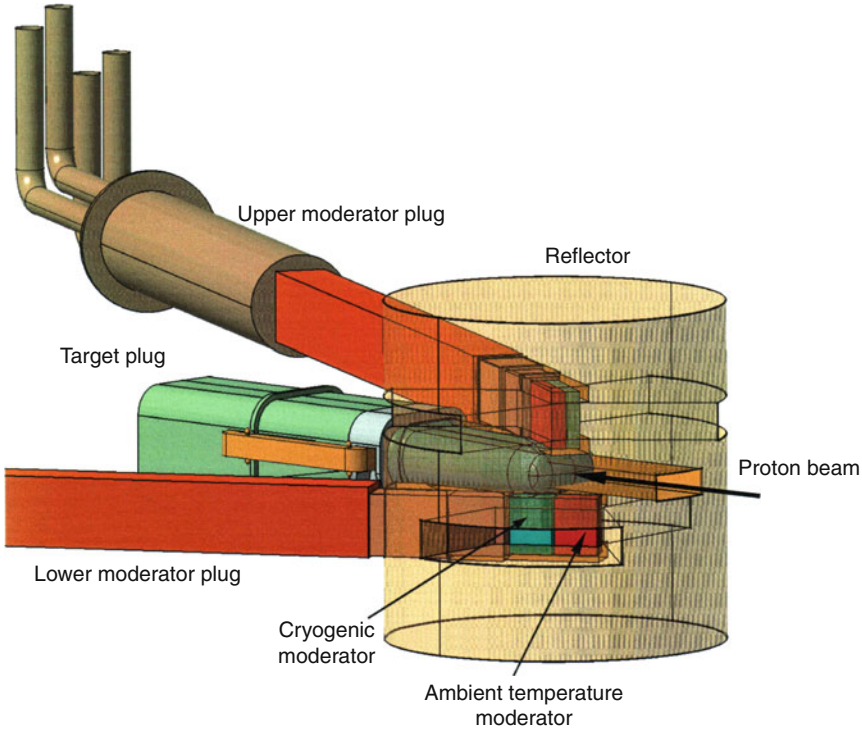


Fig. 8

Visualization of the ESS target–moderator–reflector block. Both the target (center) and the moderators (above and beneath the target) can be exchanged via horizontal plugs inside the target shielding. Proton beam injection is from the right. Moderators beneath the target are placed with their narrow faces "side by side," above with their large faces "back to back." The big cylinder (about 1.5 m diameter) is the heavy-metal reflector

from radiation damage (Chap. 22, "Radiation Damage Effects"). In order to both keep average target temperatures low and reduce specific radiation damage and loads due to dynamic effects from shock waves, a solid rotating target is conceivable and had originally been proposed for the ESS. As any solid target has to be cooled, it will inevitably be "diluted" by the coolant, whereby the primary source's luminosity will be diminished. One should therefore operate the target in its liquid state avoiding an additional cooling medium. Radiation damage would be no longer a problem with the target, but of course with its container. Refractory metals are excluded due to their high melting points. So we are left with elements like lead, bismuth, the Pb–Bi eutectic, or mercury. In fact, mercury has eventually been selected for the ESS and is realized with the US and Japanese sources, because it exhibits favorable neutron yield as shown in *Fig. 4*. The dimension of a target along the beam path will reasonably be chosen according to the range of the protons. For mercury and a proton energy of 1 GeV this is about 50 cm. Lateral target dimensions are optimized so that the moderators are not too far from the target axis (solid angle!). A typical target–moderator–reflector configuration is depicted in *Fig. 8*.

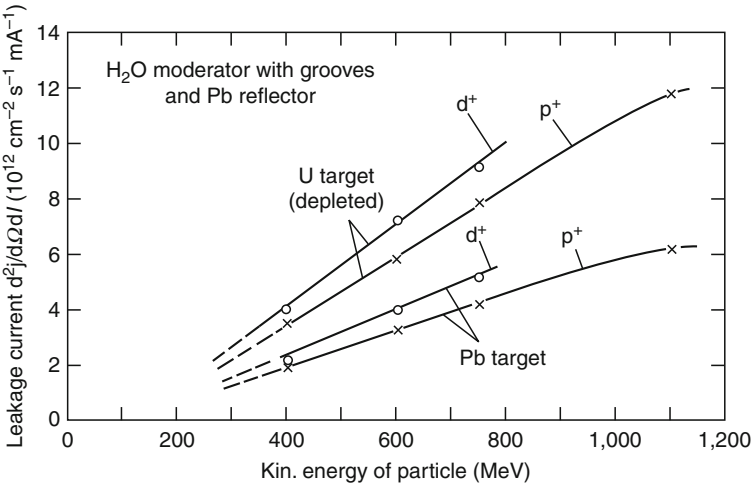


Fig. 9

Thermal time-average neutron leakage current per mA of incident charged-particle current from a water moderator placed next to targets of lead and uranium as a function of the energy of the charged particles (Bauer et al. 1983)

Remark *Only recently (2010), reservations against mercury have been posed by the authorities based on environmental arguments. For the ESS, a solid rotating target wheel like in the German SNQ project (1985, SNQ project proposal for a spallation neutron source. Kernforschungsanlage Jülich, unpublished) might be revived.*

Moderators (Sect. 3.2.3) can be placed above and below the target (*Fig. 8*). Positioning the moderators along the target axis, on the other hand, is governed by the leakage distribution of the evaporation neutrons (*Fig. 4*). Since this distribution exhibits a pronounced maximum of typically only 30–40 cm half width, even the positioning of two moderators on the same side of the target is a matter of compromise (one on each sides of the maximum or one at the maximum and the second down at the slope).

The reflector (*Fig. 8*) is a device surrounding the moderators, which returns those fast neutrons not being captured by the moderator on the first passage. A gain of a factor of 2 had been observed experimentally. Depending on the requirements, the reflector material can be moderating (beryllium or carbon) or non-moderating (lead, nickel, or other heavy elements). Moderating materials will extend the neutron pulse widths (Sect. 3.2.3).

3.2.3 The Moderators

We will now turn to the "heart" of the facility, the moderators, which were just shown in *Fig. 8* in their positions next to the target. As the upper and lower faces of the target are equivalent for symmetry reasons with respect to the emission of fast neutrons, it is obvious to place moderators on both sides. The question now is, whether D_2O as the slowing-down medium shall be used like in all modern medium- and high-flux reactors or possibly H_2O. As discussed in

Sect. 2.2, not the highest possible average neutron flux is the only reasonable demand, but rather the highest possible peak flux for a given (or requested) average flux. In that respect, H_2O is the preferred material due to its bigger slowing-down power and stronger absorption for thermal neutrons (see Eqs. 9 and 10). The reason for this seemingly paradoxical demand for stronger absorption is that the achievable neutron peak flux not only rises with the proton peak current, although not simply proportionally, but also depends on the lifetime τ of thermal neutrons in the moderator (Eq. 10).

In *Fig. 9* is shown the time-average thermal-neutron leakage from a water moderator embedded in a lead reflector (*Fig. 8*). A linear increase of the thermal-neutron leakage with proton and deuteron energy is found like the fast-neutron yield (*Fig. 2*). For protons the start of a leveling off for higher energies is observed as in *Fig. 7*. In order to determine the advantage of neutron-containing charged particles, deuterons were used as well, although not taken into further consideration for a realistic design (Bauer et al. 1983). If we choose a proton energy of 1.1 GeV and a proton current of 5 mA on a lead target like for the German SNQ project (1985, SNQ project proposal for a spallation neutron source. Kernforschungsanlage Jülich, unpublished), we deduce from the data points in *Fig. 9* a time-average thermal-neutron flux from a water moderator of

$$\Phi_{\mathrm{th}} = 6.1 \times 10^{12} \left[\mathrm{cm}^{-2}\,\mathrm{s}^{-1}\,\mathrm{mA}^{-1}\,\mathrm{steradian}^{-1}\right] \times 4\pi \times 5\,[\mathrm{mA}] \approx 3.8 \times 10^{14}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}. \quad (7)$$

Peak and Time-Average Neutron Flux

In Sect. 2 there had been emphasized the advantages of pulsed operation of spallation sources with desired thermal peak fluxes of the order of 100 times larger than the average flux of the best reactor. However, the neutron peak flux does not simply rise proportionally with the proton peak current. In order to understand the mapping of a proton pulse upon the shape of a thermal neutron pulse, a brief excursion is undertaken into the physics of neutron moderation and diffusion.

We start with a rectangular proton pulse shape. Since the spallation reaction takes place within femtoseconds we may assume that the pulse of evaporation neutrons injected into the moderator is rectangular as well. Slowing down the fast MeV neutrons to the thermal (meV) range needs about 20 collisions in H_2O and lasts for about $t_S \approx 10\,\mu s$ (Beckurts and Wirtz 1964). Therefore, even a δ-shaped pulse will be broadened after slowing down. The subsequent diffusion will broaden the pulse further.

During slowing down, a substantial fraction of the neutrons will escape the moderator, because it has to have a finite size due to the inevitable absorption. These so-called *epithermal* neutrons are lost for the thermal flux, but are altogether desired for spectroscopic means (Sect. 4.1). They constitute a big advantage of spallation sources, because by dedicated moderator shaping (and other "tricks" like decoupling and poisoning, see moderator tailoring below) *epithermal* neutrons can be made available with orders of magnitude higher fluxes than with reactors.

Thermal neutrons will by diffusion reach after some time the moderator surface and escape. The characteristic time of this diffusion process is called the lifetime τ of the neutrons in a particular moderator. It is the main quantity to characterize a pulsed source and can be determined experimentally (Bauer et al. 1981). The lifetime depends upon various parameters like type and size of the moderator as well as upon the material and size of the moderator environment, the so-called reflector. Typical values of the lifetime τ for a water moderator inside a non-moderating

reflector (e.g., lead or nickel) are between 150 and 200 μs. The fact that the lifetime τ is distinctly longer than the slowing-down time t_S simplifies the mathematical description of the pulse structure. Neglecting the slowing-down time, the neutron pulse shape is the convolution of the proton pulse shape with an exponential $e^{-t/\tau}$ describing the neutron escape. For a rectangular proton pulse of width t_p the neutron flux is therefore given as

$$\Phi(t) = \Phi_{asymp} \cdot (1 - e^{-t/\tau}) \qquad \text{for } 0 \le t \le t_p\,,$$
$$\Phi(t) = \hat{\Phi} \cdot e^{-(t-t_p)/\tau} \qquad \text{for } t \ge t_p\,,$$

where the asymptotic value Φ_{asymp}, which would be reached with uninterrupted proton peak current, is given by (with the repetition time t_{rep} as the time between two pulses)

$$\Phi_{asymp} = \bar{\Phi}_{th}\, t_{rep}/t_p\,.$$

The peak flux $\hat{\Phi}_{th}$ will be reached at $t = t_p$,

$$\hat{\Phi}_{th} = \Phi\,(t = t_p) = \bar{\Phi}_{th} \cdot \frac{t_{rep}}{t_p} \cdot (1 - e^{-t_p/\tau})\,, \tag{8}$$

where $\bar{\Phi}_{th}$ denotes the time-average flux.

Numerical Example

For a hypothetical 5.5 MW linac (1.1 GeV, 5 mA average current), operated at a repetition rate of 50 Hz and a proton pulse width of $t_p = 500$ μs, and a moderator with a neutron lifetime $\tau = 150$ μs, the expression in ❯ Eq. 8 yields

$$\hat{\Phi}_{th} = 39 \times \bar{\Phi}_{th}\,.$$

Taking the (time-average) thermal flux of the experimental result (❯ Eq. 7), which is third of the flux of the ILL reactor, the corresponding peak flux would exceed the latter already by an order of magnitude.

On the other hand, the maximum achievable peak flux for a given average proton beam power is given by the limit of ❯ Eq. 8 for vanishing proton pulse width, i.e., for $t_p \to 0$,

$$\hat{\Phi}_{max} = \lim_{t_p \to 0} \hat{\Phi} = (t_{rep}/\tau) \cdot \bar{\Phi}_{th}\,, \tag{9}$$

i.e., even a δ-shaped proton pulse current results in a finite neutron peak flux. The theoretical maximum peak-flux value

$$\hat{\Phi}_{max} = 133.3 \times \bar{\Phi}_{th}$$

would be nearly reached with a proton pulse width of one microsecond, i.e.,

$$\hat{\Phi}_{1\,\mu s} = 132.9 \times \bar{\Phi}_{th}\,.$$

As mentioned above, the slowing-down time of about $t_S = 10$ μs had been neglected. A reasonable estimate of the error of neglecting t_S is obtained by inserting in ❯ Eq. 8 a fictitious proton pulse width of the order of the slowing-down time $t_S = 10$ μs. The result is $\hat{\Phi}_{10\,\mu s} = 129 \times \bar{\Phi}_{th}$, i.e., a value only about 3% lower.

Moderator Requirements

From relation ❯ 9, we see that the limit of the peak flux is inversely proportional to the moderator lifetime. But, even with finite current pulse widths a short lifetime is important for obtaining

large peak fluxes. The lifetime τ of a thermal neutron is a measure of the escape probability from the moderator and is obviously determined by both the geometry of the moderator vessel and the absorption cross section of the moderator medium and can be written (Beckurts and Wirtz 1964) as:

$$\tau = \left(\bar{\nu}_{\text{th}} \cdot \Sigma_{\text{abs}} + 3D\pi^2/L^2\right)^{-1}, \tag{10}$$

where $\bar{\nu}_{\text{th}} = 2.2\times10^5$ cm s^{-1} is the average neutron velocity, $\Sigma_{\text{abs}} = 2\times10^{-2}$ cm^{-1} the macroscopic absorption cross section, $D = 0.36 \times 10^5$ cm^2 s^{-1} the diffusion constant for thermal neutrons and L is a typical moderator dimension of the order of 10–15 cm. The absorption cross section of H_2O is about 700 times bigger than that of D_2O. If it were only for this reason, an H_2O moderator had to be small (small L in Eq. 10), because we want to, of course, utilize the neutrons that leak from the moderator. So, a short lifetime must not entirely be due to self-absorption. As, on the other hand, H_2O possesses the largest known slowing-down density (the number of neutrons, which become thermal per cm^3 and s), an H_2O moderator anyhow does not need to be big.

For these reasons a pulsed spallation source will have small ($V \cong 1.5$ l) H_2O moderators for thermal neutrons. The corresponding lifetime of such H_2O moderators has been measured and is $\tau \approx 165$ μs (Bauer et al. 1981), which is in good agreement with a calculation using Eq. 10. Small size and absorption diminish in any case the time-average neutron yield. In order to improve this without deteriorating the peak fluxes, two tricks are used. Firstly, a moderator is enclosed by a reflector (*Fig. 8*), a strongly scattering ("reflecting") but non-moderating material, i.e., a heavy element with a large scattering cross section like lead. Secondly, the leakage probability from the moderator interior, i.e., a region of higher flux due to geometrical "buckling" ($3D\pi^2/L^2$ in Eq. 10) is enhanced by holes or grooves pointing toward the neutron beam holes. Both measures give gain factors of 2 each, whereby the reflector gain is so to speak "for free," because the anyway necessary lead or iron shielding has the same effect.

Moderator Tailoring

With two schemes called *decoupling* and *poisoning*, respectively, the lifetime τ can be effectively influenced (*Fig. 10*). *Decoupling* means to cover the moderator with a material, which is strongly absorbing thermal neutrons, e.g., boron carbide or lithium halides. Thereby neutrons returning from the reflector are prevented from entering the moderator again; in other words, the reflector is *decoupled*. Strictly speaking, decoupling is only significant with moderating reflectors, i.e., consisting of beryllium, carbon, or D_2O. The situation is different, when *poisoning* is applied. With *poisoning*, a moderator, which is geometrically large in order to intercept a large portion of fast neutrons from the target (solid angle!), is made neutronically small by placing absorbing sheets inside the moderator. Thereby the neutron pulse is shortened due to suppressing diffusion between neighboring moderator regions.

3.3 Examples of Spallation Sources

3.3.1 The US Spallation Neutron Source SNS

The SNS (www.sns.gov) is the most powerful pulsed spallation source presently (2010). It is a so-called short-pulse source, which refers to proton pulse durations in the μs range. In principle, two accelerator variants are capable of producing pulses of that length, viz., synchrotrons or a

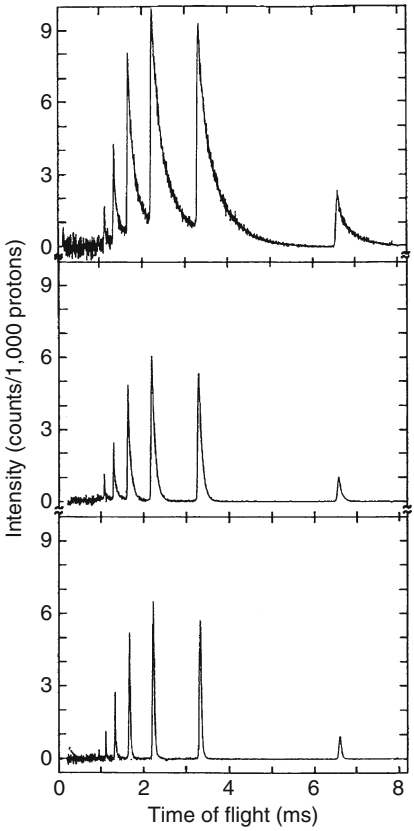


Fig. 10

Time-of-flight experiment to determine the influence of decoupling and poisoning a moderator. The peaks correspond to Bragg reflections (002) to (0012) from a pyrolytic graphite monochromator hit by the white moderator neutron beam. *Top*: no decoupling, no poisoning; *middle*: decoupled; *bottom*: decoupled and poisoned. The shorter lifetimes are clearly seen. Remarkably, the peak intensity of the individual reflections is diminished only moderately by decoupling and poisoning

combination of a pulsed linac and a compressor (accumulator) ring. The SNS represents the latter, because synchrotrons are not yet capable of delivering the proton currents demanded for the SNS.

With this concept, the linac will supply the full beam power, and the subsequent accumulator ring will compress the pulses from the linac. The basic SNS technical design parameters are:

Linac proton energy	1 GeV
Average current on target	2 mA
Average beam power	2 MW
Repetition rate	60 s^{-1}

The Components of the Accelerator

Ion Source and Linac

An ion source injects negative hydrogen (H^-) ions into a radio-frequency quadrupole (RFQ) pre-accelerator, which delivers 2.5 MeV ions to the linac proper. The linac is composed of a conventional (Alvarez) drift-tube linac (DTL), followed by a so-called coupled-cavity linac (CCL), both at room temperature and a high-energy part based on superconducting radio-frequency niobium cavities (SRF) operating at liquid-helium temperature of 2.1 K.

DTL output energy	87 MeV
CCL output energy	185 MeV
SRF output energy	1,000 MeV
Total linac length	331 m
Linac peak H^- current	52 mA
Linac pulse duration	1 ms
High-energy beam transport to ring	170 m

Accumulator (Compressor) Ring

In order to obtain a pulse width of the order of 1 μs, the long H^- linac pulse has to be wrapped into the ring through a stripper foil that strips the electrons from the negatively charged hydrogen ions to produce the protons (H^+). Negative ions are needed for injection into the ring in order to overcome problems due to Liouville's theorem when painting the momentum space during injection into the ring. Approximately 1,060 turns are accumulated and then all these protons are kicked out at once, producing a pulse of 0.695 μs width. In fact, this pulse is obviously shorter than 1/1,060 of the 1 ms long linac pulse due to 0.250 μs long gaps within the linac pulse. The gaps are necessary for a virtually loss-free extraction from the ring.

Accumulator ring diameter	79 m
Ring orbit rotation time	0.945 μs
Ring peak current	47.5 A
Pulse duration after compression	0.695 μs
Ring beam extraction gap	0.250 μs
Ring-to-target transport system length	150 m
Protons per pulse on target	2.1×10^{14}
Proton pulse duration on target	0.695 μs

Target Building

Another about 120 m long transport line will bring the compressed pulses down to the target building. The target is mercury and is located in the center of a 12-m diameter shielding block. Neutrons from the four moderators (one ambient, one composite, and two cold) are extracted through 18 beam channels to the scattering instruments. A part of the beam channels is extended by neutron guides to instruments outside the target building proper.

An artist's view of the entire complex of the US spallation source SNS is shown in *Fig. 11*. An important aspect of the versatility of an accelerator-driven neutron source is visualized in this figure, viz., a future second target building. The linac can easily be operated with a different (higher) repetition rate (at the expense of a higher power input, of course), so that freely selectable numbers can be directed to different targets.



Fig. 11

Artist's view of the US spallation neutron source SNS building complex. With its accelerator, compressor ring, and ancillary buildings it differs distinctly from a reactor laboratory. A future second target building is shown as well, which expresses the versatility of an accelerator-driven neutron source

3.3.2 The European Spallation Source (ESS)

In its final project proposal published in 2003 (ESS, 2003, The ESS project, 4 volumes, www.ess-europe.de, and F. Mezei, P. Tindemans, K. Bongardt, 2008, The 5 MW LP ESS; Best price-performance, unpublished), the ESS was designed for two target stations, with a beam power of 5 MW each. The first, a long-pulse (LP) station would have been fed by the linac directly and the second, short-pulse (SP) station by compressed pulses from two accumulator rings. The main project parameters are given in *Table 4*. The project had not been pursued further until in 2008 an initiative revived the project on the basis of an ESFRI roadmap. Meanwhile a site on the outskirts of the city of Lund in southern Sweden had been established. The parameters for the ESS had been changed considerably and are presented in *Table 5*. The major change is the restriction to a single 5 MW linac-fed LP target station.

Concluding, in *Table 2* is presented a neutronic performance comparison of selected typical neutron sources with emphasis on pulsed sources.

4 Experimental Methods at Spallation Neutron Sources

Neutron scattering is an intensity-limited technique. In other words, better resolution in all kinds of experiments is only possible with more intense sources. Obviously, the best source is that providing the highest flux. Due to engineering limits with steady state sources the

Table 4

Main parameters of the ESS as of 2003

Linac			
Beam power		10 MW	
Linac beam energy		1.334 GeV	
Ion sources		$2H^-$	
	SP		**LP**
Linac average current	3.75 mA		3.75 mA
Linac peak current	112.5 mA		112.5 mA
Linac repetition rate	50 Hz		16 2/3 Hz
Linac beam pulse duration	2 × 0.48 ms		2 ms
Two accumulator rings			
Frequency of parallel operation		50 Hz	
Number of circulating protons per ring		2.34×10^{14}	
Revolution frequency		1.2416 MHz	
Bunch length at ring ejection		0.6 μs	
Peak current		62.5 A	
Mean radius of rings		35 m	
Target type	**Flowing mercury**		**Flowing mercury**
Number of moderators (viewed faces)	2 (4)		2 (4)
Average thermal flux	3.1×10^{14} cm^{-2} s^{-1}		3.1×10^{14} cm^{-2} s^{-1}
Peak thermal neutron flux	1.3×10^{17} cm^{-2} s^{-1}		1.0×10^{16} cm^{-2} s^{-1}

Table 5

Main parameters of the revived ESS single-target-station project as of 2010

Linac	**LP**
Beam power	5 MW
Linac beam energy	1.0 GeV
Ion sources	$2H^+$
Linac average	5 mA
Linac peak current	2 × 75 mA
Linac repetition rate	16 2/3 Hz
Linac beam pulse duration	2 ms
Target type	**Flowing mercury**
Number of moderators (viewed faces)	2 (4)
Average thermal flux	3.6×10^{14} n/cm^2 s
Peak thermal neutron flux	1.1×10^{16} n/cm^2 s

introduction of pulsed operation triggered the real progress for neutron scattering. Pulsed spallation sources, in particular, even widened the frontiers such that neutrons of intermediate energies, the *epithermal* neutrons, are available with intensity orders of magnitude higher than with reactors. Moreover, all methods employing the time-of-flight technique in spectroscopy

or diffractometry open doors to research fields inaccessible with reactors. For that reason, we will not discuss in the following experiments and methods, which rely on the largest available *average* flux.

4.1 *Epithermal* Neutrons: An Important Reason for Ultra-Short Proton Pulses

The *epithermal* neutrons stem from the intermediate stage during slowing down with energies in the keV to eV regime. Due to neutron economy they are to be avoided with fission reactors, since they would be missing in the thermal flux. Moreover, they were pretty useless with continuous sources, because their monochromatization is virtually impossible by the classical Bragg reflection technique due to vanishing Bragg angles. In order to achieve sufficient energy resolution $\Delta E/E = 2\Delta t/t$ with the time-of-flight method, on the other hand, one has either to use unreasonably long flight paths (large t) or short pulses (small Δt).

Since the pulse width of *epithermal* neutrons depends on their energy, it is only limited by the proton pulse width. This energy dependence is qualitatively obvious, if one takes into account that neutrons, which left the moderator prior to being in thermal equilibrium, had the shorter duration (= pulse width) inside the moderator, the fewer collisions they made, i.e., the higher the remaining energy is. The mathematical expression for that is given by (Beckurts and Wirtz 1964)

$$\Sigma_s \, v(E) \, \Delta t_E = \text{const}\,, \tag{11}$$

where Σ_s is the macroscopic scattering cross section of the medium ($\Sigma_s = 1.33\,\text{cm}^{-1}$ for water), $v(E)$ the neutron velocity at energy E and Δt_E the variance of an initially δ-shaped pulse when reaching the energy E. For water and $E = 1\,\text{eV}$ we obtain $\Delta t_E = 0.93\,\mu\text{s}$. This is a direct justification to aim at proton pulse widths below the slowing-down time t_S (about 10 μs in water), e.g., $t_p = 1\,\mu\text{s}$.

4.2 Spectroscopy at High Energy Transfers

Neutrons with high incident energy are needed to determine energy levels of the order of several hundred meV, which are virtually not populated thermally in samples due to the Boltzmann factor. These energy levels have to be excited with energy loss for the neutron. Moreover, it is important not to transfer too large a momentum to the sample, because a large momentum transfer $\underline{Q} = \underline{k} - \underline{k}'$ ($\underline{k}$ and $\underline{k}'$ are the wave vectors before and after scattering, respectively, $Q^2 = k^2 + k'^2 - 2kk'\cos\theta$) inevitably yields the excitation of multiphonon processes or even recoil effects at individual nuclei of the sample. Both impedes the interpretation of data or even renders it impossible. In ❯ *Fig. 12* is depicted the condition that a high energy transfer $\hbar\omega = E - E' = (\hbar^2/2m)(k^2 - k'^2)$ at simultaneous small momentum transfer $\underline{Q}$ is fulfilled only with high incident neutron energies.

In addition, ❯ *Fig. 12* displays the accessible measuring range within the Q–ω plane for two typical neutron energies of a spallation source. One from the Maxwell spectrum of thermal neutrons with an energy of $E = 40\,\text{meV}$ and another for epithermal neutrons with an energy $E = 1\,\text{eV}$. The latter obviously exhibits an enormous extension of the data acquisition range within the Q–ω space. In ❯ *Fig. 13* is shown an example of a measurement with epithermal

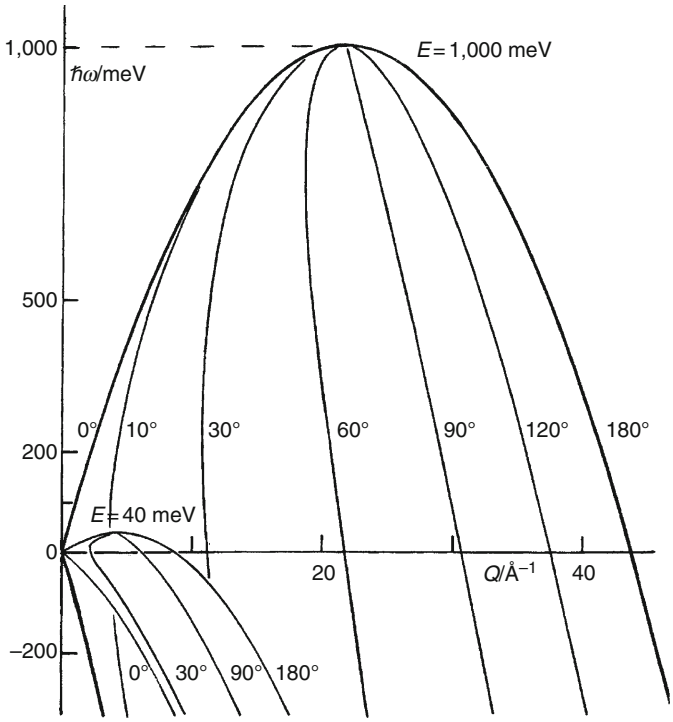


Fig. 12

Measuring area within the Q–ω plane for thermal (40 meV) and epithermal (1 eV) neutrons. The areas cover for both cases the complete scattering-angle extent $0° \leq 2\theta \leq 180°$. The set of curves relate momentum transfer $|\underline{Q}|$, energy transfer $\hbar\omega$, and scattering angle 2θ

neutrons (Fillaux et al. 1995). Since the employed time-of-flight spectrometer MARI (Multi Angle Rotor Instrument at ISIS) covers a scattering-angle range from 3° to 135°, a nearly complete overview of the dynamics is available this way. In particular, the superposition of the C–H vibration modes in coal (170 and 380 meV) with the recoil effects on hydrogen and carbon, respectively, has to be noted. To illustrate the latter, the recoil lines are indicated as well. In ❯ *Fig. 14*, the dispersion curves of a spin-Peierls system ($CuGeO_3$ at $T = 10$ K) are shown together with an excitation continuum up to 40 meV as measured at MARI as well (Arai et al. 1996).

4.3 Powder Diffractometry at Pulsed Sources

Powder diffractometry can be very efficiently performed in time-of-flight mode (Conrad et al. 2008). Sufficiently short neutron pulses are needed for the diffractometry of poly-crystalline samples in order to be able to separate by the time-of-flight technique densely packed Debye–Scherrer rings (❯ *Fig. 15*). Short neutron pulses are achieved by moderator tailoring as shown in ❯ *Fig. 10*. Moreover, according to ❯ Eq. 11, the pulse is the shorter the smaller the wavelength λ ($\lambda = h/(mv)$). Due to Bragg's law, $\lambda = 2d \sin\theta$, the lattice spacings (d spacings) in

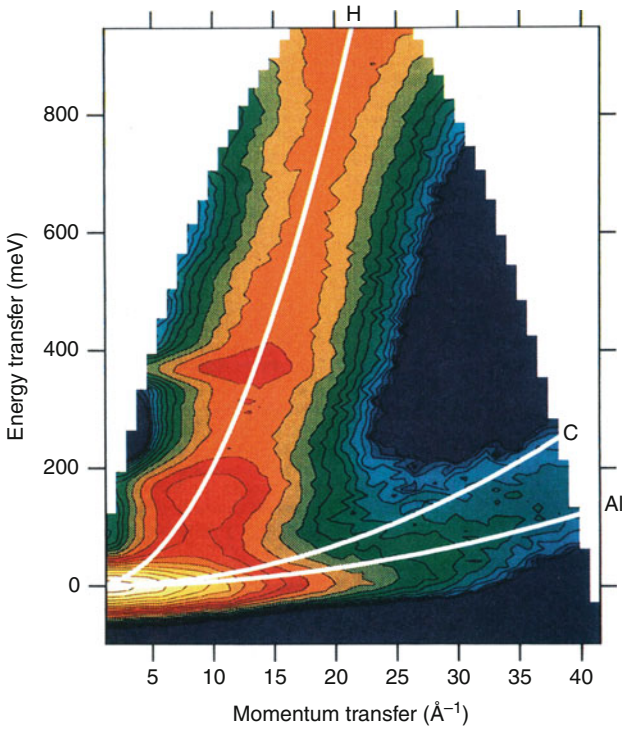


Fig. 13

Scattering law $S(Q, \omega)$ for coal measured at $T = 2$ K with incident neutron energy of 1 eV. Superimposed on inelastic peaks at 170 and 380 meV is the recoil line for mass 1 (H), as well as for masses 12 (C) and 24 (Al), respectively. Reprinted from Fillaux et al. (1995) with permission from Elsevier (2010)

Fig. 15 are proportional to the wavelength and therefore proportional to the flight time of the neutrons. The wealth of additional lines in *Fig. 15* (bottom frames) impressively shows the superiority of a pulsed spallation source like ISIS, the time-average neutron flux of which is only 7×10^{12} n/(cm^2 s) compared to a steady state reactor with a flux of 3×10^{14} n/(cm^2 s).

4.4 Neutron Powder Diffractometry in the History of Arts

If only minute quantities of certain substances are available or unique pieces of art such as single ancient coins, nondestructive methods of investigation are inevitable (see also Chaps. 33, "The Use of Neutron Technology in Archaeological and Cultural Heritage Research" and 34, "Radiation Detectors and Art" of the present Handbook). Below is shown an example of phase and texture analyses by highly efficient time-of-flight powder diffractometry. In this case the authenticity of historic coins had been the point of interest. In *Fig. 16* are shown texture polar diagrams as well as diffractograms of two seventeenth-century coins ("Ferdinand-Taler"), #205491, and #205492, of which the first could be dubbed as genuine and the second as fake.

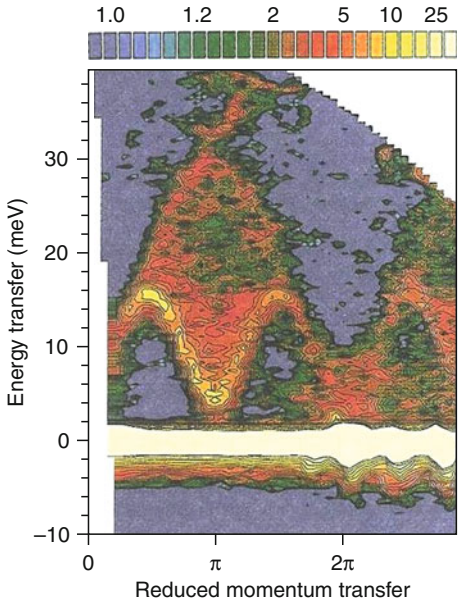


Fig. 14

Scattering from a $CuGeO_3$ single crystal at $T = 10$ K. Incident momentum $\underline{k}$ of the neutrons is perpendicular to the 1-dimensional, antiferromagnetic spin chain along the c-axis. It has to be noted that due to the low temperature the excitations with energy gain ($\hbar\omega < 0$) are missing. Reprinted from Arai et al. (1996) with permission by the American Physical Society (2010)

The verdict on the first relied on the content of silver (90%) and the observed typical texture obtained by the seventeenth-century production method of rolling. The fake one was identified by its fraction of silver (15%) and copper (85%), respectively, as well as the random crystallite distribution due to casting.

4.5 Neutron Radiography

Because of the low attenuation by most chemical elements, neutrons can advantageously be used to nondestructively investigate objects in various fields such as archaeology, arts, or technologically and industrially important items. The advantages include in particular the magnetic interactions between neutrons and magnetic moments in materials as well as the considerable variations in contrast between chemical elements and isotopes. For biological applications, radiation damage due to neutron irradiation is negligible, as compared to electrons or X-rays. Neutron radiography has been applied to authenticate paintings, examine artifacts made of metal or stone, or for quality control purposes in industries, which require precision machining such as aircraft engines. In all cases hidden structures are made visible, for instance, several layers of different paintings or fakes on the same canvas, or cracks in turbine blades. With the enormous peak intensity of pulsed sources and its inherent opportunity to perform real-time-resolved experiments, even the varying lubrication distributions inside running engines can

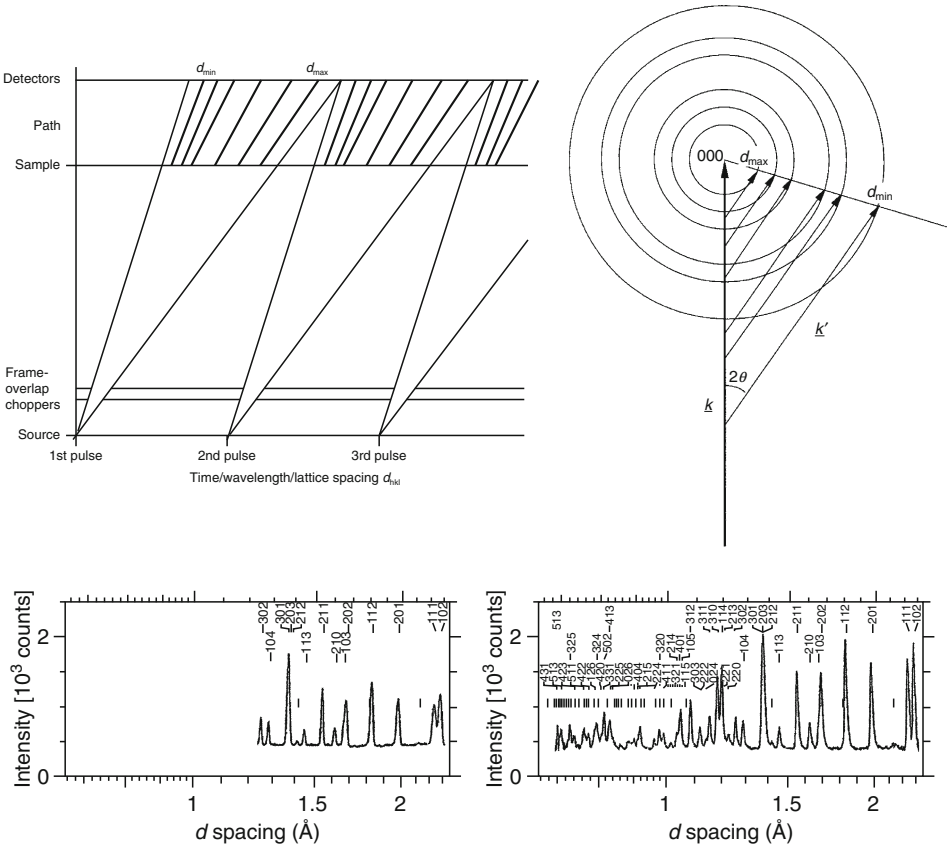


Fig. 15

Top left: **Pulsed powder diffraction as a path-versus-time diagram. The sample (polycrystal or powder) acts as a wavelength (*d* spacing) selective filter for the incident "white" pulse. The different reflections d_{hkl} appear at the detectors separated in time.** ***Top right:*** **Same as on the left, but in reciprocal-space representation with Debye–Scherrer rings and scattering triangles from a polycrystalline sample. The reflections appear at the same detector scattering angle 2θ, but at different times of arrival.** ***Bottom left:*** **Powder diffractogram of quartz, measured at the FRJ-2 reactor in Jülich and** ***bottom right:*** **At the spallation source ISIS in England. In both cases, the identical sample had been used with the same position-sensitive detector in identical distance from the sample and identical measuring time of 2 h. The wavelength extent at ISIS was 0.9–2.3 Å, the wavelength at the FRJ-2 was 1.09 Å (W. Schäfer, Mineralogisches Institut, University of Bonn, 1991, private communication)**

be studied. Besides other installations at reactors, a dedicated, powerful radiography station exists at the Swiss spallation source SINQ (Lehmann and Wagner 2010). In *Fig. 17* are shown examples of a 3-dimensional tomographic image as well as the lubricant distribution inside a motor. The contrast necessary to distinguish the latter is due to the exceptionally high scattering cross section of hydrogen compared to most other elements including all engineering metals. In neutron radiographic studies hydrogenous substances are therefore easily discernible.

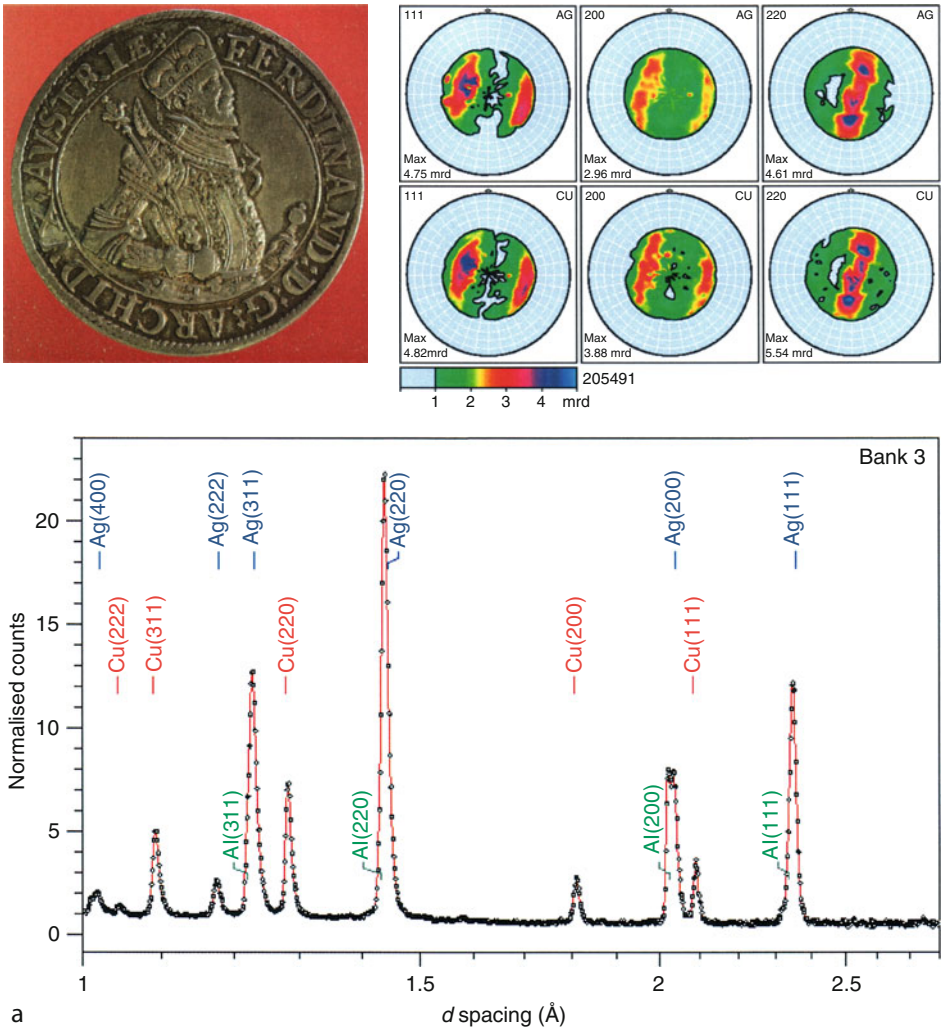


Fig. 16

Authenticity investigation of seventeenth-century coins from texture and composition determination by neutron powder diffraction: (a) genuine coin and (b) counterfeited coin (W. Schäfer, Bonn University, 2003, private communication)

5 Spallation: Accelerator-Driven Nuclear Energy

Accelerator-driven nuclear energy – most frequently cryptically called accelerator-driven systems (ADS) – summarizes ideas and concepts to apply accelerator-generated neutrons for advanced nuclear technologies. Interestingly, however, the idea even goes back to the early days of nuclear fission, when with an ADS it had been attempted to breed fissile materials such as ^{239}Pu (from ^{238}U) or ^{233}U (from ^{232}Th) in the Materials Testing Accelerator in the Lawrence Radiation Laboratory in Livermore (Lawrence 1954). Since in the 1940s and 1950s the intended

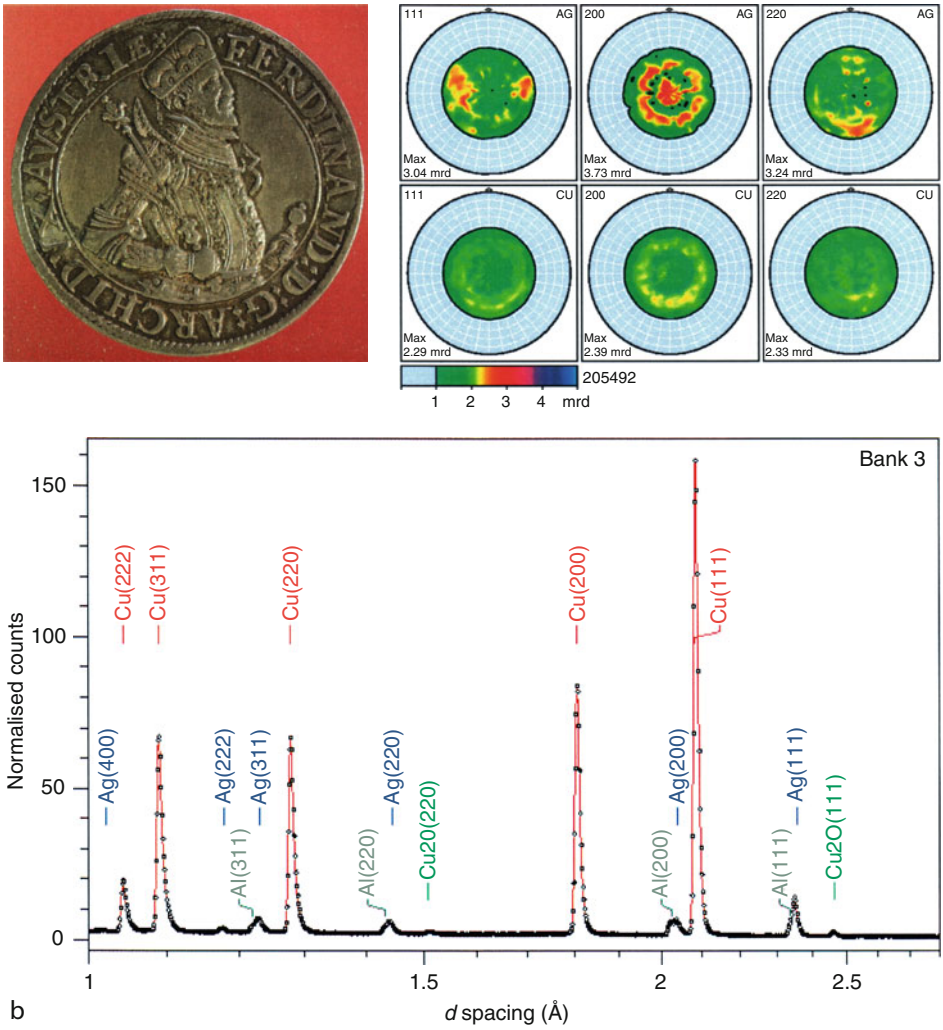


Fig. 16
(Continued)

breeding of fissile nuclei formed fertile ones as well as transmutation of actinides and fission products was found to be inefficient due to insufficient accelerators, ADS had not been pursued further for decades. With the advent in the 1990s of powerful accelerators in the MW range (cf. also Sect. 3.3) the interest in ADS revived. Meanwhile the envisaged applications comprise

- Subcritical nuclear power reactors
- Nuclear waste incinerator (transmutation of fission products, minor actinides, and plutonium from light water reactors (LWR))
- Disposal of weapons-grade plutonium
- Fertile-to-fissile conversion (thorium–uranium cycle)

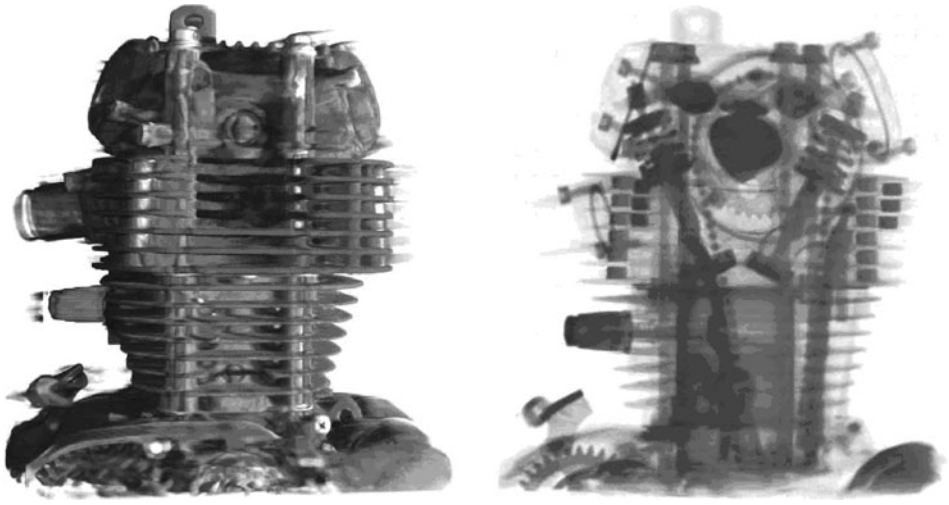


Fig. 17
Neutron radiographic image of an engine. *Left frame*: 3D tomographic image; *right frame*: Visualization of the lubricant distribution (*dark patches*) (E.H. Lehmann, PSI, 2010, private communication)

The heart of any ADS is the spallation target. If the target is surrounded by (1) a blanket assembly of nuclear fuel, such as fissile isotopes of uranium or plutonium, or by (2) ^{232}Th, which can convert to ^{233}U, or by (3) fissionable nuclear waste with added fuel, a fission chain reaction can be sustained even for a subcritical blanket (effective neutron multiplication factor $k_{\text{eff}} < 1$). In such a system, the neutrons produced by spallation would boost fission in the fuel or waste, assisted by further neutrons arising from that fission. A schematic layout of an ADS is presented in *Fig. 18*.

Up to 10% of the neutrons could come from the spallation, though it would normally be less, with the rest of the neutrons arising from fission events in the blanket assembly. An ADS is supposed to only run when neutrons are supplied to it, because it burns material that does not generate sufficiently many neutrons to maintain a fission chain reaction. Therefore, an ADS will be a nuclear reactor that could be turned off simply by shutting down the proton beam, rather than the necessity to insert control rods to absorb neutrons, which renders the fuel assembly subcritical. Since the reaction ceases when the accelerator current is switched off, ADS are considered safer than normal fission reactors.

5.1 ADS Research and Development: The Belgian MYRRHA Project

The only contemporary ADS simulation experiments started in 2009 at the Kyoto University Research Reactor Institute (KURRI) and on March 3, 2010, 100 MeV protons from a Fixed Field Alternating Gradient (FFAG) accelerator were injected for the first time into a tungsten target inside a thorium blanket (www.rri.kyoto-u.ac.jp/press/2010/Press_Release_20100312.pdf), but no major installation in the multi-MW scale has been realized so far. On the other hand,

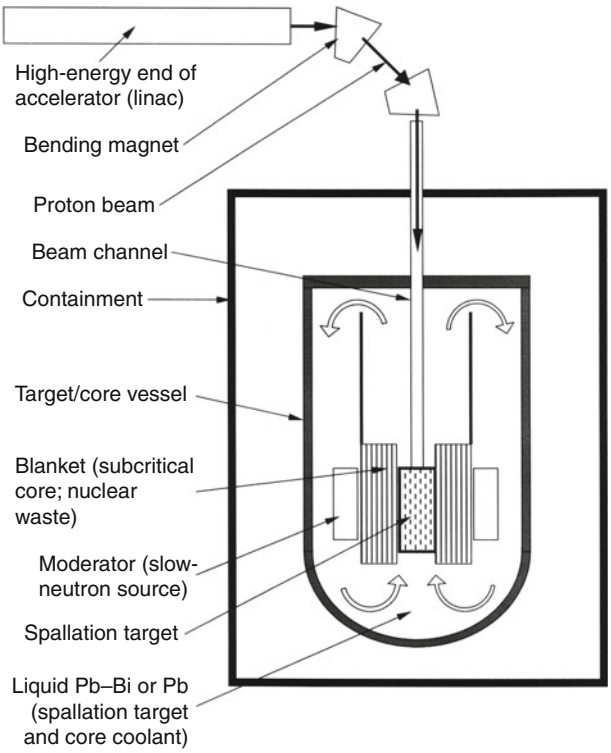


Fig. 18

Schematic layout of an ADS. If a thermal-spectrum system is considered (Bowman 1998), the moderator will be placed between target and blanket or be an integral part of the blanket

the Belgian Nuclear Research Centre (SCK.CEN) at Mol is planning the construction of an ADS research reactor called MYRRHA (Multipurpose Hybrid Research Reactor for High-tech Applications) (Nuclear power in Belgium, http://myrrha.sckcen.be/ and http://world-nuclear.org/info/inf94.html). It is pointed out by SCK.CEN that MYRRHA is not intended as an industrial waste incinerator, but to serve as a multipurpose irradiation facility for research. The first five years (2010–2014) of the project have to achieve what is called the Front End Engineering Design (FEED), which includes both the completion of supporting R&D to alleviate technical hurdles and the necessary studies to secure the licensing of MYRRHA.

Initially, MYRRHA will be a 57 MWth ADS, consisting of a proton accelerator delivering a 600 MeV, 2.5 mA proton beam to a liquid lead–bismuth (Pb–Bi) spallation target that in turn couples to a Pb–Bi-cooled, subcritical fast nuclear core (cf. *Fig. 18*). The technical details of MYRRHA are as follows.

5.1.1 The Accelerator

The number of accelerator types to choose from is limited by the need to optimize the use of existing technology and reduce the cost as well as maximize the reliability of operation as an

Table 6
Main parameters of MYRRHA

Proton beam energy	600 MeV
Accelerator current	2.5 mA
Proton beam heating	0.85 MW
Spallation neutron yield	15.3 n/p
Neutron source intensity	2.23×10^{17} n/s
Initial fuel mixture	$(U/Pu)O_2$ MOX
Initial (HM) fuel mass	680 kg
Initial plutonium enrichment Pu/(U + Pu)	35 wt%
Initial plutonium isotope vector (238/239/240/241/242)	2.33/56.87/27.00/6.10/7.69 wt%
Effective neutron multiplication factor k_{eff}	0.95
Source k factor k_s	0.96
Gain factor MF = $1/(1 - k_s)$	25.04
Source importance (source multiplication in a subcritical blanket)	1.10
Thermal power	57 MW
Specific power	66 kW/kgHM
Peak linear power (hottest pin)	253 W/cm
Average linear power (hottest pin)	146 W/cm
Maximum total flux in fast core	3.3×10^{15} n/(cm^2 s)
Maximum fast (> 1 MeV) flux in fast core	0.5×10^{15} n/(cm^2 s)
Maximum fast (> 0.75 MeV) flux in fast core	0.7×10^{15} n/(cm^2 s)

important issue. The main challenging research topic for the accelerator as part of an ADS is the improvement of the beam reliability, since an ADS is intended to be operated in a continuous way. The reliability and availability requirements are the main reasons why a linac instead of a cyclotron has been chosen for MYRRHA. The selected linac parameters are 600 MeV at 2.5 mA yielding a beam power of 1.5 MW (*Table 6*).

5.1.2 The Target

Due to the low range of 600 MeV protons in heavy-metal targets and the corresponding high power density therein, the Pb–Bi eutectic (melting point 123 °C) has been chosen as a target material. The molten eutectic target will be part of a loop, which also serves as the primary coolant of the subcritical core. In order to avoid radiation damage to the proton beam window, it had been decided to rely on a windowless solution with corresponding beam injection from above. The target will be placed in the center of the blanket, i.e., the subcritical core (cf. *Fig. 18*).

5.1.3 The Subcritical Core

The subcritical core (57 MW thermal power) is composed of MOX (mixed oxide; U/PuO_2) fuel with a plutonium content limited to 35 wt%, cooled with liquid Pb–Bi and an effective multiplication factor k_{eff} = 0.95. The core will act as a neutron multiplier of the primary

neutrons from the spallation source. The energy spectrum of the subcritical core depends on the intended application of the core. Nevertheless, a fast-neutron spectrum presents various advantages, among them are a larger excess of neutrons, which can be used for minor actinide transmutation, a reduced amount of minor actinide production in the core itself, and a better energetic yield for future energy production (see Sect. 5.2). For MYRRHA, it is envisaged to couple a fast-spectrum zone (where minor actinide transmutation studies, structural materials research, and ADS fuel studies can be performed) to a thermal spectrum region (cf. *Fig. 18* and Sect. 5.3), where radioisotope production, long-lived fission product transmutation research as well as light water reactor (LWR) fuel safety studies can be conducted.

5.2 Energy Amplifier

In 1993, Carlo Rubbia caught on to Lawrence's idea of the ADS (Lawrence 1954), because recent major advances in accelerator technology seemed to render an accelerator-driven power reactor a viable option. The so-called energy amplifier (EA) is in fact a nuclear power reactor with a subcritical core, whereby it is supposed to constitute an inherently safe design (Fernández et al. 1996). Due to the subcriticality, the core does not sustain the chain reaction, wherefore an external neutron source is needed. This external source had been immediately identified as to be based on spallation. The basic idea is that an energetic proton beam is used to boost the reaction, which in turn releases enough energy to power the particle accelerator and leaves an energy surplus for power generation ("energy amplifier"). The main features of the energy amplifier concept shall be briefly explained. For more details, the reader is referred to Fernández et al. (1996). The schematic layout is similar to that presented in *Fig. 18*.

The plant is designed to be driven by two cyclotrons in series with a final proton energy of 1 GeV and a nominal current of 12.5 mA. Due to changes of the subcriticality of the core, the two accelerators must be able to produce up to 20 mA. The subcritical core will consist of thorium fuel, which initially contains plutonium from spent light water reactor fuel as fissile material. The system is designed for an average effective neutron multiplication factor $0.95 < k_{\text{eff}} < 0.98$. Assuming an electric efficiency of the accelerator of 40% (Fernández et al. 1996), 31.25 MW electric power from the mains are required to obtain 12.5 MW nominal proton beam power. A nominal accelerator input power of 31.25 MW is predicted to produce a thermal power of the EA of 1,500 MW. In other words, if owing to the high operating temperature of the EA (550–600 °C) a 45% conversion efficiency (yielding 675 MWe) is assumed, about 5% of the produced electricity will have to be fed back to the accelerator.

In order to maximize the fission probability of the actinide inventory and to minimize the losses due to parasitic capture of thermal neutrons in fission products (cf. Sect. 5.3), a fast-neutron spectrum is chosen for the EA. Therefore, lead is chosen both as a primary coolant (no moderation!) and as spallation target for the proton beam (no separate target!). The heat produced both in the target and core will be removed by natural convection of the primary coolant. Therefore the main vessel was originally designed with a height of 30 m (later reduced to 15 m) in order to ensure a stable lead flow.

No fuel reloading or reassembling is planned for the entire lifetime of the fuel, which is estimated to be 100–150 GWd/t (approximately 5 years). Eventually the fuel will be reprocessed, the whole actinide inventory being restored into the EA, with only fission products destined for final disposal.

Variations of k_{eff} during burn-up, which will result in variations of the produced thermal power, are compensated by adjusting the proton current delivered to the target. Therefore no control rods are planned to be used in the EA.

Nevertheless, three different passive safety systems are planned to be installed. They will be triggered, if an exceptional temperature rise leads to an expansion of the coolant (lead), which will then flow through spillways into special regions of the reactor. The first system is an Emergency Beam Dump Volume (EBDV) to remotely dump the proton beam in case of an accelerator shutdown failure. The second system is the Reactor Vessel Air Cooling System (RVACS), which consists of a narrow gap between the containment and the main vessel, normally filled with thermally insulating helium gas. If this gap is filled with lead, heat from the inner core is removed via natural convection of air to ensure the decay heat removal. The third system is a scram device consisting of B_4C absorber rods, which will be pushed from underneath into the core by buoyancy forces, which are exerted by overflowing coolant lead.

5.3 Nuclear Waste Incineration

A significant part of the wastes contained in used nuclear fuel from LWRs is fission products as well as plutonium and long-lived minor actinides (particularly neptunium, americium, and curium). In recent years, interest has grown in the possibility of separating (or partitioning) the long-lived radioactive waste from the used fuel and transmuting it into shorter-lived radionuclides so that the management and eventual disposal of this waste is easier and less expensive. Since the 1990s two basically different types of waste incinerators have been proposed, viz., fast-spectrum systems (Rubbia et al. 1997) and thermal-spectrum systems (Bowman et al. 1992).

The principal design features of a fast-spectrum ADS are those of the energy amplifier (Sect. 5.2) and depicted in *Fig. 18*, but without moderators. The fast-spectrum ADS proposals feature either oxide-fueled and sodium-cooled systems or solid metallic-fueled and heavy-metal-cooled systems. The interest in fast-spectrum systems stems from the many years of effort in nearly all nuclear countries to develop fast breeder reactors. The principal design advantage of the fast-spectrum system is its favorable neutron economy and the capability to induce fission in all of the actinides. The disadvantage of the fast-spectrum ADS is the long time to reach equilibrium of LWR waste generation and ADS incineration as well as requiring a large waste-actinide inventory (Bowman 1998).

The thermal-spectrum neutron economy is less favorable than that of the fast spectrum. Parasitic capture of thermal neutrons by structural and fuel transport materials have to be minimized, wherefore elements like ^{7}Li (enriched), Be, and Zr have to be mainly utilized. Moreover, the necessary reduction of fission product absorption is one of the primary drivers toward liquid-fuel systems with the capability for online removal of fission products by rapid cycling without fuel destruction and re-fabrication. The disadvantage of rapid cycling is the large amount of material which has to be processed and the high-performance separations required (Bowman 1998). Nevertheless, the small inventory feature of the thermal-spectrum system, made possible by the large fission and capture cross sections of plutonium and minor actinides, is highly attractive for waste destruction.

The design of a thermal-spectrum ADS incinerator features a graphite-moderated liquid-fueled assembly with $k_{\text{eff}} = 0.96$ (Bowman 1998). The actinides and fission products, chemically transformed into fluorides, continuously flow through the blanket (cf. *Fig. 18*) via the carrier salt $NaF–ZrF_4$. The salt flows upward through holes in the graphite, then across the top to the outside of the system and down through heat exchangers and back into the graphite moderator.

A salt-to-graphite volume ratio of 1/20 assures a well-thermalized spectrum. A commercial 750 MWth plant will be capable of transmuting the waste from one 3,000 MWth LWR at the rate it is produced (Bowman 1998).

6 Conclusions

As far as neutron scattering is concerned, it can clearly be stated that the future of neutron sources is based on accelerators, in particular in pulsed operation mode. The big leap forward for neutron scattering with pulsed spallation sources is obvious from the following comparison as well as from *Fig. 19*. Moreover, continuous spallation sources seem to open new fields in both nuclear waste transmutation and future nuclear power plants.

Advantages of spallation neutrons

- More neutrons per reaction
- Pulsed mode easily feasible with accelerators
- Peak flux/average flux > 100 possible
- eV neutrons for neutron scattering available
- Two or more target stations with different pulse structures possible with a single accelerator
- Proton pulse lengths between microseconds and milliseconds tunable (ESS-LP long-pulse source)
- Only 10% waste heat per neutron

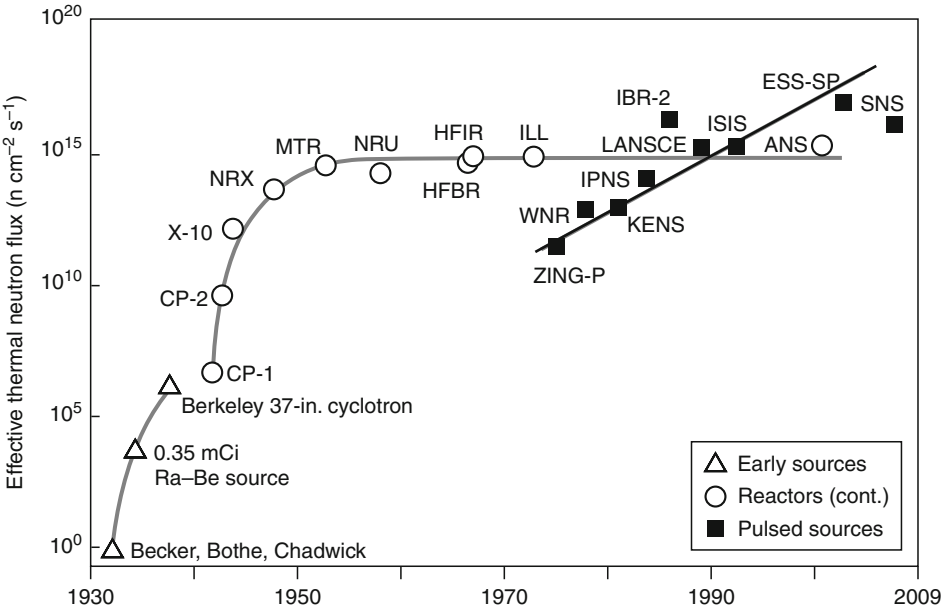


Fig. 19

The development of effective fluxes (i.e., peak fluxes for pulsed sources) of neutron sources since the discovery of the neutron in 1932 (visualization based on Brugger (1968))

- No critical assembly (no fissile elements, if ^{238}U not used; no actinides produced)
- Only 5% of saturation activity per neutron (little radioactive waste) (ESS 2002)
- Future nuclear power plants without a critical core
- Nuclear waste transmutation possible

Disadvantages and problems (for neutron scattering applications):

- Due to sophisticated accelerators more complex than a reactor but reliably functioning
- Lifetime limits of liquid-target vessels due to cavitation caused by shock waves in short-pulse operation (mitigation possible by gas injection into the liquid); no similar problems with solid targets
- Radiation damage by high-energetic protons in solid targets and target structural materials for liquid targets; no problem with windowless liquid targets and beam injection from above

7 Cross-References

Chapter 1, "Interactions of Particles and Radiation with Matter"
Chapter 7, "Accelerators for Particle Physics"
Chapter 22, "Radiation Damage Effects"
Chapter 33, "The Use of Neutron Technology in Archaeological and Cultural Heritage Research"
Chapter 34, "Radiation Detectors and Art"

References

Arai M, Fujita M, Motokawa M, Akimitsu J, Bennington SM (1996) Quantum Spin Excitations in the Spin-Peierls System $CuGeO_3$. Phys Rev Lett 77:3649

Ashmore A, Cocconi G, Diddens AN, Wetherell AM (1960) Total cross sections of protons with momentum between 10 and 28 GeV/c. Phys Rev Lett 5:576

Bartholomew GA, Tunnicliffe PR (eds) (1966) The AECL study for an intense neutron generator, AECL 2600. Atomic Energy of Canada

Bauer GS, Conrad H, Spitzer H, Friedrich K, Milleret G (1981) Measurement of time structure and thermal neutron spectra for various target-moderator-reflector configurations of a spallation neutron source. In: ICANS-V: Proceedings of the 5th meeting of the international collaboration on advanced neutron sources, Jül-Conf-45, p 475

Bauer GS, Conrad H, Grünhagen K, Spitzer H, Milleret G (1983) How much thermal neutron flux is gained using deuterons instead of protons? Proceedings of the 6th meeting of the international collaboration on advanced neutron sources, ANL-82-80, p 619

Beckurts KH, Wirtz K (1964) Neutron physics. Springer, Berlin

Bowman CD (1998) Accelerator-driven systems for nuclear waste transmutation. Annu Rev Nucl Part Sci 48:505

Bowman CD et al (1992) Nuclear energy generation and waste transmutation using an accelerator-driven intense thermal neutron source. Nucl Instrum Methods A320:336

Brugger RM (1968) We need more intense thermal-neutron beams. Phys Today 21(12):23

Carpenter JM (1977) Pulsed spallation neutron sources for slow neutron scattering. Nucl Instrum Methods 145:91

Conrad H, Brückel T, Schäfer W, Voigt J (2008) POWTEX – the high intensity time-of-flight diffractometer at FRM II for structure analysis of polycrystalline materials. J Appl Cryst 41:836

Crandall WE, Millburn GP (1958) Neutron production at high energies. J Appl Phys 29: 698

ESS (2002) The ESS Project, vol III. Technical report, chapter 4.7, ISBN 3-89336-303-3

Fernández R, Mandrillon P, Rubbia C, Rubio JA (1996) A preliminary estimate of the economic impact of the energy amplifier. CERN/LHC/96-01 (EET)

Filges D, Neef RD, Schaal H (1995) Nuclear studies of different target systems for the European Spallation Source (ESS), ICANS-XIII. Report PSI-Proc. 95–02, p 537

Fillaux F, Papoular R, Bennington SM, Tomkinson J (1995) Inelastic neutron scattering study of free proton dynamics in coal. J Non-Cryst Solids 188:161

Jackson JD (1963) Classical electrodynamics. Wiley, New York

Lawrence EO (1954) Status of the MTA process. Laboratory report LRL-102

Lehmann EH, Wagner W (2010) Neutron imaging at PSI: a promising tool in materials science and technology. Appl Phys A 99:627

Prael RE, Lichtenstein H (1989) The LAHET code system, LA-VR-89-3014. Los Alamos National Laboratory

Pynn R et al (1993) Next generation spallation neutron source, vol 2. Technical review LA-UR-93-4440. Los Alamos National Laboratory

Rubbia C, Buono S, Kadi Y, Rubio JA (1997) Fast neutron incineration in the energy amplifier as alternative to geologic storage: the case of Spain. CERN/LHC/97-01 (EET)

Further Reading

Cierjacks S (ed) (1983) Neutron sources for basic physics and applications. Pergamon Press, New York

Filges D, Goldenbaum F (2009) Handbook of spallation research. Wiley-VCH, Weinheim

Windsor C (1981) Pulsed neutron scattering. Taylor and Francis, London

31 Neutron Detection

Alfred Klett
Berthold Technologies GmbH & Co KG, Bad Wildbad, Germany

1 Introduction 761

2 Fundamental Neutron Physics 762
2.1 The Neutron 762
2.2 Basic Neutron Interactions 762
2.3 Neutron Generation 765
2.4 Neutron Moderation 765
2.5 Neutron Absorption and Shielding 766
2.6 Metrology and Dosimetric Quantities 766

3 Materials and Detector Types for Neutron Detection 768
3.1 Neutron Detection Principles 768
3.2 Active Neutron Detection Methods 768
3.2.1 Gas-Filled Detectors 768
3.2.2 Semiconductors 770
3.2.3 Scintillators 770
3.2.4 Superheated Emulsion Detectors 771
3.3 Passive Neutron Detection Methods 771
3.3.1 Track Detectors 771
3.3.2 Thermoluminescent Dosimeters 771
3.3.3 Etched-Track Detectors 772
3.3.4 Passive Superheated Emulsion Detectors 772
3.3.5 Direct Ion Storage 772
3.3.6 Other Passive Detectors 773

4 Applications of Neutron Detection 773
4.1 Neutron Dose Measurement 773
4.1.1 Introduction 773
4.1.2 Rem Counters 773
4.1.3 Tissue-Equivalent Proportional Counters 775
4.1.4 Active Personal Dosimeters 776
4.1.5 Passive Dose Measurement 776
4.1.6 Dose Measurement in Pulsed Radiation Fields 776
4.1.7 Examples of Neutron Dose Measurements 777
4.2 Spectrometry 778
4.2.1 General 778

C. Grupen, I. Buvat (eds.), *Handbook of Particle Detection and Imaging*,
DOI 10.1007/978-3-642-13271-1_31, © Springer-Verlag Berlin Heidelberg 2012

4.2.2 Bonner Spheres . . . 778
4.2.3 Time-of-Flight Spectroscopy . . . 779
4.2.4 Recoil Spectroscopy . . . 780
4.3 Neutron Activation Analysis . . . 781
4.4 Neutron Scattering . . . 781
4.5 Nuclear Medicine . . . 781
4.6 Search for Illicit Trafficking Nuclear Materials . . . 782
4.7 Reactor Instrumentation . . . 783
4.8 Fusion Monitoring . . . 783
4.9 Industrial Applications . . . 784
4.9.1 Neutron Imaging and Radiography . . . 784
4.9.2 Humidity Measurement . . . 784

5 Reference Neutron Radiation Fields . . . 784

6 Conclusion . . . 786

Acknowledgment . . . 786

References . . . 786

Further Reading . . . 789

Suppliers of Neutron Detectors . . . 789

Abstract: Neutrons are electrically neutral particles and therefore they are mainly subject to hadronic but not to electric forces. As neutrons are not directly ionizing they have usually to be converted into charged particles before they can be detected. The basic physical principles for neutron detection are the neutron's characteristic properties and several important nuclear reactions and processes.

The most important active neutron detector types are gas-filled, scintillation, and semiconducting detectors and the most important passive neutron detector types are thermoluminescent, etched-track, and nuclear-emulsion detectors. Specials techniques like the superheated emulsion detectors are unique in neutron detection.

Neutron detection covers a wide variety of applications in nuclear physics, in neutron scattering for biological, chemical, medical, and material analysis research, in metrology, in radiation protection, in nuclear energy and in the nuclear fuel cycle, in reactor instrumentation, in nuclear decommissioning and nuclear waste, in homeland security, in safeguards, in fusion monitoring, and in industrial measurements. Several important concepts and techniques as ^{3}He proportional counters, rem counters, Bonner sphere spectrometers, tissue-equivalent proportional counters, time-of-flight measurement, and neutron activation analysis are described and discussed.

A few examples and results of neutron detection in aircrew dosimetry at flight altitudes, measurements in accelerator environments, and industrial measurements illustrate the diversity of neutron detection applications. As neutron detection and measurement requires calibration facilities and procedures, neutron reference fields are also discussed.

1 Introduction

Neutron detection was earlier only a small niche in radiation detection but since about one decade its importance has considerably grown. Several fields dealing directly or indirectly with neutrons have been newly developed or extended. There is now a large number of new research facilities like synchrotron radiation sources, free electron lasers, fusion test facilities, neutron spallation sources, and other neutron sources. For medical applications in many countries the installation of hadron therapy facilities with proton or heavy-ion accelerators has been started. Nuclear decommissioning and nuclear waste is growing, intermediate storage facilities have been commissioned and spent-fuel transports require more neutron measurements. There is also more environmental monitoring and for instance an increased awareness about exposures at aircraft altitudes. Nuclear safety issues have triggered a large amount of homeland security activities and neutron detection techniques are utilized in the search for illicit trafficking nuclear materials in border control, in gate monitoring, and in the detection of special nuclear materials. Thus, recent developments and new technologies together with sometimes even reanimated classical nuclear knowledge are making neutron detection an exciting field.

2 Fundamental Neutron Physics

2.1 The Neutron

The neutron was discovered relatively late in 1932 by Chadwick, as neutron detection is not very easy (Chadwick 1932a, b). The neutron has total electric charge zero, while the proton carries one positive elementary unit of electric charge. As a consequence, the neutron is not subject to electrical force. Neutrons are primarily interacting by hadronic interactions with other particles. The hadronic interactions are much stronger than electromagnetic interactions, but they have only a short range of a few fermi. Neutrons are as neutral particles not directly ionizing when traversing through matter. They are only ionizing indirectly by generating charged particles or photons in hadronic scattering process or reactions. In the quark model the neutron is just like the proton described as a composition of three quarks glued together with gluons. Both nucleons are extended particles with electrical charge density distributions which can be characterized by mean electric and magnetic radii. The mass of the neutron is slightly exceeding the proton's mass. Both nucleons carry spin, an internal angular momentum (see *Table 1*). The rotating charge density distributions are generating the neutron's and the proton's magnetic momenta. The neutron exhibits in the absence of electrical forces the weaker magnetic interaction and is therefore also useful as a magnetic probe.

Neutrons are remarkable objects. They are stable as bound particles in nuclei, but they are unstable as free particles. They have a beta decay mode with a half-life of approximately 10 min. As neutrons are not directly ionizing, their detection is more complicated than charged-particle or photon detection. They have usually first to be converted into charged particles. Effective conversion via nuclear reactions requires in many cases slowing down of fast neutrons, because the cross sections are only sufficiently high at low neutron energies.

The particular interactions with other particles and the huge energy range over more than 15 orders of magnitude from ultracold neutrons below meV energies up to exceeding the TeV domain at large accelerators or in cosmic radiation are making neutron physics so diversified and cause a variety of phenomena and different detection techniques. In *Table 2*, there are a few common designations of neutron energy ranges. Neutrons are conventionally called slow neutrons if they are below the so-called cadmium cutoff energy at 0.5 eV and fast neutrons if they are above this value. In the terminology of the ICRP and the ICRU neutrons of all energies are considered to be strongly penetrating radiation. Thermal neutrons are very important, because there are several efficient nuclear reactions which allow for thermal-neutron detection and because thermal neutrons induce some of the frequently used fission processes.

2.2 Basic Neutron Interactions

The interactions of neutrons with other nuclei (Evans 1982; Wirtz and Beckurts 1964) can be grouped into the following processes:

- Elastic scattering
- Inelastic scattering
- Radiative neutron capture
- Other nuclear reactions
- Fission (inclusive of spontaneous fission)
- Spallation

Table 1

Physical data neutron–proton (Amsler et al. 2008)

Designation	Neutron	Proton
Mass	939.5653 MeV	938.2720 MeV
Atomic mass unit [u]	1.008664915	1.007276466
Electric charge	0	+1
Spin	½	½
Magnetic moment	$-1.913042\ \mu_N$	$+2.792775\ \mu_N$
Mean lifetime τ	885.7 s	$> 2.1 \times 10^{29}$ years
Mean electric charge radius	0.12 fm	0.88 fm

Table 2

Neutron energy ranges

Designation	Neutron energy	Wavelength λ [pm]
Ultracold neutrons	< 0.2 meV	>2,000
Cold neutrons	0.2 meV–2 meV	640–2,000
Thermal neutrons	2 meV–100 meV	90–640
Epithermal neutrons	100 meV–1 eV	28–90
Intermediate neutrons	1 eV–10 keV	
Fast neutrons	10 keV–20 MeV	
High-energy neutrons	> 20 MeV	

Table 3

Important-thermal-neutron-induced nuclear processes (Knoll 2010)

Nuclear reaction	*Q*-value [MeV]	Cross section [barn]	Natural abundance
$^{3}He\ (n, p)^{3}H$	0.764	5,330	Not applicable
$^{6}Li\ (n, \alpha)^{3}H$	4.780	940	7.4%
$^{10}B\ (n, \alpha)^{7}Li$	2.792	3,840	19.8%
$Cd(n,\gamma)$		2,450	Natural Cd
$^{157}Gd(n, \gamma)^{158}Gd$		255,000	15.7%
^{235}U fission	≈210	582	0.72%
$^{197}Au(n, \gamma)^{198}Au$		98.65	100%

These processes are not only the base for most neutron detection mechanisms but also for neutron shielding and for the generation of any radiation hazard in bio-systems. Besides elastic and inelastic scattering neutron capture and several neutron-induced nuclear reactions are of particular importance in neutron detection. The most important nuclear processes with thermal neutrons are listed in ❯ *Table 3*.

Because of the extremely short range of the hadronic forces, neutrons have to come very close within $\approx 10^{-15}$ m of a nucleus before any interaction can take place. For a neutron in normal matter, there is a lot of empty space and therefore interactions have relatively low probability

Table 4

Examples: technical data of scintillators for neutron detection

Scintillator	^{6}LiI(Eu)	^{6}Li glass	Scintillating fiber
Density [g/cm^3]	4.1	2.48–2.67	
Index of refraction	1.96	1.55–1.58	
λ_{max} [nm]	470	395	
Decay time [ns]	1,400	75–100	150
Manufacturer	SCIONIX	St. Gobain	NUCSAFE Inc.

Table 5

Materials for neutron threshold activation (Wirtz and Beckurts 1964; Knoll 2010; IAEA 1974)

Material	Reaction	Isotopic abundance [%]	Half-life [min]	γ Energy [MeV]	Yield [%]	Thresh [MeV]
Mg	^{24}Mg(n, p)^{24}Na	78.7	900	1.368	100	6.0
Al	^{27}Al(n, α)^{24}Na	100	900	1.368	100	4.9
Al	^{27}Al(n, p)^{27}Mg	100	9.46	0.84–1.01	100	3.8
Fe	^{56}Fe(n, p)^{56}Mn	91.7	153.6	0.84	99	4.9
Co	^{59}Co(n, α)^{56}Mn	100	153.6	0.84	99	5.2
Ni	^{58}Ni(n,2n)^{57}Ni	67.9	2,160	1.37	86	13.0
Ni	^{58}Ni(n, p)^{58}Co	67.9	103,104	0.81	99	1.9
Cu	^{63}Cu(n,2n)^{62}Cu	69.1	9.8	0.511	195	11.9
Cu	^{65}Cu(n,2n)^{64}Cu	30.9	762	0.511	37.8	11.9
Au	^{197}Au(n,2n)^{196}Au	100	9,792	0.33–0.35	25–94	8.6

Table 6

Materials for slow-neutron activation (Wirtz and Beckurts 1964; Knoll 2010; IAEA 1974)

Material	Reaction	Isotopic abundance [%]	Half-life [min]	Thermal cross section [barn]
Manganese	^{55}Mn (n, γ)^{56}Mn	100	154.8	13.2
Cobalt	^{59}Co (n, γ)^{60m}Co	100	10.4	16.9
Copper	^{63}Cu (n, γ)^{64}Cu	69.1	772.2	4.41
Copper	^{65}Cu (n, γ)^{66}Cu	30.9	5.14	1.8
Silver	^{107}Ag (n, γ)^{108}Ag	51.82	2.4	38.6
Silver	^{109}Ag(n, γ)^{110}Ag	48.18	0.42	90.5
Indium	^{113}In(n, γ)^{114m}In	4.23	70,560	56
Gold	^{197}Au(n, γ)^{198}Au	100	3,881	98.65

and the neutron is a very penetrating particle. The probabilities of neutron interactions are characterized by their cross sections and are usually strongly depending on the neutron energy. The Q-value is a measure for the energy released to the reaction products. The higher the Q-value the easier is detection and discrimination against gamma radiation. The most important process for fast-neutron detection and also for neutron moderation is elastic scattering on light target nuclei especially elastic n–p scattering.

Spallation is an inelastic interaction of a projectile, for instance a proton or a neutron with high kinetic energy exceeding 100 MeV with a heavy nucleus. In the first fast stage, the projectile interacts with individual nucleons of the target nucleus and several nucleons are leaving the nucleus with high energies preferably in the forward direction. In the second slower stage, the energy in the residual nucleus is distributed across the other nucleons and neutrons and other particles at a typical energy scale of several MeV are evaporated with isotropic angular distributions. Spallation is used for neutron generation in spallation sources and is also used for the detection of high-energy neutrons. Spallation target materials are for instance tungsten or lead.

2.3 Neutron Generation

The main physical processes for neutron generation are fission, fusion, and nuclear reactions. The most important neutron sources or neutron-generating facilities are as follows:

- Reactors
- Accelerators
- (α–n) radionuclide sources like ^{241}Am–Be, ^{239}Pu–Be, ^{238}Pu–Be, and ^{226}Ra–Be
- Spontaneous-fission radionuclide sources like ^{252}Cf
- Plasma neutron generators
- Fusion facilities
- Nuclear weapons

Reactors are very common as neutron sources and can deliver very high intensities. As neutrons cannot be accelerated directly by accelerators, they are generated by bombardment of appropriate target materials with charged projectiles. Nuclear (α–n) reactions are also important in neutron production. Especially if alpha particles are hitting ^{9}Be nuclei there is a large probability for the generation of a neutron and ^{12}C. The mixture of an alpha-emitting radionuclide with beryllium is therefore an excellent neutron source. Fusion processes are also efficient in neutron generation. Relatively new are plasma neutron generators which utilize the d–d or the d–t fusion reaction in a gas discharge tube. The former generates neutrons at energies around 2.5 MeV and the latter at about 14 MeV. Details of several of these processes and reactions are summarized below in Sect. 5. Another neutron source is spent nuclear fuel. There are also significant amounts of neutrons in the secondary cosmic radiation in the atmosphere.

2.4 Neutron Moderation

As there are no electric forces acting between neutrons and matter, energy can only be transferred from neutrons to other particles by hadronic interactions. As a consequence of the conservation of energy and momentum, the maximum energy transfer in elastic collisions of projectiles with target nuclei depends on the particles' masses. The maximum energy transfer in elastic neutron scattering occurs in collisions with protons or other light nuclei. This is efficient up to the ^{12}C nucleus. In order to transfer kinetic energy from neutrons to other particles elastic neutron scattering on heavier target nuclei is relatively inefficient. Multiple elastic neutron scattering in a material containing light nuclei reduces the kinetic energy of incoming neutrons considerably and is called moderation.

If moderated neutrons are in thermal equilibrium with the surrounding materials they are called thermal neutrons. The mean kinetic energy of thermal neutrons is about 0.025 eV. Neutron moderation is important because thermal neutrons can easily be detected or absorbed and because thermal neutrons can induce some of the important fission processes. Good moderators are materials with a large amount of hydrogen, for instance polyethylene or water. Deuterium is also a suitable moderator because it has a lower neutron absorption cross section than hydrogen. Of course the number of collisions required for thermalization is dependent on neutron energy. About 20 interactions are sufficient to thermalize a 1 MeV neutron in hydrogen. Beckurts and Wirtz (1964) gave a comprehensive description of neutron moderation.

2.5 Neutron Absorption and Shielding

A neutron absorber is a material with which neutrons interact significantly by nuclear reactions resulting in their disappearance as free particles. As direct absorption of fast neutrons has low probabilities efficient neutron absorbers are only existing for thermal neutrons. Therefore neutron shielding is usually a combination of moderation and successive absorption of thermal neutrons. The ^{10}B reaction or absorption on natural cadmium, which is a mixture of several isotopes, is providing excellent thermal-neutron shielding. Cadmium of a few millimeters thickness is absorbing basically all neutrons below a cutoff energy of about 0.5 eV. The cross section of ^{10}B is decreasing reciprocally to neutron velocity and boron is strongly absorbing slow neutrons. Other possible reactions are listed in ❯ *Table 3*.

Highly efficient neutron shielding for fast neutrons is either moderator material followed by layers of slow-neutron absorber or a mixture of moderating material with slow-neutron-absorbing material. Boronated polyethylene is a very useful neutron shielding material. Boron–silicone is a heat- and fire-resistant elastomer, which can be used as castable neutron shielding. Polycast is a dry mix material designed to be cast into closed containers. It is field castable, providing excellent, low-cost neutron shielding, with a hydrogen content 6% greater than that of water. Neutron putty is a nonhardening boron-loaded putty with a high hydrogen content. Neutron shielding is available as sheets, plates, rods, or pellets. Other neutron shielding materials in use are water, concrete, soil, and steel.

2.6 Metrology and Dosimetric Quantities

Radiation measurements and investigations of radiation effects require the definition of radiometric quantities (ICRU 1998a). Radiation fields are characterized by radiometric quantities which apply in free space as well as in matter. The particle number N is the number of particles that are emitted, transferred, or received. The flux is the quotient of $\mathrm{d}N/\mathrm{d}t$ where $\mathrm{d}N$ is the increment of the particle number in the time interval $\mathrm{d}t$. One of the most important quantities in neutron detection is the fluence ϕ, which is the ratio of the number $\mathrm{d}N$ of particles incident on a sphere of cross-sectional area $\mathrm{d}a$:

$$\dot{\phi} = \frac{\mathrm{d}N}{\mathrm{d}a}. \tag{1}$$

The fluence is measured in unit m^{-2}. The fluence rate is defined as

$$\dot{\phi} = \frac{d\phi}{dt} \tag{2}$$

and has unit $m^{-2}\,s^{-1}$. The distribution ϕ_E represents the fluence with respect to energy where $d\phi$ is the fluence of particles of energy between E and $E + dE$:

$$\dot{\phi}_E = \frac{d\phi}{dE}. \tag{3}$$

Dosimetric quantities should provide physical measures which are correlated with effects of ionizing radiation. The basic dose definitions are given in ❯ Chap. 10, "Radiation Protection". Operational quantities for practical measurements, both for area and for individual monitoring were introduced and further explained in ICRU Reports 39, 43, 47, and 51 (ICRU 1985, 1988, 1992, 1993). The International Commission on Radiation Protection has recommended their use in radiation protection measurements (ICRP 1991, 2007). They are based on the quantity dose equivalent and have the unit Sievert (Sv). The ICRU has provided definitions of the operational quantities at points at a depth d in phantoms made out of tissue-like materials. For strongly penetrating radiation as neutrons the depth d in the phantom is 10 mm. The operational quantities for strongly penetrating radiation are for area monitoring the ambient dose equivalent $H^*(10)$ and for individual monitoring the personal dose equivalent $H_p(10)$. The ICRU sphere with diameter 30 cm is the phantom for $H^*(10)$, while a slab phantom with dimensions 30 cm × 30 cm × 15 cm is used for the calibration to $H_p(10)$. The ICRU material has a mass density of 1 g cm^{-3} and a mass composition of 76.2% oxygen, 11.1% carbon, 10.1% hydrogen, and 2.6% nitrogen (ICRU 2001).

The relation between radiometric quantities and the operational quantities is established by fluence-to-dose-equivalent conversion factors. The operational dose-equivalent quantities $H^*(10)$ and $H_p(10)$ for neutrons are determined from the equations

$$H^*(10) = \int h^*_\phi(E)\,\phi_E\,dE, \tag{4}$$

$$H_p(10) = \int h_{p,\phi}(E)\,\phi_E\,dE, \tag{5}$$

where ϕ_E is the energy distribution of the neutron fluence and $h_{p,\phi}(E)$ and $h^*_\phi(E)$ are the corresponding energy-dependent fluence-to-dose-equivalent conversion coefficients. These coefficients were calculated by several groups and at an international level it was agreed upon the numerical values that can be found in ICRP Report 74 (ICRP 1996) and in ICRU Report 57 (ICRU 1998b). The conversion factor $h^*_\phi(E)$ as a function of neutron energy is displayed in ❯ *Fig. 1*.

The fluence response R_ϕ of a radiation detector is a useful quantity specifying its sensitivity for detection. For an irradiation in a homogeneous radiation field with fluence ϕ it is defined as (ISO 8529-1 2001)

$$R_\phi = \frac{n}{\phi}, \tag{6}$$

where n is the total count of detected events. Fluence responses are measured in units of area, usually in cm^2. The fluence response corresponds to the area of a hypothetical detector with 100% efficiency.

A radiation detector's response R_H to dose equivalent H is defined as

$$R_H = \frac{R_\phi}{h_\phi}. \tag{7}$$

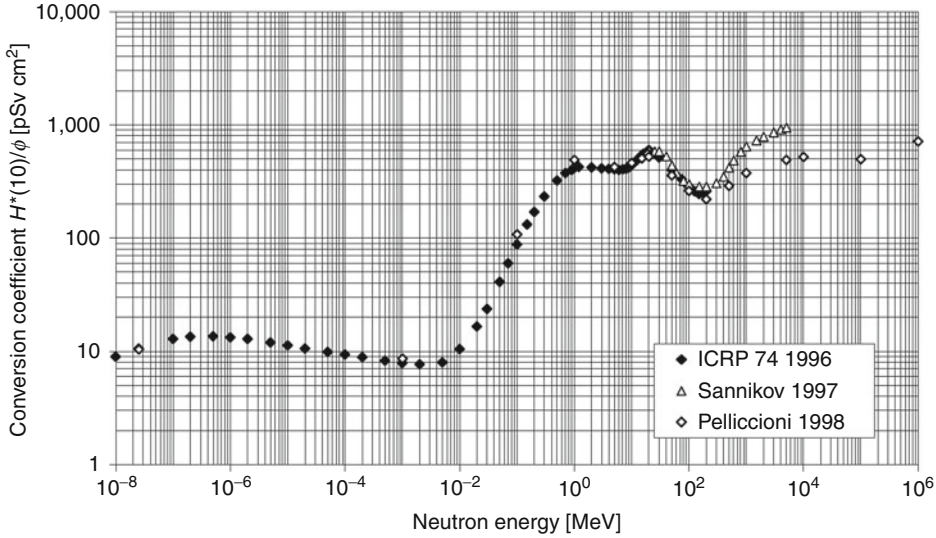


Fig. 1
Fluence-to-ambient-dose-equivalent $H^*(10)$ conversion coefficients for neutrons (ICRP 1996; Ferrari and Pelliccioni 1998; Sannikov and Savitskaya 1997)

3 Materials and Detector Types for Neutron Detection

3.1 Neutron Detection Principles

Ionizing radiation is a radiation consisting of directly or indirectly ionizing particles. A radiation detector is a device which in the presence of radiation provides a signal for use in measuring one or several quantities of the incident radiation. The detection of ionizing radiation usually utilizes ionization effects in detector materials. As neutrons are not directly ionizing they have to be converted into charged particles, which are then transferring their energy in direct ionization processes to the detector's sensitive volume. The ions are subject to charge collection and sometimes to internal amplification processes. Proportional counters, Geiger–Müller counters, GEM detectors, and photomultipliers are using avalanche charge amplification sometimes at very large gains. Pulse analysis and discrimination is easier with large signals. Other detector types like ionization chambers or semiconductors are only collecting the charge. Glenn Knoll gave an excellent and very detailed overview on radiation detection (Knoll 2010).

3.2 Active Neutron Detection Methods

3.2.1 Gas-Filled Detectors

Gas is a well suited medium for the detection of ionizing radiation. Free electric charges are mobile in gases and their recombination probability is relatively low. The conductivity of gases is low enough that high voltages can be applied to produce sufficiently strong electric fields.

Especially noble gases are well suited components of counting gases. Quenching gas admixtures are required if avalanche gas amplification is used. This holds for the detection of all types of ionizing radiation.

For neutron detection an efficient conversion process of neutrons into charged particles has to be added. There are a few gases where nuclear reactions in ❯ *Table 3* provide efficient neutron conversion. The most important counting gases in neutron detection are ^{3}He, BF_3, methane, and hydrogen. The reactions with ^{3}He and ^{10}B have sufficiently large cross sections and high Q-values to convert slow neutrons with high probability into charged particles with enough kinetic energy to exceed detection thresholds. In hydrogen or in methane neutrons are producing recoil protons. The neutron energy has to be not too low, because the recoil energy has also to be above detection threshold and in elastic scattering there is no energy contribution from a Q-value. In neutron detection the most common gas-operated detector types are proportional counters, ionization chambers, and fission chambers.

The working horse in neutron detection is certainly the ^{3}He proportional counter. Cylindrical tubes are available with diameters from a fraction of inch to several inches. There are also spherical counters and detectors with rectangular cross sections for time-of-flight spectroscopy. Tube lengths are ranging between a couple of centimeters up to several meters and ^{3}He filling pressures are up to 20 atm with small amounts of quenching gas. ^{3}He counters are rigid and can be operated at temperatures up to 200 °C. The efficiencies for thermal-neutron detection are high. The ^{3}He(n, p)^{3}H reaction releases a total of 764 keV kinetic energy as indicated by the Q-value in ❯ *Table 3*. According to the conservation of energy and momentum, the triton carries 191 keV and the proton 573 keV. Both particles are directly ionizing along their tracks through the gas volume and their total energy deposit is 764 keV. A typical pulse-height spectrum of a ^{3}He counter tube is shown in ❯ *Fig. 2*. with the full-energy-deposit peak at 764 keV. The tails at lower pulse heights correspond to events where one of the two decay particles has

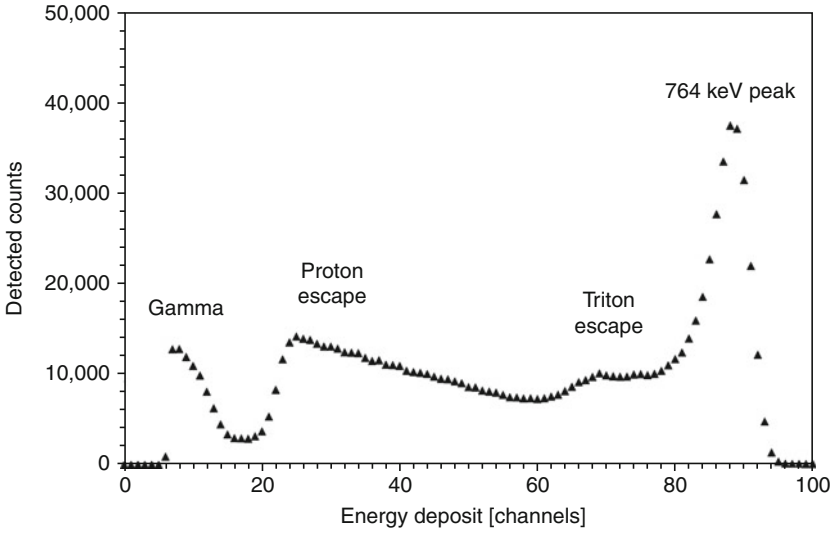


Fig. 2

Pulse-height spectrum of a ^{3}He counter tube with diameter 1″ and filling pressure 3.5 atm

hit the counter tube's wall and only a fraction of the full energy release is detected. Only one particle can hit the wall because the angle between both decay particles is 180°. The triton escape and the proton escape can clearly be seen in the pulse-height spectrum. All detected thermal-neutron signals are exceeding a minimum energy deposit. Below this minimum there is a gap. At very small pulse heights there are signals generated by gamma radiation. The discrimination threshold for the electronics is usually set above the gamma pulse heights and below the minimum neutron energy deposit.

Detectors filled with BF_3 have lower efficiencies than ^{3}He detectors due to the lower cross section and because the filling pressure is usually lower. But they have a better gamma discrimination because of the larger Q-value. Drawbacks of BF_3 are that the gas is toxic and corrosive. Most BF_3 counters are filled with pure boron tri-fluoride enriched to about 96% in ^{10}B.

Boron-lined proportional counters have a similar construction to BF_3 and ^{3}He proportional counters. However, the neutron detection is by means of a boron coating rather than boron or ^{3}He in a gaseous form, resulting in a higher neutron sensitivity. Typically, boron-lined proportional counters are used where the temperature limitations of BF_3 counters prevent their use. Detectors filled with hydrogen or with methane are used as recoil proton counters.

3.2.2 Semiconductors

The following semiconducting materials have been used in neutron detection:

- Silicon with ^{10}B coating or with ^{6}LiF film
- Gallium arsenide with ^{10}B coating
- Boron–carbide semiconductor diodes

The advantages of semiconductors for neutron detection are mainly compact size, relatively fast timing characteristics, and an effective thickness that can be varied to match the requirements of the application. Drawbacks may be the limitation to small sizes and the relative high susceptibility to performance degradation from radiation-induced damage (Knoll 2010).

3.2.3 Scintillators

The following scintillators have been used in slow-neutron detection:

- Boron-loaded plastic and liquid scintillators
- ^{6}Li scintillators (LiF, LiI, LiFZnS(Ag), glass, scintillating fibers)
- Gadolinium-loaded liquid scintillators

Advantages of scintillators are that they can be sensitive to the amount of energy deposit and that they are fast detectors which can be used for time-of-flight measurement. Scintillators are robust, easy to be operated, and relatively cheap. Their disadvantages are aging effects and radiation damage, sometimes difficulties in the light detection, for instance, with photomultipliers in the presence of magnetic fields and some scintillators are hygroscopic.

The dominant fast-neutron interaction in plastic or in liquid scintillators is the generation of recoil protons. Plastic scintillators consist out of a solid solution of organic scintillating molecules in a polymerized solvent. They are very popular because of the ease with which they

can be fabricated and shaped. Typical emission is at 400 nm. They have a large light output and short decay time and are well suited for timing measurement. Many scintillator designations are following the Saint-Gobain-type designation. A plastic scintillator for fast neutrons would be BC-720. Plastic scintillators do not allow for a good neutron–gamma separation.

The BC-501A (formerly called NE 213) is a widely used liquid scintillator with good pulse-shape discrimination properties intended for neutron detection in the presence of gamma radiation. It has extremely good timing properties and is well suited for coincidence measurements. Then there are for slow-neutron detection boron-loaded plastic scintillators (BC-454) and gadolinium (BC-521) and natural (BC-454) or enriched boron-loaded liquid scintillators (BC-523A). Also scintillators with lithium are quite common. The lithium-iodide crystal is chemically similar to sodium iodide and also hygroscopic. Liquid scintillator with lithium is also commercially available. There are as well lithium-containing glass scintillators. A new type of a scintillation detector for neutrons is scintillating glass fibers loaded with lithium. Anthracene and stilbene also have been used for neutron detection. For neutron radiography there are scintillators with phosphor screen based on ZnS(Ag) and ^{6}Li (BC-704 earlier NE-426).

3.2.4 Superheated Emulsion Detectors

Robert E. Apfel proposed in 1979 the superheated emulsion detectors, the so-called bubble technology as a new method for radiation detection (Apfel 1979). Superheated emulsion detectors are based on superheated droplets suspended in a viscoelastic gel medium, which vaporizes upon exposure to the high-LET recoils from neutron interactions. Bubbles evolved from the radiation-induced nucleation of drops give an integrated measure of the total neutron exposure. There are several different techniques to record and count the bubbles. In active devices they can be detected acoustically, by optical bubble counting, or by vapor volume measurement. Neutron spectrometry can be performed by measuring responses at different temperatures or pressures. Bubble detectors are insensitive to low-LET radiation like gammas or X-rays. Acoustical recording has the issue with discrimination of bubble pulses against noise. Francesco d'Errico published overviews concerning superheated emulsion detectors (d'Errico et al. 1995; d'Errico 2001; d'Errico et al. 2002).

3.3 Passive Neutron Detection Methods

3.3.1 Track Detectors

Passive neutron detection with nuclear track emulsions is the oldest and was once the most common method for neutron personal dosimetry (Knoll 2010; ICRU 2001; d'Errico and Bos 2004). The emulsions are relatively inexpensive, but track analysis under a microscope is laborious. New developments are focusing on automated track scanning methods. Passive radiation detectors have the advantage of being able to measure also in pulsed radiation fields where active devices may suffer from dead-time losses or pulse pile-up.

3.3.2 Thermoluminescent Dosimeters

In thermoluminescent dosimeters (TLDs) (Knoll 2010; ICRU 2001; d'Errico and Bos 2004), electrons are elevated by a radiation from the valence to the conduction band and captured

in trapping centers. Holes can also be trapped in analogous processes. The captured states are stable for longer periods. If a TLD is heated the trapped electrons or holes are re-excited and emit visible photons, which can be detected by a photomultiplier. The number of photons is a measure of the dose deposit. An exposed TLD material is thus an integrating detector for ionizing radiation. There are many TLD materials. LiF has been the most widely exploited (Knoll 2010). TLDs for neutron measurement are primarily used as albedo dosimeters, which is a dosimeter capable of measuring the fraction of neutrons reflected by a human body. Thermoluminescence detectors for neutron detection utilize typically ^{6}LiF and ^{7}LiF crystals. ^{6}LiF is sensitive to neutrons and to photons while ^{7}LiF is only sensitive to photons. The neutron contribution is calculated by determining the difference of both readings. TLDs are simple, rugged, and cheap. They have a good linearity and low detection limits. TLDs are usually processed by automatic readout.

3.3.3 Etched-Track Detectors

Etched-track detectors (Knoll 2010; ICRU 2001; d'Errico and Bos 2004) are together with TLDs the most commonly used passive neutron detectors. Charged particles, like alpha particles or protons, damage the material along their tracks, which can be made visible by chemical or electrochemical etching. These secondary particles can originate from nuclear reactions both in materials adjacent to an etched-track detector and those created inside the bulk of it. Etched-track detectors are usually processed by imaging systems which analyze and count the tracks and determine the dose. This method is insensitive to photons. In particularly over the last few tens of years, polymer etched-track detectors have been used. The most popular detector material is polyallyl-diglycol carbonate (PADC), commercially available as CR-39. ^{6}Li or ^{10}B are mainly used as converters for slow neutrons. Their response characteristics are generally sufficiently well known for neutrons with energies up to several hundred MeV. These dosimeters are generally able to determine neutron ambient dose equivalent down to a few tenths of a mSv.

3.3.4 Passive Superheated Emulsion Detectors

In passive superheated emulsion detectors or bubble detectors (ICRU 2001), the most immediate readout method is the visual inspection, a process that can be automated using video cameras and image analysis techniques. In the past decade superheated emulsions have achieved acceptance among the passive systems for personal neutron dosimetry. The detectors are considered to be the passive devices with the most accurate energy dependence of the response and the lowest detection threshold (d'Errico and Bos 2004).

3.3.5 Direct Ion Storage

Direct ion storage (DIS) is a relatively new technology. A small ionization chamber filled with air is in contact with the floating gate of a MOSFET transistor. The charge of the gate is initially set to a predetermined value. The charge generated by the radiation in the ionization chamber discharges partially the gate. The stored charge at the gate can be measured as a voltage without

modifying its value and it is proportional to the dose. Application of DIS to neutron dosimetry is possible using pairs of detectors for the separate determination of the photon and neutron dose contribution (d'Errico and Bos 2004; Fiechtner et al. 2004).

3.3.6 Other Passive Detectors

Radioluminescent glass detectors have found limited application in neutron dosimetry. A more recent technique is optically stimulated luminescence (OSL) (McKeever 2001). This method is based on laser stimulation and does not need heating of the detector material. It seems to be a breakthrough in passive radiation detection (d'Errico and Bos 2004).

4 Applications of Neutron Detection

4.1 Neutron Dose Measurement

4.1.1 Introduction

As a consequence of the 1990 recommendations of the International Commission on Radiological Protection, the operational quantities were newly defined (ICRP 1991). The relations between the neutron fluence and the two operational quantities, the ambient dose equivalent $H^*(10)$ and the personal dose equivalent $H_p(10)$, vary widely with neutron energy. The fluence response of a well-designed instrument, which will give a reading sufficiently proportional to the operational quantities, regardless of the neutron energy spectrum should have a fluence response as a function of energy that is inversely proportional to the fluence-to-dose conversion coefficients (Knoll 2010) provided by the International Commission on Radiological Protection (ICRP 1996). This is the entire secret of the art of accurate neutron dose measurement.

4.1.2 Rem Counters

The first neutron dose-rate meters based on the concept of an active thermal-neutron detector centered in an appropriate moderator with internal neutron absorbers – so-called rem counters – have been already designed in the 1960s. The Andersson–Braun counter and the Leake counter, became very popular and have been used all over the world for decades. While the Andersson–Braun counter has a cylindrical moderator and a BF_3 proportional counter tube as neutron detector (Andersson and Braun 1963, 1964) the Leake counter has a spherical moderator with a reduced weight of about 5 kg and had a LiI(Eu) crystal at the center (Leake 1966). The Leake design was later improved by replacing the crystal by a small spherical proportional counter filled with ^{3}He gas to achieve better gamma rejection properties and increased neutron sensitivity (Leake 1968).

After the publication of the ICRP60 recommendations of the International Commission on Radiological Protection (ICRP 1991), the Research Center Karlsruhe and Berthold Technologies designed a new neutron survey meter – the Berthold LB 6411 – with an energy-dependent

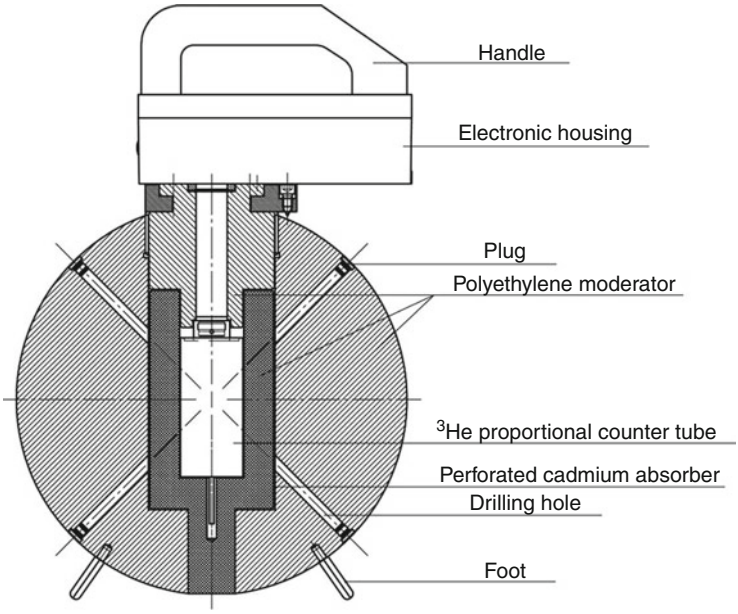


Fig. 3
Schematic drawing of the Berthold rem counter LB 6411

response optimized to the then new operational quantity ambient dose equivalent $H^*(10)$ (Klett and Burgkhardt 1997). The rem counter utilizes a cylindrical ^{3}He proportional counter tube centered in a moderating polyethylene sphere with 25 cm diameter. The energy-dependent response to $H^*(10)$ tuned with internal perforated-cadmium neutron absorbers and with boreholes is within ±30% for neutron energies between 50 keV and 10 MeV (Knoll 2010). This is standard in gamma dosimetry but it is excellent in neutron dosimetry. The $H^*(10)$ response to neutrons emitted by a bare ^{252}Cf source is approximately 3 counts/nSv, which is very high. *Figure 3* shows a drawing of the geometrical setup.

Figure 4 shows the responses of several widely used rem counters to ambient dose equivalent $H^*(10)$ which were calculated from the fluence responses given in a technical report of the IAEA (IAEA 2001). In the United Kingdom, the radiation protection group of the Health Protection Agency together with the neutron metrology group of the National Physical Laboratory have carefully assessed the performances of several instruments (Tanner et al. 2006).

For radiation protection purposes the response functions of conventional rem counters are considered to be acceptable at energies below 20 MeV. At higher energies the responses are decreasing and the instruments are underestimating ambient dose equivalent. A growing number of accelerators with high or even very high energies and enhanced interest in dose monitoring at flight altitudes triggered novel designs of instruments for extended energy ranges (Birattari et al. 1998; Fehrenbacher et al. 2007; Klett et al. 2007). Extended-range rem counters utilize layers of lead, tungsten, or other high-Z materials to convert in spallation processes high-energy neutrons into lower-energy neutrons. Response functions of several extended-range rem counters were calculated by Mares et al. (2002).

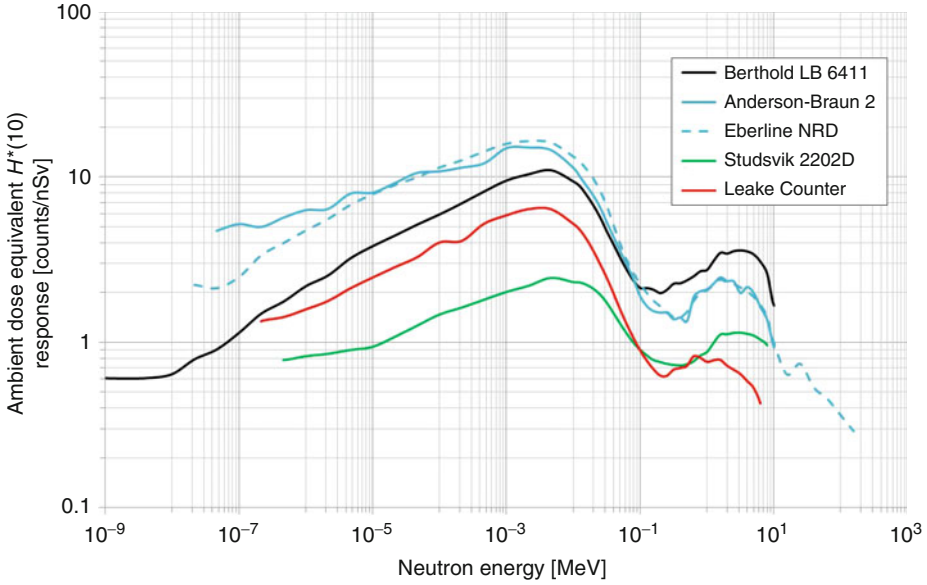


Fig. 4

Responses of several neutron survey meters to ambient dose equivalent $H^*(10)$

4.1.3 Tissue-Equivalent Proportional Counters

Tissue-equivalent proportional counters (TEPCs) allow the measurement of the probability distribution of the absorbed dose $d(y)$ in terms of lineal energy y in radiation fields. The lineal energy is defined as the ratio of the energy imparted to the matter in a volume by a single-deposition event to the mean chord length in that volume. The lineal energy can be used as an approximation of the linear energy transfer LET and the dose equivalent can be evaluated through a function $Q(y)$ which relates the quality factor to the lineal energy.

The TEPC is an important tool in microdosimetry and in some cases the only one to provide directly dose and radiation quality information in complex radiation fields. The TEPC can be used to distinguish photon and neutron contributions with good accuracy.

A tissue-equivalent proportional counter is a spherical or cylindrical detector with walls made out of tissue-equivalent material and filled with tissue-equivalent gas and operated in proportional mode. When the TEPC's gas density and its diameter is maintained at about 10^{-4} g cm^{-2}, it can simultaneously determine the absorbed dose to the tissue and of the spectrum of the pulse heights which corresponds to the energy deposition (ICRU 2001). By far the most commonly used TEPC has been the spherical counter. It has the advantage of being isotropic with respect to external radiation, but the electric field close to the wire has unfortunately not a cylindrical geometry. In order to obtain a cylindrical electric field geometry in the volume of gas amplification, Rossi proposed the so-called Rossi counter with an auxiliary helix electrode close to the wire (Rossi and Staub 1949; Rossi and Rosenzweig 1955). The sensitive volumes of TEPCs range from a few millimeters up to several centimeters in diameter. TEPCs are operated in pulse mode to record each individual event's energy deposit. The pulse height

is proportional to the charge released in the sensitive detection volume. The measurement of lineal energies from below 100 eV/μm up to more than 1 MeV/μm requires low-noise analogue electronics with a linearity over 4–5 orders of magnitude and an appropriate ADC system. TEPCs have not only been used in dose measurements at aviation altitudes and in mixed photon–neutron fields in accelerator environments, but also in investigations in radiotherapy and radiobiology (Kliauga et al. 1995; Gerdung et al. 1995).

4.1.4 Active Personal Dosimeters

A relatively new development now competing with and partially replacing the passive methods in individual dose monitoring are active personal dosimeters (APDs). In comparison to passive dosimeters they have the advantages of instant reading, audible alarm, lower detection limits, data memory, and communication capabilities with other hosts. There are several neutron APDs on the market, which have one or several silicon diodes or silicon strip detectors for neutron detection. The semiconducting detectors are combined with neutron converters or with layers of material for neutron activation, for instance, silver. Some use separate photon channels for the subtraction of gamma doses. The main issue with neutron APDs is their poor energy-dependent response. There were comparisons of instruments on the market and summaries published (Bolognese-Milsztajn et al. 2004; Luszik-Bhadra 2007). Recent measurements with electronic personal neutron dosimeters for high neutron energies were reported by Luszik-Bhadra (2007).

4.1.5 Passive Dose Measurement

There is a large amount of passive dose measurement in individual monitoring of occupational exposure. The passive neutron detectors that are used are described above. A good overview about the state of art of the available techniques was given by F. d'Errico and A.J.J. Bos (2004).

4.1.6 Dose Measurement in Pulsed Radiation Fields

Many accelerators or other radiation generators operate in pulsed mode. It is well known that active radiation detectors are subject dead to time effects and exhibit limitations in pulsed radiation fields (Knoll 2010). These limitations cannot easily be overcome without the development of new active detection technologies. Measurement of pulsed radiation is usually done with passive detectors.

There are now a few developments of new technologies based on the activation of radionuclides by pulsed radiation fields. One of these designs utilizes the neutron-induced activation of the nuclides ^{8}Li, ^{9}Li, and ^{12}B with short half-lives below 200 ms on the target nucleus ^{12}C in the detector materials. The decay products are detected in a time-resolved measurement. The instrument is mainly intended for radiation protection at accelerators with high energies and accomplishes the measurement of even very short and intense pulsed neutron fields (Klett and Leuschner 2007; Klett et al. 2010).

Luszik-Bhadra published another design of a new monitor for pulsed fields based on the activation of silver. The device comprises four silicon diodes in a 12″ polyethylene moderator sphere, two diodes covered on both sides with Ag, and two diodes covered with tin. The decay products of the activation products ^{109}Ag and ^{110}Ag are beta particles which are detected by the semiconductors. The detectors covered with silver are sensitive to neutrons and photons, while the detectors covered with tin are only sensitive to photons. The neutron dose is determined by subtraction (Luszik-Bhadra 2010; Leake et al. 2010).

4.1.7 Examples of Neutron Dose Measurements

The intensities of radiation levels at flight altitudes are exceeding-ground-level intensities by two orders of magnitude. The exposure of aircrews is comparable with or even larger than the exposure of workers classified as occupationally exposed. Primary galactic and solar particles – mainly protons – are interacting with the atmosphere and are generating secondary particles with a complicated composition. At flight altitudes of civil aircrafts about 50% of the ambient-dose-equivalent contribution is from neutrons, about 35% is from photons, electrons, and muons, and about 15% is from protons. Accurate dose measurements in these mixed fields with energies ranging from keV up to even exceeding the TeV domain is difficult. The recommendation by the International Commission on Radiological Protection (ICRP) in 1990, that exposure to cosmic radiation in the operation of jet aircraft should be recognized as occupational exposure, initiated a large number of new dose measurements onboard aircraft. A EURADOS working group has brought together all recent, available, preferably published, experimental data and results of calculations, mainly from laboratories in Europe (Lindborg et al. 2004). The reported results have been obtained using a variety of instrument types like rem counters, TEPCs, and Bonner sphere spectrometers. The results obtained are in good agreement almost all within ±25% of the mean values. During the time period 1995–1998 at temperate northern latitudes in 10 km altitude measured ambient-dose-equivalent rates for neutrons were about 3 μSv/h and the total about 5 μSv/h. The total exposure on a typical trans-Atlantic flight is about 50 μSv (Luszik-Bhadra 2007).

Another interesting example of neutron dose measurements was an international project investigating complex workplace radiation fields at European high-energy accelerators and thermonuclear fusion facilities. This study included all common types of existing neutron detection techniques in an environment with mixed radiation fields and high energies. The relevant techniques and instrumentation employed for monitoring neutron and photon fields around high-energy accelerators were reviewed with some emphasis on recent developments to improve the response of neutron-measuring devices beyond 20 MeV. It was investigated which type of area monitors to be employed (active and/or passive) and how they should be calibrated. The influence of the pulsed structure of the beam on the instruments and the needs and problems arising for the calibration of devices for high-energy radiation are addressed. The major high-energy European accelerator facilities are reviewed along with the way workplace monitoring is organized at each of them. The facilities taken into consideration are research accelerators, hospital-based hadron therapy centers, and thermonuclear fusion facilities. The issues of calibration are discussed and an overview of the existing neutron calibration facilities was provided (Bilski et al. 2006; Rollet et al. 2009; Silari et al. 2009).

4.2 Spectrometry

4.2.1 General

Neutrons appear in nature, in laboratories, or in nuclear facilities covering a very large energy range from ultracold up to ultrahigh energies at accelerators up to the TeV domain. As the interaction of neutrons with matter usually strongly depends on their energy, spectral information is needed in order to describe the occurring process. Spectrometry measurements are needed to characterize neutron fields. Commonly used measurement methods are Bonner sphere measurement, time-of-flight measurement, nuclear recoil measurement, neutron-induced nuclear reactions, methods based on activation and on threshold effects, and neutron diffraction. An excellent overview about neutron spectrometry for radiation protection was provided by Thomas (2004).

4.2.2 Bonner Spheres

In 1960, the Bonner sphere spectrometer (BSS) was first described by Bramblett, Ewing, and Bonner (Bramblett et al. 1960). Of the many types of neutron spectrometers that have been developed this multi-sphere system has been used by more laboratories than any other. It is easy to operate, it has an almost isotropic response, it covers energies from thermal up to GeV neutrons, and it can be used with active or with passive detectors. A BSS is consisting of several moderating spheres with different diameters and a thermal-neutron detector which is assembled in the centers of these spheres. The spheres are usually made out of polyethylene and each sphere with the thermal-neutron detector has a sensitivity to neutrons over a broad energy range. However, the sensitivity for each sphere peaks at a particular neutron energy depending on the sphere diameter. From the measured readings of a set of spheres, information can be derived about the spectrum of a neutron field (Thomas and Alevra 2002; Thomas and Klein 2003).

Several types of thermal-neutron detectors have been used. In the original Bonner sphere spectrometer a small ^{6}LiI(Eu) scintillator was used. Various cylindrical and spherical proportional counter tubes filled with BF_3 or ^{3}He are obvious alternatives. Several groups investigated the use of the SP9 spherical ^{3}He proportional counter produced by Centronic Ltd. UK. It has a diameter of 32 mm and a gas pressure of about 2 atm. The characteristics of BSSs with this detector are well established. Typical moderator sphere diameters are between 3″ and 18″ with the number of spheres in a set ranging between 6 and 12 spheres (Thomas and Alevra 2002).

If sphere i has response function $R_i(E)$ and is exposed in a neutron field with the spectral fluence $\phi(E)$, then the sphere reading M_i is obtained mathematically by folding $R_i(E)$ with $\phi(E)$:

$$M_i = \int R_i(E)\phi(E)\,\mathrm{d}E. \tag{8}$$

This integral extends over the range of neutron energies present in the field. Good approximations of $R_i(E)$ can be obtained by simulation calculations supported by measurements in well-characterized reference neutron fields. Information about the spectrum $\phi(E)$ can be extracted by unfolding. However, because the total number of spheres is limited the solution may provide a poor representation of the spectrum with important features smeared out. Additional a priori information on the spectrum is useful (Thomas and Alevra 2002).



Fig. 5
Bonner sphere spectrometer NEMUS of the PTB with five of the ten polyethylene spheres in the back, a spherical ^{3}He proportional counter in the center, and parts of the modified spheres in the foreground *left* and *right*

The PTB NEMUS system, the INFN Frascati BSS, and the NPL BSS are a few examples where detailed descriptions, measurements, and intercomparison data were published (Wiegel and Alevra 2002; Bedogni and Esposito 2009). *Figure 5* shows the components of the PTB NEMUS Bonner sphere spectrometer. BSSs have an excellent energy range, good sensitivity, isotropy and photon discrimination, simple but time-consuming operation, and poor energy resolution. According to a neutron field's intensity and time structure various types of active or passive thermal-neutron detectors can be selected. The data analysis requires complex unfolding of the measured data. For instance FRUIT (Frascati Unfolding Interactive Tool), an unfolding code for Bonner sphere spectrometers, was developed under the Labview environment at the INFN-Frascati National Laboratory and is available from the authors upon request (Bedogni et al. 2007). An excellent overview on Bonner sphere neutron spectrometry was provided by Thomas and Alevra (2002).

4.2.3 Time-of-Flight Spectroscopy

Time-of-flight spectroscopy is based on measurement of the time it takes for a neutron to travel a known distance. From this the neutron's velocity and the energy can be calculated. The measurement needs a start and a stop signal. The former can for instance be derived from a pulsed neutron generation process while the latter is generated when the neutron arrives at a distant neutron detector. The indication of the start can be generated by time-pick-up signals from an accelerator, by an appropriate detector, or by a beam chopper. To minimize uncertainties time-of-flight spectroscopy requires precise time measurement. Therefore the detectors from which

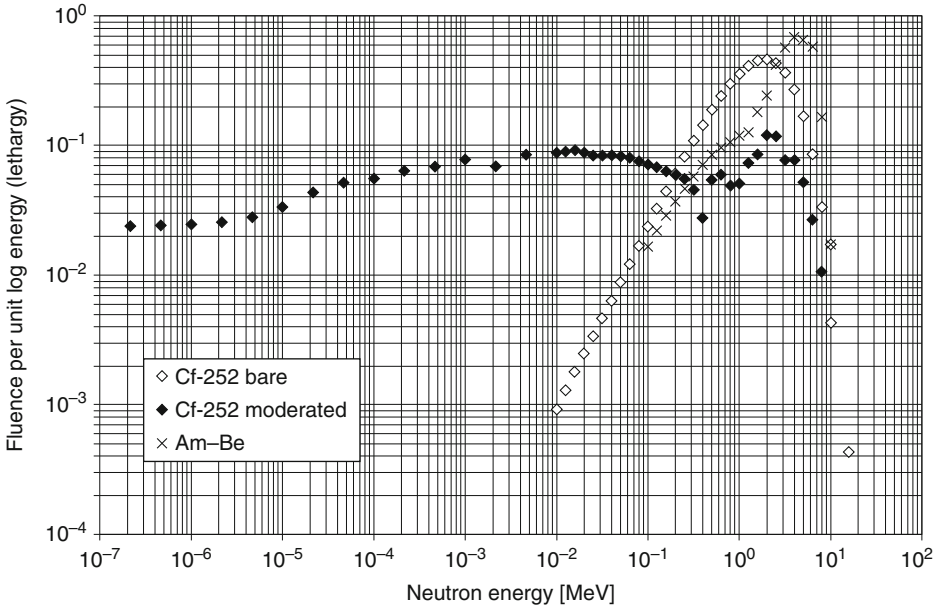


Fig. 6
Neutron spectra of bare ^{252}Cf, moderated ^{252}Cf, and ^{241}Am–Be (IAEA 2001)

the time information is derived have to be very fast. Excellent timing characteristics have for instance plastic scintillators. The flight paths have to be very long and can be as long as the order of magnitude of 100 m.

4.2.4 Recoil Spectroscopy

Another neutron detection technique with spectral sensitivity is based on elastic neutron scattering on light target nuclei. During the interaction a fraction of the neutron's energy is transferred to the recoil nucleus. As the recoil nucleus is directly ionizing it deposits its energy in the detector materials. The maximum energy $E_{\max}$ which can be transferred from a neutron with kinetic energy E_n to a recoil nucleus with mass A in units of neutron mass is (Knoll 2010):

$$E_{\max} = \frac{4A}{(1+A)^2} E_n. \tag{9}$$

The maximum fractional energy transfer in elastic neutron scattering is 1 for hydrogen, 0.64 for ^{4}He, and 0.22 for ^{16}O. Therefore only light nuclei are of primary interest with hydrogen being the best choice. Recoil proton spectroscopy can be performed with detectors with a substantial amount of hydrogen in the detector material. The easiest detector would be a scintillator containing hydrogen as organic crystals, plastic scintillators, or liquid scintillators. Liquid scintillators have the advantage of the possibility for neutron–gamma discrimination. Another detector type for recoil spectroscopy would be gas recoil proportional counters filled

with hydrogen, methane, or with helium. The energy distribution for all scattering angles in elastic neutron–proton scattering is approximately a rectangular function. This is the detector's response function. The neutron spectrum has to be determined by deconvolution (Knoll 2010).

4.3 Neutron Activation Analysis

Neutron activation analysis (NAA) is one of the most sensitive analytical techniques to determine concentrations of many elements in a variety of materials. It is based on neutron activation and requires irradiation of the specimen with neutrons. This creates artificial radioisotopes of the elements of interest. Usually reactors are used but other types of neutron sources as previously discussed can also be used, if the energy and intensity requirements are met. The decay products of the artificial radioisotopes are then measured and analyzed. Preferably gamma spectroscopy allows for the identification of nuclides and for quantitative measurement of concentrations. As it is a nuclear method NAA does not depend on the chemical form of the sample. NAA is a nondestructive technique and requires usually very small amounts of sample. Many elements can be determined at the same time (Alfassi 1990).

There are a few important experimental conditions for NAA, first of all the kinetic energy of the neutrons used for irradiation. Generally neutron activation is performed with thermal neutrons, but there are also nuclear reactions used, where higher neutron energies are required. The intensity of the neutron field and the cross sections of the selected activation processes are important parameters. The nuclear decay products can be measured during or after neutron irradiation.

4.4 Neutron Scattering

There is a lot of research utilizing neutron scattering in biology, biotechnology, medicine, nanotechnology, and in research on catalysis, drugs, energy, magnetism, molecular structure, polymers, and superconductors all over the world. Some of the leading institutes are the Institut Laue–Langevin in Grenoble, the Rutherford–Appleton Laboratories in Oxford with ISIS, the FRM II reactor in Munich, the Research Center Jülich, the Paul Scherrer Institute in Würenlingen with SINQ, the KENS Neutron Scattering Facility at KEK in Japan, the Oak Ridge National Laboratory with their new spallation neutron source, and the Los Alamos Neutron Science Center LANSCE. These research centers are using special techniques like elastic and inelastic scattering, diffractometry, time-of-flight measurement, small-angle neutron scattering, reflectometry, or measurements of polarized neutrons. The international scientific community benefits from applying these sophisticated large installed detector systems, which employ physical principles and detection techniques that were discussed here. The details of these spectrometers and techniques are very elaborate and are not subject of this article. More information about these detector systems and research disciplines can be obtained from the Web sites of the neutron scattering laboratories.

4.5 Nuclear Medicine

In radiotherapy with neutron beams, the estimation of the neutron doses to the organs surrounding the target volume is particularly challenging. For instance at the Louvain-la-Neuve (LLN)

facility these doses were investigated. The transport of a 10 cm × 10 cm beam through a water phantom was simulated with the Monte-Carlo code MCNPX and measurements of the absorbed dose and of dose equivalent using an ionization chamber and superheated-drop detectors were performed (Benck et al. 2002).

Boron neutron capture therapy (BNCT) is a cancer treatment method where after the delivery of a suitable boron compound to tumor cells, the tumor is irradiated with slow neutrons. The boron concentration in the tumor has to exceed the boron concentration in normal tissue considerably, which can be achieved by a number of compounds. A reactor or an accelerator has to deliver large thermal-neutron fluences of the order of magnitude of 10^{12} n/cm^2 to get sufficient results of these irradiations (Gahbauer et al. 1997). Accurate measurements of neutron fluences and dose distributions as well as Monte-Carlo calculations are the base of treatment planning. Neutron fluence and absorbed-dose measurements were for instance performed with activation foils and paired ionization chambers (Binns et al. 2005).

4.6 Search for Illicit Trafficking Nuclear Materials

Since the demise of the former Soviet Union and even more since the new terrorism, the search for illicit trafficking or hidden nuclear material became an important new application in radiation detection. In the beginning the main focus was in gamma detection, but soon neutron detection was also included, because plutonium, a material used for nuclear weapons, is a significant source of fission neutrons (Kouzes 2005).

Neutron detection is very selective for the indication of dangerous nuclear materials. Plutonium is extremely hazardous and hard to be detected, because it is not very difficult to shield the alpha, beta, and photon emissions. But the even-numbered plutonium isotopes exhibit significant spontaneous fission yields. For instance 1 g of ^{238}Pu is emitting 2,660 fission neutrons per second. Therefore a neutron detector for the search for illicit trafficking or hidden nuclear material needs a maximum of sensitivity in the fission-neutron energy region. Rem counters are for these applications not sensitive enough and perform poorly, because their "dose tuning" is based on a tremendous amount of neutron filtering and absorption. A well-designed detector's energy-dependent response should be optimized to fission neutrons at a maximum of sensitivity. This can be achieved by a moderator of reasonable size in which a large thermal-neutron detector is located. An example of a highly sensitive handheld detector is described in Klett (1999).

Neutron detection is now widely used in security applications like access and exit control of nuclear facilities, vehicle monitoring, border control, monitoring in harbors and airports, in waste storage management, and in safeguard activities. The Austrian Research Center Seibersdorf in cooperation with a team of International Atomic Energy Agency IAEA experts and supported by the World Custom Organization (WCO) and by INTERPOL has performed the Illicit Trafficking Radiation Detection Assessment Program (ITRAP). The aim of the study was to work out the technical requirements and the practicability of useful monitoring systems. International suppliers and manufacturers of radiation detection equipment from nine different countries have participated. The study covered fixed-installed monitoring instruments, pocket-type instruments, and handheld instruments. Neutron monitoring should be included in the

fixed-installed systems; it was desirable for the handhelds and not necessary for the pocket-type instruments (Beck 2000).

For the detection of special nuclear materials (SNM) there are neutron coincidence counters in use. They have arrays of neutron detectors – usually large ^{3}He proportional counters in moderators optimized for fission-neutron detection – covering a container from several sides. There is active interrogation and passive measurement. For active interrogation a neutron source is used to induce fission in a fissile material under investigation. Passive measurement measures neutrons emitted by the sample without external irradiation.

Homeland security applications consumed large amounts of ^{3}He since a decade. Since about 2008 there is now a worldwide shortage in ^{3}He supply and many groups are now developing neutron detection alternatives (Kouzes et al. 2010)

4.7 Reactor Instrumentation

Reactor instrumentation requires mostly slow-neutron detection at high intensities and under extreme conditions of reactor operation. Neutron intensities have to be measured in core up to 10^{14} cm^{-2} s^{-1} and out of core up to 10^{10} cm^{-2} s^{-1}. There are high pressures and temperatures which can be as high as 300 °C. Because of a lower gamma sensitivity, gas-filled detectors is a preferable choice.

Boron ionization chambers can be tailored to measure the required range of neutron flux. Uncompensated boron ion chambers are generally used in regions of high neutron flux where the gamma flux is only a small share of the total radiation level. Fission chambers can be used in pulse or in direct-current mode. Fission chambers include a fissile material normally ^{235}U. The fission fragment's large energy deposits are generating the detector signals. Fission chambers in pulse mode are ideal in mixed fields because gamma discrimination is easy in pulse mode. So-called self-powered detectors utilize a material with a high cross section for neutron capture with subsequent beta decay. The beta decay current is measured without external bias voltage. Overviews and more details about reactor instrumentation can be found in Knoll (2010) and Boland (1970).

4.8 Fusion Monitoring

Neutron spectrometry is a tool for obtaining fusion plasma information such as ion temperature and fusion power. Neutron spectrometry measurements for diagnostics at the Joint European Torus (JET) between 1983 and 1999 were reported by Jarvis (2002). A wide variety of spectrometer types with nuclear emulsions, NE213 liquid scintillators, hydrogen ionization chambers, recoil proton counters, ^{3}He ionization chambers, silicon detectors, and diamond detectors have been tested with varying degrees of success. Magnetic proton recoil spectrometers are successful in monitoring the d–d and the d–t reactions at 2.5 MeV and 14 MeV, respectively. Investigations about the time resolutions of several different neutron spectrometry techniques and an upgraded magnetic recoil spectrometer for ITER were recently described by Andersson (2010).

4.9 Industrial Applications

4.9.1 Neutron Imaging and Radiography

As neutron interactions with atoms and molecules are very different from X-ray interactions, the neutron is sensitive to other aspects of matter. For instance, in investigating in automotive industries with imaging techniques X-rays would illuminate the metallic structure of an engine while neutrons would rather take a picture from the oil. Neutron imaging requires position-sensitive neutron detectors.

Neutron transmission radiography (NR) is based on the attenuation of radiation passing through a sample. Details of samples can be made visible, if the attenuation is different in different materials. As neutron detectors track etch foils, a combination of a neutron converter layer (Gd, Dy) and X-ray film, or a combination of a neutron-sensitive scintillator and a CCD camera or position-sensitive ^{3}He detectors have been used. A new development is amorphous-silicon flat panels. They contain Gd as neutron absorber and BaFBr:Eu^{2+} as the agent which provides the photoluminescence. An imaging plate scanner is extracting the digitized image information from the plates by de-excitation caused by a laser signal.

4.9.2 Humidity Measurement

Because water is an excellent neutron moderator it is possible to measure humidity with neutrons. If fast neutrons emitted by a neutron source are penetrating humid matter, there is thermalization depending on the amount of water. A typical neutron humidity measurement setup comprises a fast-neutron source and a detector for thermal neutrons close to each other. The probe is positioned in or close to the sample material, which could be coal, coke, sand, sinter, soil, or lime sand bricks. The measurement is online and continuously without direct contact with the sample. The measurement is not affected by temperature, pressure, pH value, or optical characteristics of the material and determines of the amount of water molecules, irrespective of their physical or chemical binding. Humidity measurement with neutrons is mainly used by chemical, cement, ceramics, coal, iron, and steel industries.

5 Reference Neutron Radiation Fields

Reference neutron radiation fields are very important for the calibration of neutron detectors. The following types are commonly used:

- Neutrons from radionuclide sources, including sources in a moderator
- Neutrons generated by nuclear reactions with charged particles from accelerators
- Neutrons from reactors

These fields have usually unidirectional beams and cover neutron energies in the range from thermal up to several hundred MeV. There are quasi-monoenergetic neutron radiations for determining the response of neutron-measuring devices as a function of energy and there are neutron fields with wide spectra for calibration of instruments.

Table 7

Commonly used ISO reference neutron radiations (ISO 8529-1 2001; ISO 8529-2 2000; ISO 8529-3 1998) with fluence-averaged energies *E* and fluence-to-dose conversion factors $h^*_\phi(E)$ according to ICRP74 (ICRP 1996)

Source/Generation	Half-life [years]	Reaction	Energy [MeV]	$h^*_\phi(E)$ [pSv cm^2]
^{252} Cf (D_2O moderated)	2.65	Spontaneous fission	0.550	105
^{252}Cf	2.65	Spontaneous fission	2.130	385
^{241}Am–Be	432	(α, n)	4.160	391
Reactor or accelerator		^{9}Be(d, n)X/thermal column	2.5×10^{-8}	10.6
Sc-filtered reactor beam			2×10^{-3}	7.7
Accelerator		^{45}Sc(p, n)^{45}Ti	24×10^{-3}	19.3
Accelerator		T(p, n)^{3}He and ^{7}Li(p, n)^{7}Be	0.144	127
Accelerator		T(p, n)^{3}He and ^{7}Li(p, n)^{7}Be	0.250	203
Accelerator		T(p, n)^{3}He and ^{7}Li(p, n)^{7}Be	0.565	434
Accelerator		T(p, n)^{3}He	1.200	425
Accelerator		T(p, n)^{3}He	2.5	416
Accelerator		D(d, n)^{3}He	5.0	405
Accelerator		T(d, n)^{4}He	14.8	536
Accelerator		T(d, n)^{4}He	19.0	584
Accelerator		Université Catholique de Louvain (UCL)	33 and 50	
Accelerator		TSL Uppsala	200	
Accelerator		CERN/CERF (Mitaroff and Silari 2002)	$<1 \times 10^3$	

An instrument under calibration is placed in a free-in-air radiation field of known fluence rate and the reading is recorded. Neutron scattering from the air, by the walls, floor, and ceiling should be minimized and corrected for. The room used for irradiation should be as large as possible and measurements with a shadow cone help to take into account the scattered neutrons' contributions. The international standard ISO 8529 about reference neutron radiations covers in its three parts the general principles, the commonly used radiation fields which are listed in *Table 7*, and the calibration procedures (ISO 8529-1 2001; ISO 8529-2 2000; ISO 8529-3 1998).

The facilities providing neutron radiations traceable to national standards are for instance the Physikalisch-Technische Bundesanstalt PTB in Germany, the National Physical Laboratory NPL in the United Kingdom, and the National Institute of Standards and Technology NIST in the United States.

6 Conclusion

As neutrons are neutral and not directly ionizing particles, neutron detection is more difficult than photon or charged-particle detection. Neutrons can only be detected after conversion into charged particles. There are several efficient nuclear processes to convert neutrons into charged particles, among them especially elastic n–p scattering and nuclear reactions on ^{3}He, ^{6}Li, and ^{10}B target nuclei. These processes are utilized in a variety of neutron detection techniques which are used in research, nuclear medicine, industry, and in many other fields.

Acknowledgment

Thanks to Dr. Burkhard Wiegel/PTB Braunschweig for the photo and print permission of the NEMUS spectrometer.

References

Alfassi ZB (1990) Activation analysis, vol I and II. CRC, Boca Raton

Amsler C et al (2008) Particle data group. Phys Lett B667:1

Andersson IÖ, Braun J (1963) A neutron rem counter with uniform sensitivity from 0.25 eV to 10 MeV. In: Proceedings of the IAEA Symposium on Neutron Dosimetry, vol II. STI/PUB/69. IAEA, Vienna, pp 87–95

Andersson IÖ, Braun J (1964) Nukleonik 6:237

Andersson Sundèn E (2010) Neutron spectrometry techniques for fusion plasmas, instrumentation and performance. Thesis, Uppsala University, Sweden

Apfel RE (1979) The superheated drop detector. Nucl Instrum Methods 162:603–608

Beck P (2000) Final report: ITRAP Illicit Trafficking Radiation Detection Assessment Program Austrian Research Centers. Seibersdorf, Austria

Bedogni R, Esposito A (2009) Measurements of Neutron Spectrum in the high-energy DAΦNE accelerator complex with an extended range Bonner sphere spectrometer radiation measurements and instrumentation. Nucl Technol 168:615–619

Bedogni R, Domingo C, Esposito A, Fernández F (2007) FRUIT: an operational tool for multisphere neutron spectrometry in workplaces. Nucl Instrum Methods A580:1301–1309

Benck S, D'Errico F, Denis J-M, Meulders JP, Nath R, Pitcher EJ (2002) In-phantom spectra and dose distributions from a high-energy neutron therapy beam. Nucl Instrum Methods A476: 127–131

Bilski P, Blomgren J, d'Errico F, Esposito A, Fehrenbacher G, Fernandez F, Fuchs A, Golnik N, Lacoste V, Leuschner A, Sandri S, Silari M, Spurny F, Wiegel B, Wright P (2006) Complex workplace radiation fields at European high-energy accelerators and thermonuclear fusion facilities, Yellow report CERN-2006–007. CERN, Geneva

Binns PJ, Riley KJ, Harling OK (2005) Epithermal neutron beams for clinical studies of boron neutron capture therapy: a dosimetric comparison of seven beams. Radiat Res 164(2):212–220

Birattari C, Esposito A, Ferrari A, Pelllicioni M, Rancati T, Silari M (1998) The extended range neutron rem counter LINUS: overview and latest developments. Radiat Prot Dosim 76:135–148

Boland JF (1970) Nuclear reactor instrumentation. Gordon & Breach, New York

Bolognese-Milsztajn T, Ginjaume M, Luszik-Bhadra M, Vanhavere F, Wahl W, Weeks A (2004) Active personal dosemeters for individual monitoring and other new developments. Radiat Prot Dosim 112(1):141–168

Bramblett RL, Ewing RI, Bonner TW (1960) A new type of neutron spectrometer. Nucl Instrum Methods 9:1–12

Chadwick J (1932a) Possible existence of a neutron. Nature 129:312

Chadwick J (1932b) The existence of a neutron. Proc R Soc A 136:692

d'Errico F (2001) Radiation dosimetry and spectrometry with superheated emulsions. Nucl Instrum Methods B184:229–254

d'Errico F, Bos AJJ (2004) Passive detectors for neutron personal dosimetry: state of the art. Radiat Prot Dosim 110(1–4):195–200

d'Errico F, Agosteo S, Sannikov AV, Silari M (2002) High-energy neutron dosimetry with superheated drop detectors. Radiat Prot Dosim 100:529–532

d'Errico F et al (1995) Active neutron spectrometry with superheated drop (Bubble) detector. Radiat Prot Dosim 61:159–162

Evans RD (1982) The atomic nucleus. Krieger, New York

Fehrenbacher G, Kozlova E, Gutermuth F, Radon T, Schütz R (2007) Neutron dose measurements with the GSI ball at high-energy accelerators. Radiat Prot Dosim 125(1–4):209–212

Ferrari A, Pelliccioni M (1998) Fluence to dose equivalent conversion data and effective quality factors for high energy neutrons. Radiat Prot Dosim 76(4):215–224

Fiechtner A, Boschung M, Wernli C (2004) Present status of the personal neutron dosemeter based on direct ion storage. Radiat Prot Dosim 110(1–4):213–217

Gahbauer R, Gupta N, Blue T, Goodman J, Grecula J, Soloway AH, Wambersie A (1997) BNCT: status and dosimetry requirements. Radiat Prot Dosim 70(1–4):547–554

Gerdung S, Pihet P, Grindborg JE, Roos H, Schrewe UJ, Schuhmacher H (1995) Operation and application of tissue-equivalent proportional counters. Radiat Prot Dosim 61(4):381–404

IAEA (2001) Compendium of neutron spectra and detector responses for radiation protection purposes, Supplement to Technical Reports Series No. 318, Technical reports series No. 403. International Atomic Energy Agency, Vienna

IAEA International Atomic Energy Agency (1974) Handbook on nuclear activation cross-sections, Technical reports series No. 156. IAEA, Vienna

ICRP (The International Commission on Radiological Protection) (1991) 1990 recommendations of the International Commission on Radiological Protection, Annals of the ICRP, v.21, no. 1–3; ICRP publication, 60. Pergamon, Oxford

ICRP (The International Commission on Radiological Protection) (1996) Conversion coefficients for use in radiological protection against external radiation, Annals of the ICRP, v 26, no. 3–4; ICRP publication, 74. Pergamon, Oxford

ICRP (The International Commission on Radiological Protection) (2007) The 2007 recommendations of the International Commission on Radiological Protection, Annals of the ICRP, v 37, no. 2–4; ICRP publication, 103. Elsevier, Oxford

ICRU (The International Commission on Radiation Units and Measurements) (1985) Determination of dose equivalents resulting from external radiation sources, ICRU Report 39. ICRU, Bethesda

ICRU (The International Commission on Radiation Units and Measurements) (1988) Measurement of dose equivalents from external radiation sources, Part 2, ICRU Report 43. ICRU, Bethesda

ICRU (The International Commission on Radiation Units and Measurements) (1992) Measurements of dose equivalents from external photon and electron radiations, ICRU Report 47. ICRU, Bethesda

ICRU (The International Commission on Radiation Units and Measurements) (1993) Quantities and units in radiation protection dosimetry, ICRU Report 51. ICRU, Bethesda

ICRU (The International Commission on Radiation Units and Measurements) (1998a) Fundamental quantities and units for ionizing radiation, ICRU Report 57. ICRU, Bethesda

ICRU (The International Commission on Radiation Units and Measurements) (1998b) Conversion coefficients for use in radiological protection against external radiation, ICRU Report 60. ICRU, Bethesda, MD

ICRU (The International Commission on Radiation Units and Measurements) (2001) Determination of operational dose equivalent quantities for neutrons, ICRU Report 66. J ICRU 1(3) Table A.42 page 119

International Organization for Standardization (ISO 8529-1) (2001) Reference neutron radiations, 1st edn, Part 1: Characteristics and methods of production. International Organization for Standardization, Geneva

International Organization for Standardization (ISO 8529-2) (2000) Reference neutron radiations, 1st edn, Part 2: Calibration fundamentals of radiation protection devices related to the basic quantities characterizing the radiation field. International Organization for Standardization, Geneva

International Organization for Standardization (ISO 8529-3) (1998) Reference neutron radiations, 1st edn, Part 3: Calibration of area and personal dosimeters and determination of their response as a function of neutron energy and angle of incidence. International Organization for Standardization, Geneva

Jarvis ON (2002) Neutron spectrometry at JET (1983–1999). Nucl Instrum Methods A476(1–2):474–484

Klett A, Burgkhardt B (1997) The new remcounter LB6411: measurement of neutron ambient dose equivalent H*(10) according to ICRP60 with high sensitivity. IEEE Trans Nucl Sci 44(3): 757–759

Klett A (1999) Plutonium detection with a new fission neutron survey meter. IEEE Trans Nucl Sci 46(4):877–879

Klett A, Leuschner A (2007) A pulsed neutron monitor IEEE 2007 Nuclear Science Symposium and Medical Imaging Conference, Conference Records Oct 27–Nov 3. Honolulu, Hawaii

Klett A, Mayer S, Theis C, Vincke H (2007) A neutron dose rate monitor for high energies. Radiat Meas 41(Suppl 2):279–282

Klett A, Leuschner A, Tesch N (2010) A dose meter for pulsed neutron fields: 11th Neutron and Ion Dosimetry Symposium NEUDOS-11, Cape Town, South Africa, 12–16 October 2009. Conference Proceedings – Radiat Meas 45(10): 1242–1244, Elsevier

Kliauga P, Waker AJ, Barthe J (1995) Design of tissue-equivalent proportional counters. Radiat Prot Dosim 61(4):309–322

Knoll (2010) Glenn F, Radiation detection and measurement, 4th edn. Wiley, New York

Kouzes RT (2005) Detecting illicit nuclear materials. Am Sci 93:422–427

Kouzes RT, Ely JH, Erikson LE, Kernan WJ, Lintereur AT, Siciliano ER, Stephens DL, Stromswold DC, Van Ginhoven RM (2010) Neutron detection alternatives to ^{3}He for national security applications, Nucl Instrum Methods Phys Res A 623:1035–1045

Leake JW (1966) A spherical dose equivalent neutron detector. Nucl Instrum Methods 45: 151–156

Leake JW (1968) An improved spherical dose equivalent neutron detector. Nucl Instrum Methods 63:329–332

Leake JW, Lowe T, Mason RS, White G (2010) A new method of measuring a large pulsed neutron fluence or dose exploiting the die-away of thermalized neutrons in a polyethylene moderator. Nucl Instrum Methods A613:112–118

Lindborg L, Bartlett DT, Beck P, McAulay IR, Schnuer K, Schraube H, Spurny F (2004) Cosmic radiation exposure of aircraft crew: compilation of measured and calculated data. Final report of the EURADOS WG 5. European Commission, Directorate-General for Energy and Transport, Radiation Protection Issue No. 140, Luxembourg, ISBN 92-894-8448-9

Luszik-Bhadra M (2007) Electronic personal neutron dosemeters for high neutron energies: measurements, new developments and further needs. Radiat Prot Dosim 126(1–4): 487–490

Luszik-Bhadra A (2010) New neutron monitor with silver activation: 11th Neutron and Ion Dosimetry Symposium NEUDOS-11, Cape Town, South Africa, 12–16 October 2009. Conference Proceedings – Radiat Meas 45(10):1258–1262, Elsevier

Luszik-Bhadra M, Bolognese-Milsztajn T, Boschung M, Coeck M, Curzio G, Derdau D, D'Errico F, Fiechtner A, Kyllönen J-E, Lacoste V, Lievens B, Lindborg L, Lovefors Daun A, Reginatto M, Schuhmacher H, Tanner R, Vanhavere F (2007) Summary of personal neutron dosemeters results obtained within the EVIDOS project. Radiat Prot Dosim 125(1–4): 293–299

Mares V, Sannikov AV, Schraube H (2002) Response functions of the Andersson-Braun and extended range rem counters for neutron energies from thermal to 10 GeV. Nucl Instrum Methods A 476:341–346

McKeever SWS (2001) Optically stimulated luminescence dosimetry. Nucl Instrum Methods B184(1–2):29–54

Mitaroff A, Silari M (2002) The CERN-EU high energy reference field (CERF) facility for dosimetry at commercial flight altitudes and in space. Radiat Prot Dosim 102(1):7–22

Rollet S, Agosteo S, Fehrenbacher G, Hranitzky C, Radon T, Wind M (2009) Intercomparison of radiation protection devices in a high-energy stray neutron field. Part I: Monte Carlo simulations. Radiat Meas 44:649–659

Rossi HH, Rosenzweig WA (1955) Device for the measurement of dose as a function of specific ionization. Radiology 64:404

Rossi BB, Staub HH (1949) Ionization chambers and counters. McGraw-Hill, New York

Sannikov AV, Savitskaya EN (1997) Ambient dose equivalent conversion factors for high energy neutrons based on the ICRP60 recommendations. Radiat Prot Dosim 70(1–4): 383–386

Silari M, Agosteo S, Beck P, Bedogni R, Cale E, Caresana M, Domingo C, Donadille L, Dubourg N, Esposito A, Fehrenbacher G, Fernández F, Ferrarini M, Fiechtner A, Fuchs A, García MJ, Golnik N, Gutermuth F, Khurana S, Klages Th, Latocha M, Mares V, Mayer S, Radon T, Reithmeier H, Rollet S, Roos H, Rühm W, Sandri S, Schardt D, Simmer G, Spurny F, Trompier F, Villa-Grasa C, Weitzenegger E, Wiegel B, Wielunski M, Wissmann F, Zechner A, Zielczynski M (2009) Intercomparison of radiation protection devices in a high-energy stray neutron

field. Part III: instrument response. Radiat Meas 44:673–691
Tanner RJ, Molinos C, Roberts NJ, Bartlett DT, Hager LG, Jones LN, Taylor GC, Thomas DJ (2006) Practical implications of Neutron Survey Instrument Performance Report HPA-RPD-016. Health Protection Agency, Chilton, Didcot, United Kingdom
Thomas DJ (2004) Neutron spectrometry for radiation protection. Radiat Prot Dosim 110(1–4): 141–149
Thomas DJ, Alevra AV (2002) Bonner sphere spectrometers – a critical review. Nucl Instrum Methods Phys Res A Accel Spectrom Detec Assoc Equip 476(1–2):12–20
Thomas DJ, Klein H (2003) Introduction (to Handbook on neutron and photon spectrometry techniques for radiation protection - special issue) Radiat Prot Dosim 107(1–3): 13–21
Wiegel B, Alevra AV (2002) NEMUS – the PTB Neutron Multisphere Spectrometer: Bonner spheres and more. Nucl Instrum Methods Phys Res A 476:36–41
Wirtz K, Beckurts KH (1964) Neutron physics. Springer, Heidelberg

Further Reading

Anderson IS, McGreevy RL, Bilheux HZ (eds) (2009) Neutron imaging and applications – a reference for the imaging community. Springer, New York
Blomgren J, Lindborg L (eds) (2007) Tenth International Symposium on Neutron Dosimetry: progress in dosimetry of neutrons and light nuclei light nuclei, Uppsala, 12–16 June 2006, Radiation Protection Dosimetry v 126, nos 1–4. Oxford University Press, Oxford
Bottolier-Depois J-F, Beck P, Reitz G, Rühm W, Wissmann F (eds) (2009) Cosmic radiation and aircrew exposure, Radiation Protection Dosimetry v 136, no 4. Oxford University Press, Oxford
Grupen C, Shwartz B (2008) Particle detectors, 2nd edn. Cambridge University Press, Cambridge
Kiefer H, Maushart R (1972) Radiation protection measurement. Pergamon, Oxford
Klein H, Thomas D, Menzel HG, Curzio G, D'Errico F (eds) (2002) NEUSPEC 2000: proceedings of the International Workshop on Neutron Field Spectrometry in Science, Technology and Radiation Protection; Pisa, 4–8 June 2000, Nuclear instruments & methods in physics reseach, Section A, Accelerators, spectrometers, detectors and associated equipment; v. 476, no. 1–2. Elsevier, Amsterdam
Kleinknecht K (1998) Detectors for particle radiation, 2nd edn. Cambridge University Press, Cambridge
Leo RW (1994) Techniques for nuclear and particle physics experiments, 2nd revised edition. Springer, Berlin
Menzel HG, Chartier JL, Jahr R, Rannou A (eds) (1997) Neutron dosimetry: proceedings of the Eighth Symposium, Paris, 13–17 November 1995, Radiation Protection Dosimetry v 70, nos 1–4. Nuclear Technology, Ashford
Rózsa S (1987) Radiometrische Messungen in der Industrie – Grundlagen und Meßmethoden. Franzis-Verlag GmbH, München
Schmitz T, Waker AJ, Kliauga P, Zoetelief H (eds) (1995) Design, construction and use of tissue equivalent proportional counters, Radiation Protection Dosimetry v 61, no 4. Nuclear Technology, Ashford
Thomas DJ, Klein H (eds) (2003) Neutron and photon spectrometry and techniques for radiation protection, Radiation Protection Dosimetry v 107, nos 1–3. Nuclear Technology, Kent

Suppliers of Neutron Detectors

ALOKA Co. Ltd., 6-22-1 Mure, Mitaka-shi, Tokyo, 181-8622, Japan
Berthold Technologies GmbH & Co KG, Calmbacherstrasse 22, Bad Wildbad, 75323, Germany
BTI Bubble Technology Inc., Chalk River, Canada
Canberra Canberra Industries Inc., 800 Research Parkway, Meriden, CT, 06450, USA
Centronic Limited, Centronic House, King Henry's Drive Croydon CR9 0BG, United Kingdom
Framework Scientific, LLC
General Electric Company GE-Reuter-Stokes, 3135 Easton Turnpike Fairfield, CT, 06828-0001, USA

Innovative American Technology (IAT), 4800 Lyons Technology Park Drive, Coconut Creek, FL, 33073, USA

John Caunt Scientific Ltd., PO Box 1052, Oxford, OX2 6YE, United Kingdom

Laboratory Impex Systems Ltd, 15 Riverside Park, Wimborne, Dorset BH21 1 QU, United Kingdom

Landauer Inc., 2 Science Road, Glenwood, Illinois, 60425-1586, USA

LND, INC, Nuclear Radiation Detectors, 3230 Lawson Boulevard, Oceanside, NY, 11572, USA

Ludlum Measurements Inc., 501 Oak Street, Sweetwater, TX, 79556, USA

MIRION, Lieu-dit Calés Route d'Eguiès, F-13113 Lamanon, France

NucSafe Inc., 601 Oak Ridge Turnpike, Oak Ridge, TN 37830, USA http://www.nucsafe.com

Polimaster International, 112, M.Bogdanovich St., Minsk, 220040, Republic of Belarus

ROTEM, Rotem Industrial Park, Mishor Yamin, D.N Arava 86800, Israel

SAIC Inc., 1710 SAIC Drive, McLean, VA 22102, USA

Saint-Gobain Crystals, 7900 Great Lakes Pkwy, Hiram, OH, 44234-9681, USA

Scionix Holland B.V., Regulierenring 3, Bunnik, 3981 LA, The Netherlands, www.scionix.nl

TA Technical Associates, 7051 Eton Avenue Clanoga Park, CA, 912303, USA

Thermo Fisher Scientific Inc., 81 Wyman Street, Waltham, MA, 02454, USA http://www.thermo.com

Toshiba Electron Tubes & Devices Co., Ltd. 1385 Shimoishigami, Otawara, Tochigi, Japan

TSA Systems Ltd., 14000 Mead Street, Longmont, CO, 80504-9698 USA

32 Instrumentation for Nuclear Fusion

Rudolf Neu
Max-Planck-Institut für Plasmaphysik, Garching, Germany

1 Introduction . . . 793

2 Basic Nuclear Fusion . . . 794

3 Diagnostic of Fusion Plasma . . . 798

4 Radiation Measurements . . . 799
4.1 Thermography . . . 800
4.2 Continuum Radiation . . . 800
4.2.1 Bolometry . . . 800
4.2.2 Soft-X-Ray Diagnostic . . . 801
4.3 Line Radiation . . . 801
4.3.1 Passive Spectroscopy . . . 802
4.3.2 Charge-Exchange Spectroscopy . . . 804
4.3.3 γ Spectroscopy . . . 805
4.4 Particle Measurements . . . 805
4.4.1 Charge-Exchange Neutrals . . . 805
4.4.2 Neutron-Rate Measurements . . . 806
4.4.3 Neutron Spectroscopy . . . 806
4.4.4 Charged-Particle-Loss Diagnostic . . . 807

5 Special Requirements for ITER and Burning-Plasma Devices . . . 807
5.1 Spectroscopic Systems and Bolometry . . . 808
5.2 Fusion Products . . . 808

6 Conclusions and Outlook . . . 809

7 Cross-References . . . 809

Acknowledgments . . . 809

References . . . 810

C. Grupen, I. Buvat (eds.), *Handbook of Particle Detection and Imaging*,
DOI 10.1007/978-3-642-13271-1_32, © Springer-Verlag Berlin Heidelberg 2012

Further Reading 810

Major Fusion Devices 810

Abstract: To characterize a fusion plasma in an adequate way and to understand its complex behavior as complete as possible, a large number of different plasma parameters must be determined simultaneously. Most of them are local quantities varying with the radial coordinate and in time. The spatial resolution aimed at is 1–5 cm and the required time resolution varies from μs to ms depending on the measured quantity. The demanded accuracy of all the measuring systems is typically 1–10%. The major challenge for fusion plasma diagnostic originates from the harsh environment they are exposed to. When proceeding to future burning-plasma devices, these burdens will further increase mainly due to the strongly increased neutron fluence and energy load during the lifetime of the diagnostic. This contribution sketches basic nuclear fusion and concentrates on diagnostic areas where radiation detectors are involved, factoring out completely the wide field of electromagnetic measurements and laser-aided methods.

1 Introduction

Investigating hot magnetized plasmas in the frame of fusion research, a large number of different diagnostic methods must be applied. Common to almost all diagnostic techniques is that plasma probing must be conducted without any material contact between the measuring instrument and the plasma, due to the high temperatures and power fluxes involved. The physics principles underlying the various techniques originate from almost all fields of physics. They are complementing one another with respect to dynamic range or concerning certain information gained by other diagnostic systems necessary to interpret the measurements. The whole diagnostic system has redundancy which means that a given physical quantity is measured simultaneously by a number of different measuring systems. They are based on different physics principles, ensuring that systematic errors do not remain unrecognized. Although a large number of diagnostic methods are well developed and established, growing new fields of research demand for additional information, or higher spatial and temporal resolutions, stimulating the development of new methods and even more refined techniques. Typically, more than 50 different diagnostic systems are installed at modern large fusion experiments. In first instance, they provide the information necessary for the safe operation of the fusion experiment and the protection of its components. Moreover, because all present-day fusion devices are devoted to research, they constitute the backbone for the scientific work conducted.

To characterize the plasma in an adequate way and to understand its complex behavior as complete as possible, a large number of different plasma parameters must be determined simultaneously. These are the density of the electrons and the ions, the composition and density of certain impurities in the plasma, the temperature of the electrons and of the ions which can differ significantly depending on the heating scenarios conducted, the total stored energy, the plasma pressure, the radiation loss, currents and electric fields in the plasma and many others. Most of them are local quantities varying with the radial coordinate. Examples are the electron temperature and density being typically maximal at the plasma axis and approaching low values at the plasma edge varying over orders of magnitude. Most of the quantities are time dependent with characteristic times varying significantly. While fluctuating quantities are characterized by time scales of the order of 1 μs, certain currents in the plasma or the energy content vary slower by up to six orders of magnitude. The demanded accuracy of all the

diagnostic measuring systems is typically 1–10%. The spatial resolution aimed at is 1–5 cm and the time resolution of the order of μs. The following short introduction to the subject concentrates on the areas where radiation detectors are involved, factoring out completely the wide field of electromagnetic measurements and laser-aided methods. For a deeper insight into the physical principles of the described diagnostics, the reader should refer to Lochte-Holtgreven (1995) and Hutchinson (2002).

2 Basic Nuclear Fusion

Any energy production from nuclear reactions is based on differences in the nuclear binding energy. An explanation of the structure for the general trend of the nuclear binding energies was given by C. F. von Weizsäcker in 1935. Starting from the very limited range of the strong nuclear force, he assumed that each nucleon just influences its nearest neighbors. The binding energy per nucleon would thus be constant. The smaller binding energies for smaller nuclei are due to the relatively large surface-to-volume ratio. The nucleons at the surface have missing partners, and thus, their contribution to the total binding energy of the nucleus is reduced. The decrease of binding energy per nucleon for nuclei beyond $A \approx 60$ is due to the repulsive Coulomb force of the large amount of positive protons. The finer structures in the nuclear binding energy are due to quantum-mechanical effects, i.e., at certain so-called "magic" proton and neutron numbers, the nucleus formed is a very stable configuration. This is roughly comparable to the stable electron configurations of the noble gases, where electron shells are completed. The first magic number is 2, which is manifested as a most remarkable example of a local maximum for the helium nucleus with 2 protons and 2 neutrons. From the characteristics of the nuclear binding energy, it is clear that there are two ways of gaining nuclear energy: either by transforming heavy nuclei into medium-size nuclei (this is done by fission of uranium, which is described in Chap. 30, "Spallation – Neutrons Beyond Nuclear Fission") or by fusion of light nuclei into heavier ones. In particular, the fusion of hydrogen isotopes into stable helium offers the highest energy release per mass unit.

The main fusion reaction in the Sun starts from protons leading through intermediate steps to an α particle (He nucleus), two positrons, and two neutrinos and the release of the energy:

$$4\,\mathrm{p} \longrightarrow \alpha + 2\mathrm{e}^{+} + 2\nu_{\mathrm{e}} + 26.7\ \mathrm{MeV}.$$

Since the fusion reaction in the Sun involves the weak interaction to transform protons to neutrons, the reaction rates are far too low to allow energy production on a terrestrial scale. Possible candidates for using fusion energy on earth are the following reactions (D and T denoting the hydrogen isotopes deuterium and (the heaviest) tritium, with one and two neutrons, respectively):

$$\begin{aligned}
&\mathrm{D} + \mathrm{D} \longrightarrow {}^{3}\mathrm{He} + \mathrm{n} + 3.27\ \mathrm{MeV}\ (50\%)\\
&\qquad\quad \text{or}\quad \mathrm{T} + \mathrm{p} + 4.03\ \mathrm{MeV}\ (50\%),\\
&\mathrm{D} + {}^{3}\mathrm{He} \longrightarrow {}^{4}\mathrm{He} + \mathrm{p} + 18.35\ \mathrm{MeV},\\
&\mathrm{D} + \mathrm{T} \longrightarrow {}^{4}\mathrm{He} + \mathrm{n} + 17.59\ \mathrm{MeV},\\
&\mathrm{p} + {}^{11}\mathrm{B} \longrightarrow 3\,{}^{4}\mathrm{He} + 8.7\ \mathrm{MeV}.
\end{aligned}$$

The D–T reaction has by far the largest cross section at the lowest energies. This makes the D–T fusion process the most promising candidate for an energy-producing system. The special

role of D–T reactions can be explained by a resonance in the compound nucleus ^{5}He which has an excited state just 64 keV above the sum of the masses of deuterium and tritium. Since tritium is an unstable radioactive isotope, decaying to ^{3}He with a half-life of 12.3 years, it does not exist in significant amounts. Therefore, tritium has to be produced with nuclear reactions of the neutrons from the D–T reaction and lithium:

$$\begin{aligned} \mathrm{n} + {}^6\mathrm{Li} &\longrightarrow {}^4\mathrm{He} + \mathrm{T} \quad + 4.8\ \mathrm{MeV}, \\ \mathrm{n} + {}^7\mathrm{Li} &\longrightarrow {}^4\mathrm{He} + \mathrm{T} + \mathrm{n}' \quad - 2.5\ \mathrm{MeV}. \end{aligned}$$

Deuterium and lithium are very abundant and widespread in the earth's crust and ocean water.

Since the range of the nuclear force is of the order of the dimensions of the nuclei, the two colliding nuclei have to "touch" each other for a fusion reaction to occur. However, elastic Coulomb scattering has a much larger cross section, which leads rather to small-angle scattering than to the fusion reaction. A way of overcoming this problem is to confine a thermalized state, a so-called plasma of deuterium and tritium ions at energies of about 10 keV. Since the average energy of particles at a certain temperature is about kT, where k is the Boltzmann constant, temperatures are often given in units of electron volt (1 eV $\triangleq 1.16 \times 10^4$ K) in fusion research. If the plasma is confined long enough, the particles thermalize as a result of many Coulomb scattering processes and thus entail a Maxwellian velocity distribution:

$$f(v) = n \left(\frac{m}{2\pi k T} \right)^{3/2} \cdot \exp\left(-\frac{m v^2}{2\,k T} \right),$$

where f is the number of particles in the velocity interval between v and $v + \mathrm{d}v$, n is the density of particles, m is their mass, and kT is their temperature. The reaction rate per unit volume, R, can be written as

$$R = n_\mathrm{D} \cdot n_\mathrm{T} \cdot \langle \sigma v \rangle$$

with v now being the relative particle velocity and $\langle \sigma v \rangle$ being the reaction parameter, which is calculated by integration over the distribution function of deuterium and tritium yielding

$$\langle \sigma v \rangle = \frac{4}{(2\pi m_\mathrm{r})^{1/2} (kT)^{3/2}} \int \sigma(\varepsilon_\mathrm{r}) \cdot \varepsilon_\mathrm{r} \cdot \exp\left(-\frac{\varepsilon_\mathrm{r}}{kT} \right) \mathrm{d}\varepsilon_\mathrm{r},$$

where m_r is the reduced mass, and ε_r the relative kinetic energy. ❯ *Figure 1* shows the reaction parameter for some important fusion reactions. At temperatures of interest, the nuclear reactions result predominantly from the tail of the distribution.

A plasma can be confined by magnetic fields. In linear configurations, the end losses are by far too large to reach the necessary thermal insulation, which is described by the so-called energy confinement time τ_E. From power balance considerations – comparing the fusion power inherent in the α particles to losses by radiation and heat conduction – a value of τ_E on the order of several seconds is required. The end losses can be completely avoided in a toroidal system. However, in such a device with purely toroidal magnetic field, the field curvature and gradient result in a vertical drift which is in opposite directions for ions and electrons. The resulting electric field causes an outward $\mathbf{E} \times \mathbf{B}$ drift of the whole plasma, and therefore, such a simple magnetic field configuration will be unstable. To avoid this charge separation, it is necessary to twist the magnetic field lines by additional magnetic field components, where single field lines map out so-called flux surfaces. On these flux surfaces, plasma transport is fast, as it is always parallel to $\mathbf{B}$, and therefore, plasma parameters usually are constant on a given flux surface. Perpendicular to the flux surfaces, transport is hindered because particle motion perpendicular

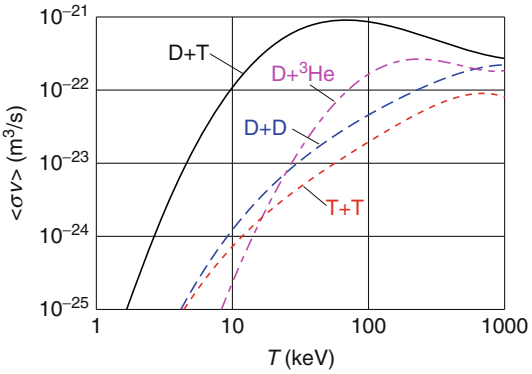


Fig. 1
Reaction parameter $\langle \sigma v \rangle$ as a function of ion temperature T_i for different fusion reactions (Bosch and Hale 1992)

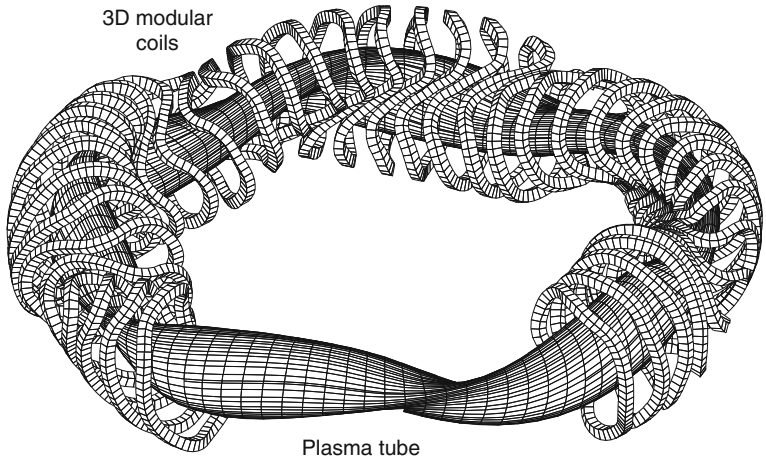


Fig. 2
Schematic view of the modular stellarator W7-X with non-planar coils

to **B** is restricted by the Lorentz force, and therefore, large radial temperature and density gradients can be maintained in such a configuration allowing to reach central plasma parameters necessary for fusion.

Two different principles for twisting the magnetic field lines are under investigation worldwide. The stellarator was invented in 1951 by Lyman Spitzer, Jr. in Princeton. Here, the twist of the field lines is created by external coils wound around the plasma torus. These external coils have the advantage that the current can be controlled from outside, and can flow continuously, but this kind of device is very challenging from the engineering point of view. *Figure 2* sketches a modern "modular" stellarator, where the planar toroidal coils and the helical coils of a classical stellarator have been replaced by one complex, but modular system of non-planar coils.

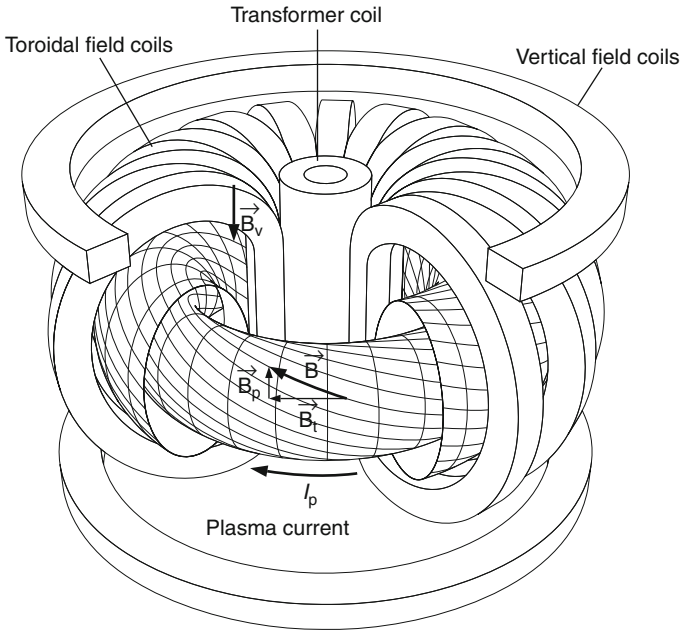


Fig. 3

Schematic view of a tokamak. The innermost cylindrical coil is the transformer coil for inducing a plasma current. The toroidal coils above and below the machine create a vertical field for plasma shaping and position control

The second approach is the tokamak proposed by two Russian physicists, Tamm and Sakharov, in the year 1952 and realized by Artsimovich. The word *tokamak* itself is derived from the Russian words for toroidal chamber with magnetic field. The tokamak concept is outlined in *Fig. 3*. The toroidal magnetic field is provided by external coils, and the necessary twist is produced by means of an electric current in the plasma itself which gives rise to the poloidal component of the twisted magnetic field. This current is produced by induction, the plasma acting as the secondary winding of a transformer. Tokamaks have proven to be very successful in improving the desired fusion plasma conditions and the today's most successful experiments are based on this principle. However, a transformer can induce the (DC-)plasma current only during a finite time, while, as mentioned before, a stellarator may principally run completely steady-state. For truly continuous tokamak operation, alternative current drive methods are being developed. Another disadvantage of the required large plasma current is the potential danger of so-called disruptions: Uncontrolled very fast (~1 ms) plasma current quenches which can give rise to large forces on the device. A recent review on the status of tokamak research is given in Wesson (2004).

In a future fusion reactor, the hot plasma will be surrounded by the first wall and the blanket. The latter is filled with lithium to produce the tritium, as discussed before. The majority of the thermal energy of the plant is delivered here by neutron moderation, because the neutrons carry four-fifth of the energy released in the D–T fusion reaction. A shield is provided

behind the blanket to stop the neutrons not captured by the blanket in order to reduce the heat and radiation loads to the cold structures of the superconducting magnets (the application of superconduction is mandatory for fusion reactors to obtain a positive energy balance).

3 Diagnostic of Fusion Plasma

To investigate in detail the behavior of fusion plasmas with their extreme parameters, a large variety of (partially complementary) diagnostic methods have been developed and are being used in present-day experiments. In general, the diagnostic systems of a fusion experiment constitute a major part of it, with regard to hardware as well as manpower. The principles of these methods come from all areas of physics. A selection of the ones relevant to the scope of this contribution is discussed in the following. The basic physics principles of plasma diagnostics are described extensively in Hutchinson (2002). In ❯ *Fig. 4*, the most important plasma parameters and the methods with which they are deduced are presented. ❯ *Figure 5* shows some principle plasma parameters of a typical discharge in the largest German tokamak ASDEX Upgrade. As described above, the plasma is only confined in a tokamak if a current is flowing. In the shown discharge (#25834), the current is ramped up to its final value of 1 MA until $t = 1.1$ s and is ramped down again at its termination after $t = 6.0$ s. The line (of sight) averaged electron density (same insert), which is measured by interferometry, reaches a maximum value of about 9×10^{19} m^{-3}, which is typical for fusion plasmas. The insert below gives the trajectories of the

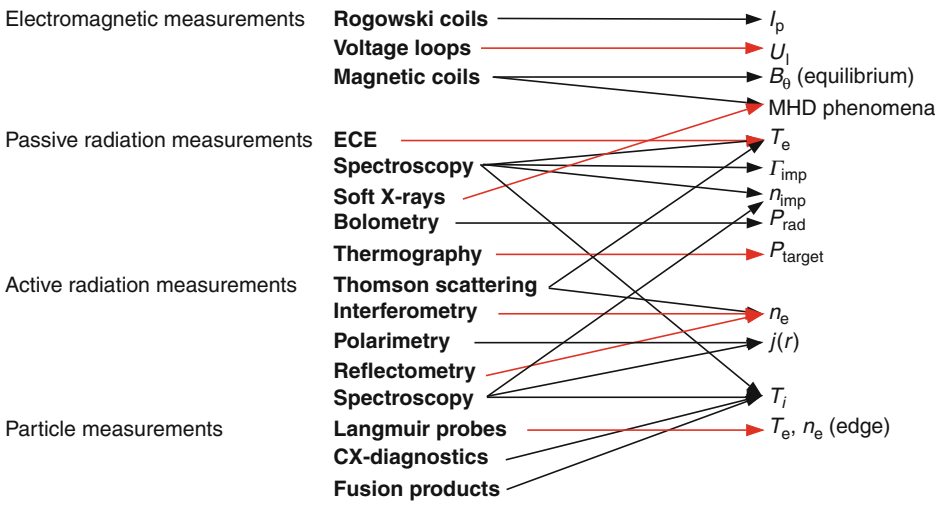


Fig. 4

Physics parameters of relevance in fusion plasmas (*right*) together with the diagnostic method used for their deduction (*left*). I_p, plasma current; U_l, loop voltage; B_θ, (poloidal) magnetic field; MHD, magneto-hydro-dynamics; T_i, T_e, ion and electron temperatures; n_e, n_{imp}, electron and impurities densities; P_{rad} radiated power, P_{target}, power load to the plasma-facing components, $j(r)$, plasma current density

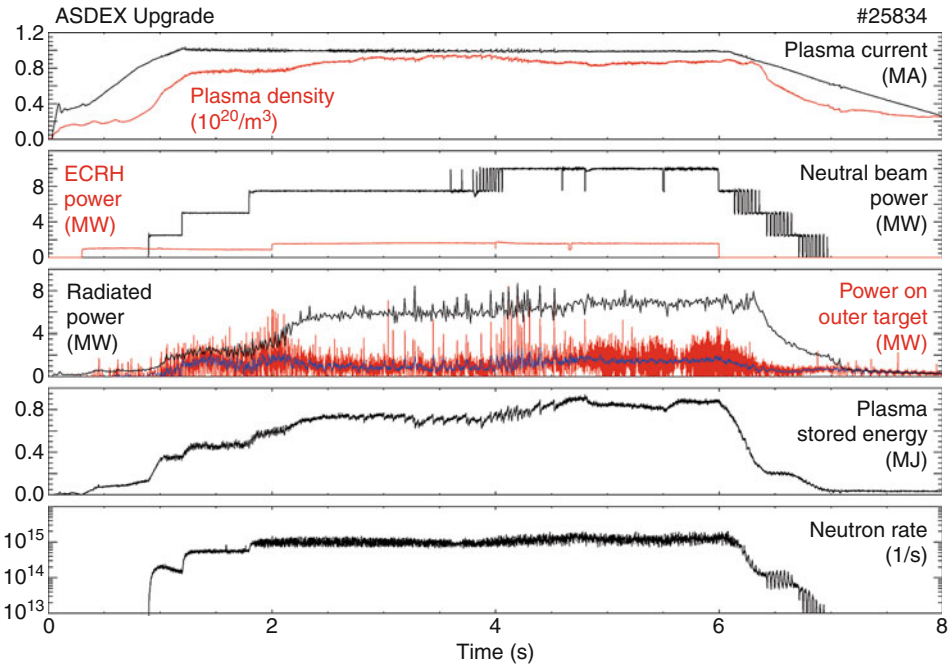


Fig. 5

Temporal evolution of several parameters of a tokamak discharge (for details see text)

auxiliary heating power. The discharge is heated by neutral beam injection (NBI) at a power of 10 MW and Electron Cyclotron Resonance Heating (ECRH) of about 2 MW. The next insert presents the power radiated from the plasma and the power convected/conducted to the outer so-called divertor target, where the main plasma–wall interaction appears. The sum of radiated and convected/conducted power must equal the total heating power – the missing part in the balance is conducted/convected to other plasma facing components. Note that there is no significant contribution from the fusion reaction in this plasma discharge since the experiment is performed within D only. The plasma stored energy is a measure for the confinement at a given heating power. Similarly, the neutron rate (lowest insert) also provides a rough measure of the central plasma parameters, as can be deduced from the equation on the neutron rate given above. Throughout the different phases of the discharge, the neutron rate varies by several orders of magnitude (logarithmic scale). The emission of radiation and particles from a fusion plasma can roughly be divided into two categories. The first one is due to plasma and to atomic physics processes as scattering, excitation, ionization, and recombination, the second category originates from (nuclear) fusion and subsequent processes. From the viewpoint of their technical implementation, a division in radiation and particle measurements seems to be more practical in the context of this handbook.

4 Radiation Measurements

Magnetically confined plasmas are optically thin, except for electron cyclotron radiation (≈100 GHz range) and in some cases for the hydrogen Lyman-α radiation. The radiation spectrum

ranges from the infrared to the X-ray region and contains a lot of information about the plasma composition. Different mechanisms for the emission of radiation can be distinguished:

- Thermal radiation from hot surfaces.
- Radiation from free electrons in the field of ions, so-called bremsstrahlung, or free–free radiation, since it is a transition between two unbound states of the electron, resulting in a continuous spectrum.
- Recombination radiation from free electrons captured by an ion (free–bound radiation). This spectrum is also continuous, but it contains edges resulting from electrons with negligible energy recombining into a specific energy level of the ion.
- Line radiation from bound electrons in an excited atom or an ion.
- Line radiation from nuclear reactions of fast particles in the plasma.

4.1 Thermography

Power handling is one of the primary functions and most challenging problems for tokamak plasma facing components (PFCs). Infrared thermography is an important tool that can quantify the surface temperature and the heat-load profiles. The surface temperature is deduced from the absolute measurement of the thermal radiation in the mid-to-long-wavelength infrared ≈3–8 μm taking into account the emissivity of the surface. As detectors, linear or two dimensional CMOS arrays of InSb or HgCdTe are used. The detectors are mounted outside the vacuum vessel, and the IR radiation is transmitted through sapphire or Ge optics. Temporal resolution in the sub-ms range has to be achieved to allow the measurement of fast plasma events depositing power to the PFCs. The IR camera measurements are benchmarked directly with two-color pyrometry or indirectly by thermocouples being built into the PFCs measuring the temperature increase of the whole component.

4.2 Continuum Radiation

Because the bremsstrahlung and recombination radiation spectra fall off exponentially with the frequency ν, the slope of this spectrum can be used to deduce T_e, without any need for an absolute calibration. Usually semiconductor detectors (sensitive in the keV range) are used with a pulse-height analysis system to measure spectra as sketched in ❯ *Fig. 6*.

The bremsstrahlung intensity depends on the ion species, and if different ion species are present in the plasma, all their contributions have to be added, and the intensity is proportional to the so-called effective charge Z_{eff}, where $Z_{\mathrm{eff}}^2 \approx \frac{\sum_i n_i Z_i^2}{n_e}$. A calibrated measurement can therefore be used to determine Z_{eff}, which is a good measure of the degree of cleanliness of the plasma and an important parameter for the fusion power output.

4.2.1 Bolometry

The energy loss from a plasma due to radiation is often a major contribution to the energy balance and is therefore an important parameter in describing the plasma (see ❯ *Fig. 5*). The

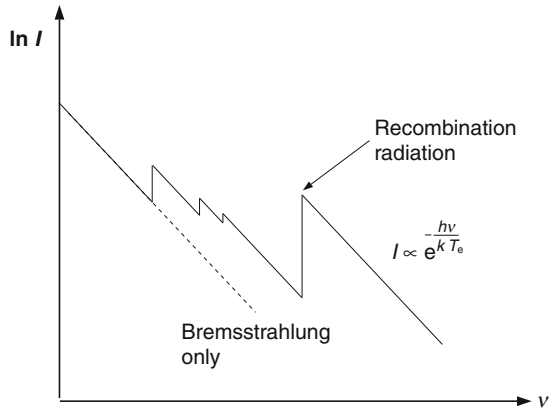


Fig. 6
Schematic view of the bremsstrahlung and recombination spectra from a fusion plasma

easiest way to measure the total radiation loss is to use a bolometer. This is a detector specifically designed to have a flat spectral response over a wide energy range, mainly in the UV region, where the main radiative energy loss occurs.

Usually, a bolometer consists of a thin foil (gold absorber on thin mica or kapton foils with gold meanders on their back side) that absorbs the energy. The temperature rise of this foil is then equal to the total energy flux divided by the bolometer's thermal capacity. This finite capacity, however, limits the time resolution of this technique to the order of ms.

4.2.2 Soft-X-Ray Diagnostic

A much better time resolution (fraction of μs) can be reached by using semiconductor detectors, which are sensitive in the soft X-ray region ($\hbar\omega \sim T_e$). The absolute calibration of these detectors is more problematic, because they change their sensitivity in time under neutron and plasma irradiation, but they are ideally suited to measure dynamic processes (such as magneto-hydrodynamic (MHD) instabilities) in the plasma.

Bolometers, as well as soft-X-ray detectors, measure line-integrated signals that do not allow spatial resolution. Because both detector elements can be built very small, usually a large number of them is used in a pinhole camera to get spatial resolution. If several such cameras view the plasma from different directions (see ❯ *Fig. 7*, left-hand side), the local emissivity in the plasma can be deduced from the line-integrated signals by an inversion algorithm (tomography).

4.3 Line Radiation

Line radiation from electrons bound in atoms or ions occurs at distinct wavelengths which are specific to this atom (or ion). Therefore, line radiation can in principle always be attributed to a

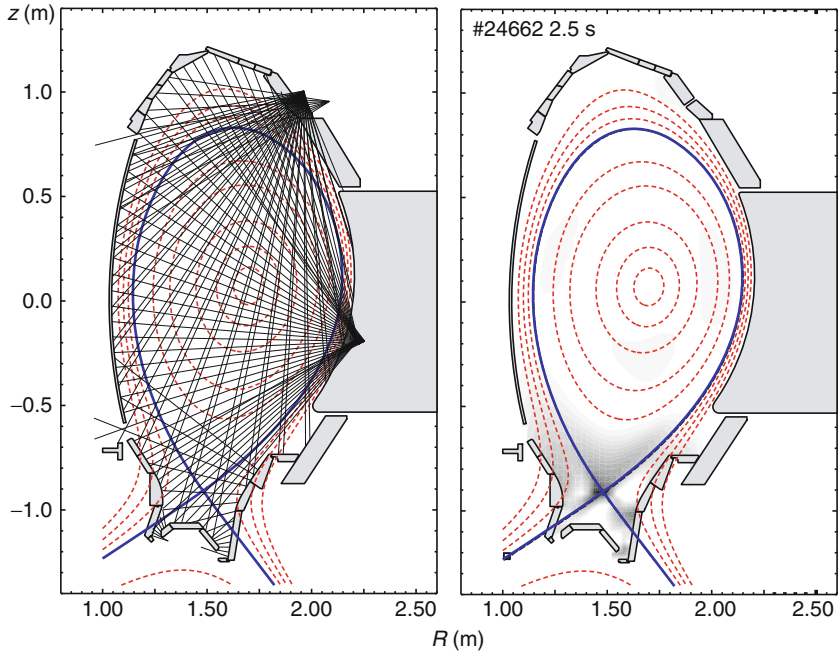


Fig. 7
Typical arrangement of lines of sight of a bolometer diagnostic (*left*) for tomographic reconstruction of the radiation profile (*right*) at ASDEX Upgrade

certain ion species. For this reason, spectroscopy is ideally suited to investigating the behavior of impurities in a fusion plasma. The main objectives are:

- Identification of the elements present in the plasma
- Measurement of the impurity influx from walls and plasma facing components from the line emission of excited atoms
- Determination of total impurity concentrations
- Investigations of transport processes by comparing measured impurity concentration profiles and their temporal evolution with transport models
- Determination of plasma parameters from line shapes and line ratios

4.3.1 Passive Spectroscopy

The electron temperature (T_e) is usually peaked in the center of the plasma. Since the charge state of ions strongly increases with T_e, the different ionization stages of an element in the plasma are usually ordered in a shell structure, as shown in ❯ *Fig. 8* for argon. For a central electron temperature of 1.3 keV ($\approx 1.5 \times 10^7$ K) as shown in the figure, all ionization states up to hydrogen-like Ar (Ar^{17+}) are abundant. Lighter species will be fully ionized and only heavy species (like metals) will be partially ionized (and therefore able to radiate) in the inner part of the plasma.

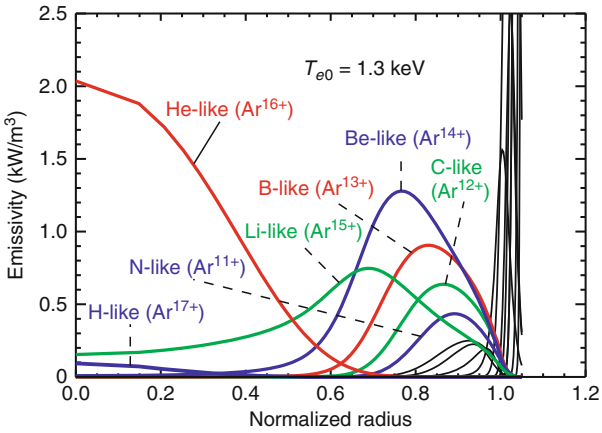


Fig. 8
Emissivity profiles of different Ar ions in an ASDEX Upgrade discharge with a central electron temperature of $T_e \approx 1.3$ keV

As the excitation energy strongly increases with the charge state ($\propto Z^2$), it becomes clear that plasma spectroscopy has to cover a wide range of wavelengths. For the investigation of excited atoms and low ionization stages at and around the plasma edge, visible lines can be used. In the plasma center, the highly charged ions emit X-rays, and in a large part of the plasma, vacuum ultraviolet lines are emitted.

To cover this whole range, different spectrometers involving quite different techniques have to be used, see also Bitter et al. (2007).

Visible and ultraviolet (700–200 nm)
Light can be transferred via glass fibers or normal optical components (for $\lambda \leq 350$ nm quartz has to be used), gratings or prisms are used to disperse the spectra and the light is detected with photomultipliers and modern CCD cameras. The spectral resolution can be made very high ($\lambda/\Delta\lambda \approx 10^4$).

Vacuum ultraviolet (VUV, 200–30 nm)
In this range, light is strongly absorbed even in air and can only be guided in evacuated tubes. Gratings can still be used, and windowless photomultipliers, scintillators, channeltrons or micro-channel plates with subsequent scintillators and CCD cameras are used for detection.

Soft X-rays (100–1 nm)
The detectors are the same as in the VUV range, but to enhance the reflectivity of the gratings, they are used at grazing angles of incidence (1–20°). The spectral resolution, however, is relatively low ($\lambda/\Delta\lambda \approx 200$–$2000$).

X-rays (2–0.1 nm)
In this spectral range, Bragg reflection on crystals is used to disperse the spectra. This results in a very good spectral resolution ($\lambda/\Delta\lambda \sim 10^4$). Beryllium windows can be used and air again becomes transparent. Scintillators and proportional counters are used as detectors.

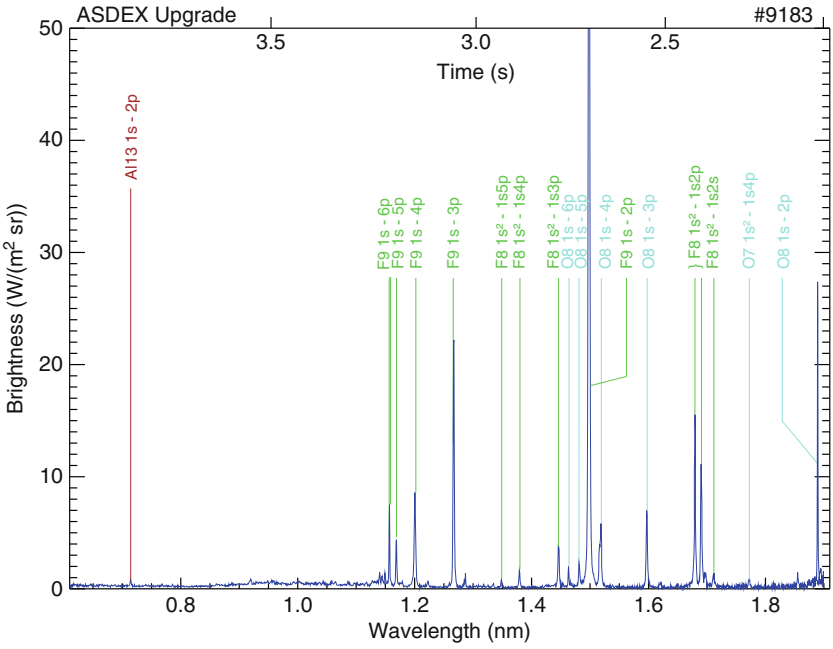


Fig. 9

X-ray spectrum of light He-like and H-like ion species emitted in discharge #9183 at ASDEX Upgrade. Please note, the spectrum is taken by a rotating crystal spectrometer, which means that each wavelength is measured at a different time during the discharge (time scale at the top)

For focussing devices (with toroidally or spherically bent crystals) position-sensitive detectors as multi-wire proportional counters, back-illuminated CCDs, CCDs with image intensifiers have to be employed. Recently, new pixel detectors became available (Pacella et al. 2007; Ince-Cushman et al. 2008). These detectors, either based on gas amplification or on semiconductor elements, allow the readout of single pixels with individual amplification and discrimination. *Figure 9* shows the spectrum of He-like and H-like O ions and F ions (an indication for H-like Al is also visible). The spectrum is taken by a Bragg crystal spectrometer with a rotating detector/crystal arrangement allowing to monitor a huge wavelength range during a plasma discharge.

4.3.2 Charge-Exchange Spectroscopy

In the previous subsection on line radiation, it was discussed that light ions radiate only at the plasma edge because they are fully ionized further inward. This problem can be overcome by injecting a beam of neutral atoms. In collisions of these beam atoms with the plasma ions, the latter can take over an electron (charge exchange) and subsequently emit radiation. Energy and angular-momentum conservation results in the excitation of high-lying Rydberg

levels, which emit visible radiation even from the heavier ion species. Usually the high-power heating beams (NBI heating, see ❯ *Fig. 5*) are also used for charge-exchange spectroscopy. This technique allows the measurement of the light-ion density in the plasma core. Since the radiation is emitted in the visible range, it can be detected with high spectral resolution (similar to "passive spectroscopy"), which allows the determination of the ion temperature T_i from Doppler measurements and of the plasma rotation.

4.3.3 γ Spectroscopy

Nuclear-reaction γ-ray diagnostic is a powerful technique used for studies of the fast ion behavior in fusion devices. Intense γ lines are emitted when fast ions (p, d, t, ^{3}He, and ^{4}He) react either with plasma fuel ions or with main plasma impurities, such as beryllium, boron, carbon, and oxygen. Intensity and 2-D source profiles from the ^{9}Be(^{3}He,nγ)^{11}C, ^{9}Be(^{3}He,pγ)^{11}B, and ^{12}C(^{3}He,pγ)^{14}N reactions are used in the European tokamak device JET as an indicator of the efficiency and spatial distribution of ion cyclotron resonance frequency (ICRF) power deposition in ^{3}He-minority heating scheme (Kiptily et al. 2002). To identify the existence of fusion α particles, the measurements of γ-rays from the reaction ^{9}Be(α,n γ)^{12}C can be used. The γ-ray measurements are performed in the energy range ≈1–20 MeV using NaI(Tl), BGO, and NE226 detectors with typical diameters of 5–12.5 cm (Krasilnikov et al. 2007).

4.4 Particle Measurements

4.4.1 Charge-Exchange Neutrals

The ions in the plasma are in general well confined by the magnetic fields, but in collisions with neutrals also present in the plasma, they can take over an electron and become a neutral atom that can leave the plasma. As in these charge-exchange collisions both particles keep their energy, the neutral atoms leaving the plasma represent the energy distribution of the plasma ions. Neutral particle analyzers consist of two major components, a stripping cell, where in a low-pressure gas the neutral particles are ionized again, and a magnetic and electrostatic energy analyzer to measure their energy spectrum. For plasma ions with a Maxwellian energy distribution, this spectrum falls off with energy E as $\exp(-E/T_i)$ and allows determination of T_i. The diagnostic is especially useful for the detection of deviations from the Maxwell distribution, e.g., due to additional heating of the plasma. However, it is a line-integrated measurement and a set of different lines of sight is necessary to unfold the measurements to determine an ion temperature profile $T_i(r)$. Furthermore, the density of the neutrals decreases toward the plasma center and is generally very low. Both issues can be overcome by using a beam of neutral atoms injected into the plasma. This increases the neutral atom density in general, and results in a localized source of the charge-exchange neutrals. With increasing size and density of the plasmas, the charge-exchange diagnostic becomes very problematic in respect to information about the plasma center since re-ionization of the atoms becomes large and strongly influences the spectrum of the neutral atoms leaving the plasma. For measurements near the plasma edge, it remains very valuable even for high-density plasmas.

4.4.2 Neutron-Rate Measurements

Usually, fusion plasmas nowadays work with hydrogen or deuterium, and therefore, fusion reactions between deuterium ions (D) can already occur via two different reaction branches, thereby creating highly energetic protons (p), tritium nuclei, t, ^{3}He nuclei, and neutrons (n) (see ❯ Sect. 2). Neutrons easily escape from any plasma and have been used as a diagnostic tool from the very beginning of fusion research. Since neutrons do not suffer from collisions in the plasma, they carry the full information about the fusion process. The reaction parameter $\langle \sigma(v) \cdot v \rangle$ (see ❯ Sect. 2) is a strong function of T_i. Therefore, the absolute measurement of the neutron rate is a direct measurement of T_i, provided the deuterium density is known. As most of the fusion reactions come from the ions in the high-energy tail of the velocity distribution, this method works only for Maxwellian plasmas. Small deviations from the Maxwellian energy distribution (i.e., even a small fraction of suprathermal particles in the tail of the distribution) can significantly alter the fusion reaction rate.

In thermal plasmas, the neutron emission rate of 1 s^{-1} is equivalent to the fusion power of 1.06×10^{-12} W in a D–T plasma and 0.44×10^{-12} W in a D–D plasma, respectively. Therefore, it will be used as a feedback parameter for fusion output control in a future fusion reactor and it has already been employed for control in JET (Jarvis et al. 1990) and TFTR (Hendel et al. 1990) tokamaks. The neutron rate is typically measured outside the vacuum vessel, and the neutrons have therefore to penetrate the vessel walls and other structures. To make the measurements independent of the spectrum, they are further moderated (for example, with polyethylene) and detectors which are only sensitive to thermal neutrons are used. Performed in such a way, the measurement is quasi volume-integrated. Since the neutron emission profile is strongly peaked, it can be assumed to be a toroidal line source, and the neutron detectors can be absolutely calibrated for this case by moving a neutron source along the plasma axis in the vacuum vessel. For absolute calibration of DD neutron measurements, typically a ^{252}Cf neutron source is used which is occasionally checked by activation measurements. The detectors must have a wide dynamic range, fast response time, and must not be sensitive to hard X-rays and γ-rays. Often a stack of detectors with different sensitivities is used. BF_3 and ^{3}He proportional counters, ^{235}U and threshold ^{238}U fission chambers are the most common neutron detectors having a time response on the order of ≈100 ns. The fission chambers are operated in counting, Campbell, and current modes. In addition, semiconductor Si and natural diamond detectors are also successfully used in counting mode for neutron flux monitoring.

4.4.3 Neutron Spectroscopy

Spectra of unscattered neutrons reflect the velocity distributions of the reacting fuel ions. For a plasma with ion temperature T_i, this results in a Gaussian-shaped neutron energy spectrum with an energy width of the neutron peak $\sim T_i$. Plasmas with essential fast ion component due to auxiliary heating by NBI or ICRH will produce neutron spectra with non-Gaussian tails and Doppler energy shift. Fast confined α particles can give rise to alpha knock-on neutron emission in DT plasmas with a high-energy tail. Therefore, neutron spectroscopy provides information of the following plasma parameters: ion temperature, reaction rates, fusion powers for D–D and D–T reaction components, fuel ion densities in the core, plasma rotations, relative densities of super-thermal ions and their energy distributions, density and shape of fusion α-particle energy distribution. Different types of neutron spectrometers are used in present-day fusion devices.

Magnetic Proton Recoil (MPR) and Time of Flight for Optimized Rate (TOFOR) large neutron spectrometers are successfully used in JET for detailed high-count-rate neutron spectrometry (temporal resolution ≈50 ms). Natural diamond, stilbene, and NE-213 (see, for example, Zimbal et al. (2007)) compact neutron spectrometers also are successfully applied for neutron spectrometry at other fusion devices (temporal resolution ≈300–400 ms). Typical energy resolutions achieved for all the spectrometers is in the range of 2–3% (Krasilnikov et al. 2007 and references therein). Different to neutron-rate measurements described in the previous paragraph a direct view to the plasma has to be provided with as little as possible absorbing and scattering material in the line of sight. The collimation of the spectrometers, however, requires strong shielding around the detector to restrict its view into the plasma to the line of sight.

4.4.4 Charged-Particle-Loss Diagnostic

Besides the neutron measurements also the measurement of charged-fusion-product losses provides important information on the source rate and profiles as well as the loss mechanisms. Since the production of a sufficient population of alpha particles (e.g., in DT) is limited in present-day devices, it has long been realized that the 3 MeV proton and the 1 MeV triton, produced in one of the DD fusion branches, should possess very similar single-particle behavior, since both gyroradius and slowing-down time are very similar to the 3.5 MeV αs. In the attempt to understand their behavior, new techniques were developed to study directly the confinement of charged fusion products (CFPs) (including energetic ions such as beam and RF-tail ions) primarily through the measurement of particle flux to the first wall. The most common detector systems for these charged-particle measurements are Faraday foil collectors and scintillator probes. In the Faraday foil detectors, the particles are detected as current to ground, whereas in scintillation detectors, the light is transported through fiber guides to externally mounted fast CCD cameras or photomultipliers. The combination of the entrance slit at the diagnostic head with the rather large magnetic field of the fusion devices acts as a velocity analyzer.

5 Special Requirements for ITER and Burning-Plasma Devices

Depending on their location in a burning-plasma device, diagnostics will experience different levels of radiation, nuclear heating, etc. For ITER, which will be the first fusion device which should demonstrate a burning DT plasma, the highest level will be experienced by components near the first wall where the neutron-flux levels will be up to 3×10^{18} n m^{-2}s^{-1}, the dose rate up to 2×10^3 Gy s^{-1}, and the plasma radiation up to 500 kW m^{-2}. The neutron and radiation-dose levels are typically some orders of magnitude higher than the maximum reached on present machines. The neutron heating will be typically 1 MW m^{-3} compared to essentially zero on existing machines. Relative to the conditions in present machines, probably the most significant extrapolation is in the pulse length: The pulse lengths will be up to several thousand seconds, that is, 10–1000 times higher than that typical on present-day machines. Combined with the higher flux levels and planned high number of discharges, this means that the end-of-life fluence levels will be about 10^5 times higher (Costley et al. 2005). As a consequence, many phenomena new to diagnostic design can occur and have to be taken into account, for example, radiation-induced conductivity and radiation-induced absorption and luminescence

in optical materials. Further, the nuclear environment sets stringent demands on the engineering of the diagnostic systems, as for example containment of tritium and vacuum integrity, ability to withstand high pressures, minimization of activation and remote handling, and maintenance.

5.1 Spectroscopic Systems and Bolometry

An extensive array of spectroscopic instrumentation will be installed in ITER covering the visible-to-X-ray wavelength range. Both passive and active measurement techniques will be employed. A common design requirement for the optical systems in the visible and infrared (IR) regions is to provide access with high optical throughput while maintaining neutron shielding. This is achieved by using labyrinths in shielding blocks mounted in the ports. At least the first few elements of the optical systems have to be mirrors because the high levels of radiation lead to enhanced absorption in refractive components. Few existing tokamak diagnostics incorporate in vacuo mirrors and the complexity in ITER of maintaining optical stability in a nuclear heated system, using a water coolant prescribed for the massive bulk heat removal of the blanket, significantly impacts the engineering design. The mirrors that face the plasma can suffer both erosion and deposition depending on their location and the plasma conditions. The erosion arises from bombardment by energetic ions and atoms formed through the charge-exchange process while the deposition is due to the erosion of nearby first wall material. The maintenance of the performance of the first mirrors is a major challenge for the implementation of the optical systems.

The vacuum ultraviolet spectrometers for the main plasma impurities require direct coupling to the tokamak since windows are not available for this wavelength range. For these systems, the radiation has to be collected through apertures in the blanket shield modules and conducted through straight pipes to the vacuum-sealed instruments. Each pipe is fitted with an isolation valve and means to ensure venting into an enclosure/vault capable of withstanding high pressures.

Owing to the high neutron fluence, gold-based bolometers used in present-day devices will fail because of the transmutation of Au to Hg and the embrittlement of the mica or kapton foils. Materials reported to be more radiation resistant are Pt and SiN. Thus, the detector development for ITER currently focuses on Pt absorbers on SiN membranes (Meister et al. 2010).

5.2 Fusion Products

In order to demonstrate the generation of fusion power in ITER, it will be necessary to measure the total neutron source strength with high accuracy and high reliability. In order to make these measurements, several fusion-product diagnostics are planned: internal and external neutron flux monitors, radial- and vertical-viewing neutron cameras, neutron spectrometers, neutron activation systems, and lost-alpha detectors. The implementation of flux detectors both inside and outside the vacuum vessel is not expected to pose significant problems although those mounted inside the vacuum vessel will potentially be susceptible to cable problems and adequate precautions have to be taken. Also, the presence of fissile material must be properly managed. In order to make an accurate measurement of the total neutron source strength, measurements of all regions, where significant fusion reactions are occurring, are needed.

However, because of the long ports on ITER, only the central region of the plasma can be readily probed with collimators and detectors, which have good access for repair, replacement, and calibration. In order to determine the spatial profile of the neutron source, it is necessary to measure the emission along multiple lines of sight in two different directions so that tomographic reconstruction can be carried out.

Various instabilities in the plasma could give rise to a loss of fusion alpha particles with consequential reduction in the fusion heating and so diagnostics capable of measuring lost alphas are required. The use of Faraday cup detectors and scintillator probes, as described above, could be difficult in ITER, because of disturbing background signals in the case of the scintillators and neutron-induced noise and cable problems in the case of the Faraday detectors.

6 Conclusions and Outlook

This short overview of plasma diagnostics has tried to treat the basic plasma diagnostic techniques (i.e., the most commonly used ones), and therefore, a lot of specialized techniques could not been mentioned at all. Great efforts are still being made to develop new techniques and improve existing ones. In the next major device, ITER, an extensive diagnostic system is still required. While the physics of the operation of the diagnostics, as established on past and present machines is, in many cases, directly applicable to ITER, the technology and engineering of the implementation of the systems differ substantially and involve many challenges. They arise from several different aspects and especially the unavoidable location of many of the diagnostic components inside the vacuum vessel and in the ports where they are subject to high levels of radiation and heating. This means that many phenomena new to diagnostic design can occur and have to be considered. In addition, the designs of the diagnostic systems have to meet stringent requirements associated with the nuclear environment and allow full remote control maintenance and repair.

A lot of details on processes and measurement techniques used in the instrumentation for nuclear fusion can be found under "Cross-References," see below.

7 Cross-References

- Chapter 9, "Calibration of Radioactive Sources"
- Chapter 12, "Tracking Detectors"
- Chapter 15, "Scintillation Counters"
- Chapter 16, "Semiconductor Counters"
- Chapter 26, "Accelerator Mass Spectrometry and Its Applications in Archaeology, Geology and Environmental Research"
- Chapter 29, "Particle Detectors in Materials Science"
- Chapter 30, "Spallation – Neutrons Beyond Nuclear Fission"

Acknowledgments

The author wants to thank the ASDEX Upgrade team for supplying sample data and figures.

References

Bitter M, Hill K, Scott S, Paul S, Ince-Cushman A et al (2007) AIP conference proceedings, vol 988. Melville, New York, pp 155–164

Bosch H-S, Hale GM (1992) Improved formulas for fusion cross-sections and thermal reactivities. Nucl Fusion 32(4):611–631, Erratum (1993) 33(12):1919

Costley A, Sugie T, Vayakis G, Walker C (2005) Technological challenges of ITER diagnostics. Fusion Eng Des 74:109–119

Hendel HW, Palladino RW, Barnes CW et al (1990) In situ calibration of TFTR neutron detectors. Rev Sci Instrum 61:1900

Hutchinson I (2002) Principles of plasma diagnostics. Cambridge University Press, Cambridge

Ince-Cushman A, Rice JE, Bitter M, Reinke ML, Hill KW et al (2008) Spatially resolved high resolution x-ray spectroscopy for magnetically confined fusion plasmas. Rev Sci Instrum 79: 10E302

Jarvis O, Sadler G, van Bell P, Elevant T (1990) In-vessel calibration of the JET neutron monitors using a ^{252}Cf neutron source: Difficulties experienced. Rev Sci Instrum 61:3172

Kiptily V et al (2002) γ-ray diagnostics of energetic ions in JET. Nucl Fusion 42:999

Krasilnikov A, Sasao M, Kaschuck Y, Kiptily V, Nishitani T et al (2007) AIP conference proceedings, vol 988. Melville, New York, pp 249–258

Lochte-Holtgreven W (1995) Plasma diagnostics. AIP, New York

Meister H, Eich T, Endstrasser N, Giannone L, Kannamuller M et al (2010) Optimization of a bolometer detector for ITER based on Pt absorber on SiN membrane. Rev Sci Instrum 81(10):10E132

Pacella D, Romano A, Pizzicaroli G, Gabellieri L, Bellazzini R et al (2007) AIP conference proceedings, vol 988. Melville, New York, pp 197–200

Wesson J (2004) Tokamaks, 3rd edn. Clarendon Press, Oxford

Zimbal A, Reginatto M, Schuhmacher H (2007) AIP conference proceedings, vol 988. Melville, New York, pp 323–326

Further Reading

Chen FF (2010) Introduction to plasma physics and controlled fusion. Springer-Verlag, New York

Griem H (2005) Principles of plasma spectroscopy. Cambridge University Press, Cambridge

Kunze HJ (2009) Introduction to plasma spectroscopy. Springer, Heidelberg

Orsitto FP et al (eds) (2008) Burning plasma diagnostics. In: AIP conference proceedings, vol 988. Melville, New York

Piel A (2010) Plasma physics: an introduction to laboratory, space, and fusion plasmas. Springer, Berlin

Major Fusion Devices

ALCATOR-C tokamak, MIT Plasma Science and Fusion Center, U.S.A. http://www.psfc.mit.edu/cmod/. Accessed 27 February 2011

ASDEX Upgrade tokamak, MPI für Plasmaphysik Garching, Germany. http://www.ipp.mpg.de/ippcms/eng/for/projekte/asdex/index.html. Accessed 27 February 2011

DIII-D tokamak, General Atomic, Fusion Group, U.S.A. http://fusion.gat.com/global/DIII-D. Accessed 27 February 2011

EAST tokamak, Institute of Plasma Physics, Academia Sinica, Hefei, China. http://east.ipp.ac.cn/ENGLISH/HT-7U.htm. Accessed 27 February 2011

ITER, International Collaboration, Aix en Provence, France. http://www.iter.org/. Accessed 27 February 2011

JET - Joint European Torus, Eur. Fusion Development Agreement, hosted by Culham CCFE, UK. http://www.jet.efda.org/. Accessed 27 February 2011

JT-60U tokamak, JAEA, Naka, Japan. http://www-jt60.naka.jaea.go.jp/english/index-e.html. Accessed 27 February 2011

KSTAR tokamak, NFRI, Daejeon, Korea. http://www.nfri.re.kr/english/research/kstaroperation01.php. Accessed 27 February 2011

LHD stellarator, NIFS, Toki, Japan. http://www.lhd.nifs.ac.jp/en/lhd/. Accessed 27 February 2011

MAST spherical tokamak, CCFE, Culham, United Kingdom. http://www.fusion.org.uk/MAST.aspx. Accessed 27 February 2011

NSTX spherical tokamak, Princeton Plasma Physics Laboratory, U.S.A. http://www.pppl.gov/nationalsphericaltorus.cfm. Accessed 27 February 2011

Tore Supra tokamak, CEA, Fusion Site at Cadarache, France. http://www-drfc.cea.fr/cea/ts/description/ts_description01.htm. Accessed 27 February 2011

Wendelstein 7-X stellarator, MPI für Plasmaphysik Greifswald, Germany. http://www.ipp.mpg.de/ippcms/eng/for/projekte/w7x/index.html. Accessed 27 February 2011

33 The Use of Neutron Technology in Archaeological and Cultural Heritage Research

Dudley Creagh
University of Canberra, Canberra, Australia

1 Introduction 814

2 Technology 814
2.1 Neutron Sources 815
2.2 Detectors 817

3 Neutron Interactions and Neutron Scattering Techniques 818
3.1 Coherent Neutron Scattering 818
3.1.1 Small-Angle Neutron Scattering (SANS) 819
3.1.2 Neutron Reflectometry and Grazing-Incidence Diffraction (GID) 820
3.1.3 Neutron Diffraction 822
3.2 Incoherent Scattering (Neutron Activation Analysis (NAA)) 824
3.3 Total Neutron Scattering: Imaging 825
3.3.1 Two-Dimensional Imaging 825
3.3.2 Three-Dimensional Imaging: Computed Tomography 826

4 Selected Applications 828

5 Conclusions 830

Acknowledgments 830

References 830

C. Grupen, I. Buvat (eds.), *Handbook of Particle Detection and Imaging*,
DOI 10.1007/978-3-642-13271-1_33, © Springer-Verlag Berlin Heidelberg 2012

Abstract: Nations define themselves by their history and their customs. Their history is determined by both archaeological and archival evidence. The continuing development of a national culture is essential for the formation of a national identity. Both archaeological sites and cultural heritage artifacts are important to many nations because of income earned through tourism. This chapter discusses the use of neutron technology, one of a number of possible technologies, in the study of archaeological and cultural heritage artifacts. In particular descriptions of Neutron Activation Analysis, Neutron Diffraction, and Neutron Imaging Techniques will be given, and selected applications of these techniques to archaeology and cultural heritage artifacts will be given.

1 Introduction

Archaeological sites and cultural heritage artifacts are of great importance for the defining of the characteristics of nations to their people. Viewed from another perspective, travel to countries by tourists interested in viewing national monuments, national buildings, national museums, and the contents thereof brings much needed stimulus to national economies. For some nations, tourism contributes substantially to the creation of national wealth.

To conserve and preserve the cultural heritage of nations effectively, it is necessary to have in place appropriate systems of conservation practice. In the conservation of artifacts it is of paramount importance that the structure and composition of the artifact is thoroughly understood. Then, and only then, it is possible to formulate appropriate conservation and preservation strategies.

Conservation scientists adhere to the basic principle "primum non nocere": first, do no harm. Therefore, the rule is to subject whatever is to be conserved to noninvasive, nondestructive testing. They use many techniques to study the objects. Radiation sources employed include: electromagnetic (mm-Wave, THz, IR, X-ray, γ-ray); electron beam (electron microscopy in all its forms: transmission (TEM), scanning (SEM), convergent beam electron diffraction (CBED), used in conjunction with energy-dispersive X-ray fluorescence (EDXRF), and Electron Energy Loss Spectroscopy); particle beam X-ray emission (PIXE), particle beam γ-ray emission (PIGE), and accelerator mass spectrometry (AMS). How these techniques are applied in practice is described in several texts (Creagh and Bradley 2000, 2006, 2007).

In these cited reference books the role of neutrons in archaeological and cultural heritage studies received little attention. The sole contribution was by Kockelmann et al. (2000). It must be stressed that the solution of problems in conservation science usually requires the use of several complementary techniques, in their case, synchrotron radiation, see, for example, Creagh (2007).

This chapter deals with the use of several techniques that use neutron technology to solve problems in cultural heritage science: neutron activation analysis (NAA), neutron diffraction (ND), and neutron tomography (NDT).

2 Technology

The use of neutrons for experimentation is usually more tightly controlled than other radiation types. Neutrons interact strongly with hydrogenous material, and have the capacity to cause severe damage in organisms. (See Chap. 10, "Radiation Protection"; Chap. 22, "Radiation

Damage Effects"). Systems that use neutron radiation sources, therefore, must be well shielded against radiation leakage. Neutron sources may be: nuclear reactors, spallation sources, members of the actinide family of the periodic table that spontaneously emit neutrons, composite sources (these comprise a particle emitter and a beryllium target: ${}^{4}_{2}\mathrm{He} + {}^{9}_{4}\mathrm{Be} \rightarrow {}^{12}_{6}\mathrm{C} + {}^{1}_{0}\mathrm{n}$), and sealed gaseous discharge tubes (Chap. 25, "Technology for Border Security") (ionized deuterium ($\mathrm{D} = {}^{2}_{1}\mathrm{H}$) is accelerated into a target loaded with tritium ($\mathrm{T} = {}^{3}_{1}\mathrm{H}$): $\mathrm{D} + \mathrm{T} \rightarrow {}^{4}_{2}\mathrm{He} + {}^{1}_{0}\mathrm{n}$). The neutrons have energies typically in the range 1 meV ("thermal neutrons") to 14 keV ("fast neutrons").

General techniques for the detection of neutrons are discussed by Klett (Chap. 31, "Neutron Detection"). Detectors vary considerably in energy response, linearity of response with count rate, energy resolution, spatial resolution, size, and complexity of the readout systems.

In this chapter, a brief discussion will be given of detectors that are used in imaging systems.

2.1 Neutron Sources

The nuclear reactor is the most commonly used neutron source in research. Nuclear reactors are devices in which chain reactions are initiated, controlled, and sustained in a stable condition. There are many types and designs of nuclear reactors (Lewis 2008; for a summary of types see *Nuclear Reactor Technology*: http://en.wikipedia.org/wiki/Nuclear_Reactor_Technology). It is not feasible here even to summarize this information. Spontaneous fission (spontaneous fission is a form of radioactive decay, possible only for elements with atomic masses greater than 230, in which the nucleus divides into two parts of approximately equal mass, with the simultaneous emission of a neutron) is possible in atoms such as ${}^{235}_{92}\mathrm{U}$, ${}^{239}_{94}\mathrm{Pu}$, or ${}^{242}_{94}\mathrm{Pu}$. Thus there is, in a rod of fissionable material, always a flux of neutrons that can interact with nuclei to cause a reaction producing two or more neutrons (one possible reaction is: ${}^{1}_{0}\mathrm{n} + {}^{235}_{92}\mathrm{U} \rightarrow {}^{236}_{92}\mathrm{U} \rightarrow {}^{141}_{56}\mathrm{Ba} + {}^{92}_{36}\mathrm{Kr} + 2{}^{1}_{0}\mathrm{n}$). These neutrons could interact with other atoms causing an increase in the neutron flux. But, effective absorption of the neutrons occurs at low neutron energies, so the neutrons have to be slowed down ("moderated"). This can be done in a number of ways. In the reactor (OPAL Reactor, Ansto, Sydney, Australia: http://www.ansto.gov.au/OPAL) shown in the schematic diagram (*Fig. 1*), cooling water flowing through the fuel assemblies in the core and heavy water ($\mathrm{D_2O}$) in the reflector vessel surrounding the core provide moderation of the neutron energy.

The core is $(0.35 \times 0.35 \times 0.6)\ \mathrm{m}^3$ in size and contains 16 fuel elements, each $(0.08 \times 0.08\ \mathrm{m}^2)$ cross section of 20% ${}^{235}_{92}\mathrm{U}$ enriched uranium. The neutron flux is controlled by the raising and lowering of five hafnium control rods.

The small size of the reactor core maximises the flux of neutrons available for the irradiation of materials and research. Neutron guide tubes set into the reactor enable the neutrons to be directed to equipment, which can be used for a wide range of scientific purposes.

In this chapter the use of techniques such as neutron activation analysis (NAA), neutron diffraction (ND), reflectometry (NR), grazing-incidence diffraction (GID), small-angle scattering (SANS), and imaging (NDT) will be described. Applications of these techniques in the study of items of archaeological and cultural heritage significance will be given in a later section (see, also, the web site: http://ancient-charm.neutron-eu.net/ach/).

Nuclear reactors provide a temporally constant flux of moderated neutrons with an approximately Maxwellian distribution of energies of a mean energy of about 25 meV. The energy of the neutron is related to its wavelength ($E = 81.8/\lambda^2$; here E is in units of meV and λ is in Å). Useful flux is delivered over the wavelength range 0.5–5 Å.

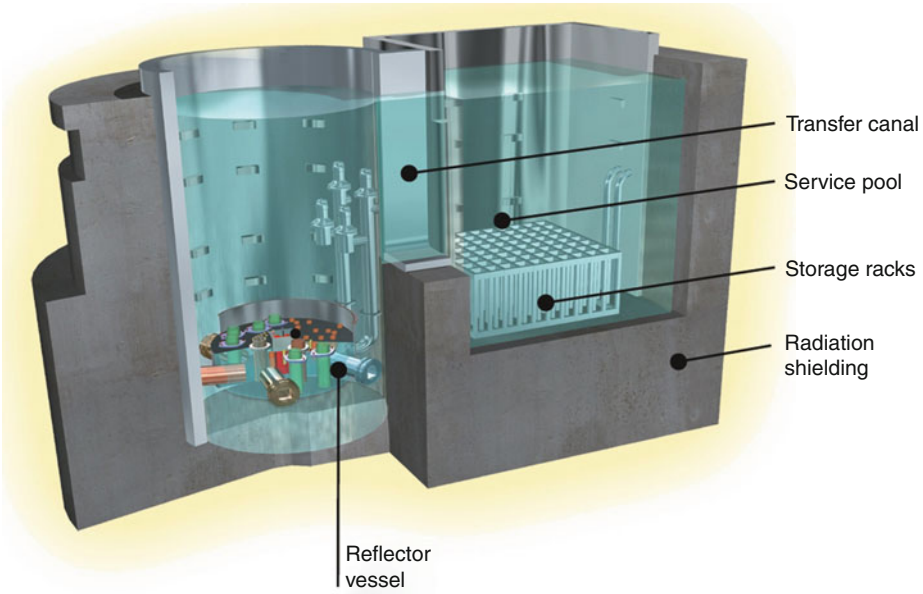


Fig. 1
Schematic view of the OPAL reactor

Another source of high-intensity neutron flux is the neutron spallation source (spallation is the process whereby an energetic particle causes, in its interaction with a nucleus, the emission of a number of particles of similar size, see also Chap. 30, "Spallation – Neutrons Beyond Nuclear Fission"). Spallation sources are particle accelerators in which negatively charged hydrogen ions are accelerated in a linear accelerator (LINAC: see Burkhardt, Chap. 7, "Accelerators for Particle Physics") to energies of ~1 GeV. The ions are stripped of their electrons and concentrated into highly intense (~2 MW) pulses of protons of ~1 μs duration and frequency ~50 Hz. These then interact with a target containing heavy elements (the UK facility (ISIS) at Daresbury uses tantalum; the US facility at Oakridge (SNS) uses mercury), and neutrons are generated by spallation. After moderation these neutron pulses can then be used in the same types of scientific experiments as those at nuclear reactors.

Spallation sources provide the neutron flux as high-intensity pulses, with pulses of 1–50 μs duration being provided at separations of 10–100 ms. Measurements are taken using time-of-flight (TOF) techniques. The desired energy band in a neutron bunch is selected using rotating mechanical shutters ("choppers"). The mean velocity of the neutron bunch is related to its wavelength by the de Broglie equation ($\lambda = h/(mv) = h/\sqrt{2mE}$), where h is Planck's constant, λ the neutron wavelength, E its energy, and v its velocity.

Each pulse contains a distribution of neutron velocities, the corresponding wavelength range being 0.4–4 Å, with a mean of ~1 Å. Traveling to the detector takes a different time for each neutron wavelength. Separation of the wavelengths is effected by plotting a graph of counts versus time.

Chap. 1, "Interactions of Particles and Radiation with Matter," Chap. 30, "Spallation – Neutrons Beyond Nuclear Fission," and Chap. 31, "Neutron Detection" will discuss the types of neutron interaction experiments that can be undertaken with a spallation source.

Of the other sources, composite sources are used to study the composition of materials in confined spaces, such as the internal walls of an oil well. The sealed tube system using the D–T reaction has similar uses. These are applications of the neutron activation analysis. As well, the D–T source has been used in imaging systems (Eberhardt et al. 2005).

2.2 Detectors

A detailed description of the neutron detectors will be given elsewhere (Chap. 31, "Neutron Detection"). Here only the detectors commonly used in neutron imaging systems are discussed.

The simplest form of detector consists of a scintillation material (Katigiri 2004) that is coupled to a device which will count the light flashes arising from the interaction of neutrons with the scintillator materials. For example, a scintillator such as glass or plastic blocks containing $^{10}_{4}B$ or $^{6}_{3}Li$ may be coupled to a photo multiplier system. Flashes from the interaction of the neutrons with the dopants are amplified by a photomultiplier tube or an avalanche photodiode and processed using conventional multichannel analyzer or pulse-height analysis systems. In this application, many scintillation detectors are mounted in an array, and the object is illuminated by a divergent beam. Or the image may be formed using a ZnS screen, which is viewed by a CCD TV camera. Imaging plates can be used, such as those used in X-ray radiography, and may find application in some instances. (Imaging plates were originally developed for X-ray radiography. They are large (A3 or A2 size) polymer sheets coated with a phosphor of Eu^{2+}-doped BaF_2. Interaction with particles causes electrons to be raised to metastable states in color centers in the phosphor. When illuminated by a laser the electrons return to their ground state with the release of photons, which are then detected with a scintillation detector.) Smaller devices utilize Complementary Metal Oxide Semiconductor (CMOS) detector arrays (Chap. 21, "New Solid State Detectors"; Chap. 16, "Semiconductor Counters").

Whatever method of detection is used, systems must have good spatial resolution, a short exposure time, a large active area onto which the image is projected, a large number of pixels per line scanned (which is related to the spatial resolution), a large and linear dynamic range, and a short readout time.

The type of experiment to be undertaken will determine which of the detector configurations to be used. *Table 1* lists these imaging systems and their configurations. Because of the rapid development of scintillation materials and counting systems, the figures in this table are to be taken as indicative rather than absolute.

Table 1

Characteristics of four common imaging systems

System	Scintillator CCD camera	Imaging plates	Amorphous silicon Flat panel	CMOS pixel detector
Spatial resolution (μm)	100–500	25–100	127–750	200
Exposure for a suitable image (s)	10	20	1–10	0.1–50
Detector area (typical) (cm × cm)	25 × 25	20 × 40	30 × 40	3.5 × 8
Pixels per line	1000	6000	1750	400
Dynamic range	10^5 (Linear)	10^5 (Linear)	10^3 (Nonlinear)	10^5 (Linear)
Readout time (s)	2–100	300	0.03–1	0.2

3 Neutron Interactions and Neutron Scattering Techniques

When a beam of neutrons of energy E and intensity I_0 passes through a material of thickness t it suffers a loss in intensity given by the Beer–Lambert law $(I/I_0 = \exp(-\mu_l t))$, where μ_l is the linear absorption coefficient. The cross section σ for an atom is related to μ_l by: $\mu_l = (1/V)\Sigma\sigma_i$, where the summation is carried out over all the nuclei in the volume V. As mentioned by Creagh (Chap. 25, "Technology for Border Security") most objects are not homogeneous in composition and structure. If the neutron beam traverses a number of materials with different thickness the composite linear absorption is given by $(\mu_l = \Sigma\mu_i t_i)$, where the sum is over all the different species of nuclei present.

The absorption coefficient comprises both elastic and inelastic processes. Neutrons have no charge. (Neutrons are baryons, and comprise one up quark and two down quarks. They have no charge, and a mass slightly larger than a proton. They do, however, have a magnetic moment (spin ½). As free entities they are unstable with a half life of ~15 min; $n^0 \rightarrow p^+ + e^- + \bar{\nu}_e$. Here $\bar{\nu}_e$ is an antineutrino. This is the simplest example of the phenomenon of β decay.) Interaction occurs directly inside the nucleus. In elastic collisions the neutron leaves the nucleus with almost no change in kinetic energy and the nucleus remains in the same configuration. There is a change in the momentum, however, and the observed absorption is due to scattering of the neutrons away from the direction of the incident beam.

Inelastic scattering arises from the disturbance of energy states within nuclei by the neutron. Many pathways exist for nuclear reactions involving neutrons and nuclei. These range from the creation of a stable nucleus with a higher atomic mass to excitation events in which the nucleus is perturbed into a higher excited state to which it relaxes emitting a γ-ray, to the creation of intermediate nuclei which transform to new nuclei by the emission of α or β particles, to the spallation of further neutrons, and for the actinide nuclei, in the production of fission.

The applications of interest in this chapter use either elastic scattering, or inelastic scattering in which prompt γ-rays are generated, or inelastic scattering in which intermediate nuclei with lifetimes greater than several days are produced.

3.1 Coherent Neutron Scattering

Elastic scattering and neutron scattering experiments are, broadly speaking, similar to those that use X-rays. In these experiments the neutron wavelength generally remains fixed. The parameter of interest is the total scattering angle (2θ). What is measured is the intensity of the scattered radiation at a scattering angle of θ. Much of what will be written about neutrons applies as well to X-rays. The difference between the two is that the coherent (or elastic) neutron length, which is related to the scattering cross section by the relation $\sigma = 4\pi bc^2$, varies significantly, and irregularly, between nuclei of different atomic mass, whereas the X-ray scattering cross section is a smooth analytic function of the atomic number (Creagh: Chap. 25, "Technology for Border Security"). As well, the scattering length can be negative for some nuclei. For example, deuterium and hydrogen have scattering lengths of 6.674 fm and −3.747 fm, respectively.

In what follows, several experimental techniques involving elastic scattering of neutrons will be described.

It is important to remember that, in all scattering processes, whether of neutrons or electromagnetic radiation, the dimension of the object relative to the wavelength of the radiation determines the nature of the scattering.

Note that it is necessary to determine the atomic composition of the specimen using X-ray Fluorescence Spectrometry (XRF) before undertaking any coherent neutron experiments because neutrons can activate some atoms and therefore cause them to become radioactive.

3.1.1 Small-Angle Neutron Scattering (SANS)

Small-angle scattering (SANS) is said to occur for small angles of scattering (defined by a momentum transfer, $q = (4\pi/\lambda)\sin\theta$, of less than about 1. The scale of the interaction is such that the object under investigation is very much larger than the wavelength of the neutron beam. The application of scattering theory leads to an equation for the scattering of an ensemble of identically shaped particles in a homogenous medium of the form

$$I(q) = 4\pi \int_0^\infty \frac{p(r)\sin(qr)}{qr}\,\mathrm{d}r,$$

where $p(r)$ is the pair-distance distribution function (PDDF), $p(r) = r^2\, V\gamma(r)$. Here V is the volume and $\gamma(r)$ is the correlation function (Porod 1951).

The "particles," which are referred to, are fluctuations in scattering density in the beam: the difference between the local scattering density and the average scattering density, $\Delta\rho$. The fluctuations may be, for example, colloidal particles in a liquid, or at the other extreme they may result from a distribution of voids in an alloy.

The total intensity in a SANS pattern is an invariant quantity. At the high-resolution region of the pattern, if the interface is smooth, the intensity distribution $I(q)$ is proportional to Sq^4, where S is the total surficial area.

The total scattering length is the intensity at $q = 0$, which is given by:

$$I(0) = 4\pi \int_0^\infty p(r)\,\mathrm{d}r.$$

The shape of the particles has effect, and scattering from them may be described by a scattering radius of gyration, which is entirely analogous to the mechanical radius of gyration. Tabulations of radii of gyration exist for an extremely wide variety of shapes (Glatter and May 2004).

Figure 2 shows the scattering from spherical, oblate, and prolate spheroidal particles. There is a marked difference in shape between these theoretical curves. Though instrumental broadening effects tend to smear the curves, it is usually possible to determine the shapes of monodisperse "particles." The problem becomes more difficult for poly-disperse systems in which particles of different sizes and shapes are involved.

The SANS technique is used extensively in the study of biological molecules, biological membranes, viruses, surfactants, mechanical defects in materials, colloids, alloy segregation, and proteins, and applications as well in cultural heritage science.

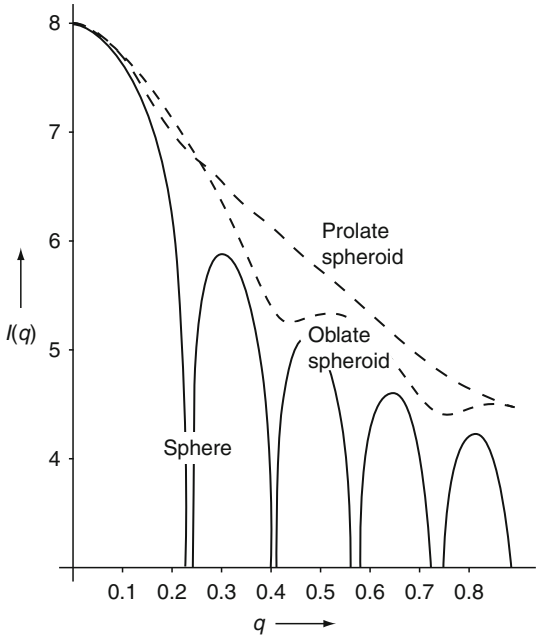


Fig. 2
The intensity of scattering form spherical, oblate spheroidal, and prolate spheroidal particles

3.1.2 Neutron Reflectometry and Grazing-Incidence Diffraction (GID)

When a neutron beam encounters an interface, like all waves, it is reflected and refracted according to Snell's law ($\cos \alpha_i / \cos \alpha_f = v_1/v_2 = n_2/n_1$) (Fermon et al. 2009). The extent to which it is reflected or refracted depends on the refractive index and the angle of incidence. For an interface between a material of mean nuclear potential, V_{av}, and air, the refractive index for coherent scattering is given by the relation: $n = \sqrt{1 - V_{av}/E} = 1 - \lambda^2 N b_c / 2\pi$. If absorption (nuclear reaction) processes and magnetic effects are involved, the formula for the refractive index must be modified (Sears 1989).

Figure 3 shows the configuration used for X-ray reflectometry experiments. For specular reflection the angle of incidence α_i is equal to the angle of reflection α_f. In reflectivity experiments α_i is usually fixed, and the angle α_f is varied. Surfaces are not necessarily homogeneous and may be ordered in some way. If the surfaces exhibit order, crystalline diffraction effects can occur. θ_B is the angle of incidence corresponding to a diffraction peak occurring at θ_{2B}. The two are related by Bragg's law ($2d_{hkl} \sin \theta = n\lambda$; where d_{hkl} is the interplanar spacing, θ is half the total angle of scattering, and λ is the wavelength) (Andersen et al. 2004). When diffraction occurs, grazing-incidence diffraction (GID) is said to occur.

Figure 3 shows only the processes occurring during reflection. Whether n is less than 1 (total external reflection) or greater than 1 (refraction), a wave field with a component normal to the surface exists, although for $n < 1$, the wave field penetrates only a few microns. If there

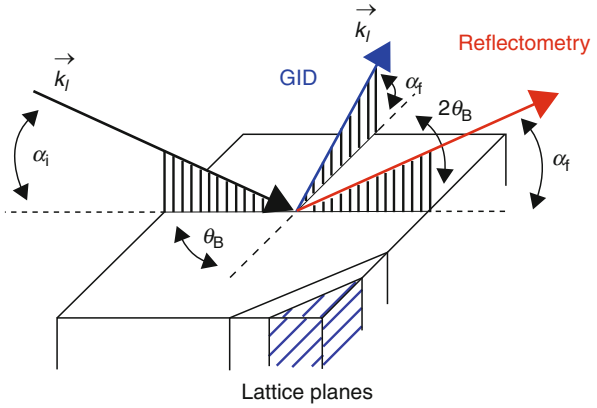


Fig. 3
The geometry of reflectivity and grazing-incidence diffraction experiments (GID)

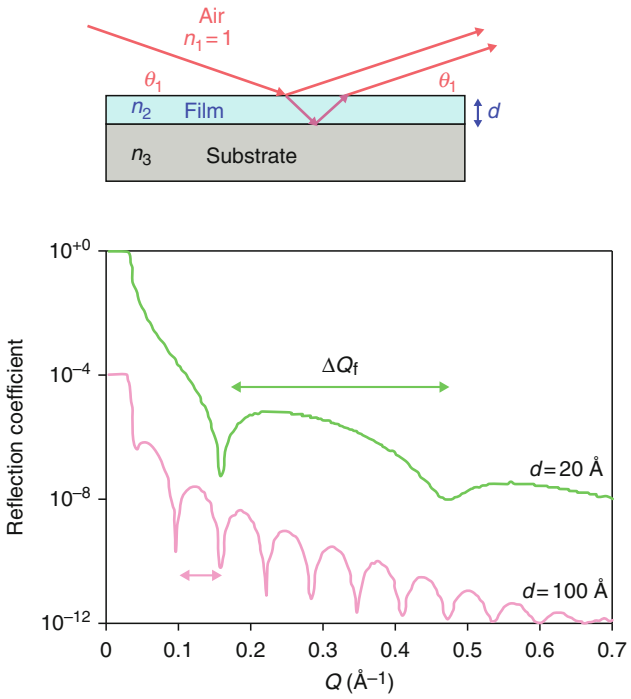


Fig. 4
Interference effects arising from the interaction of waves reflected by a buried interface

are other surfaces lying parallel to the surface, further processes of reflection and refraction occur, and constructive and destructive interference can occur between the reflected waves from the several interfaces. (These are referred to as Kiessig fringes. The spacing between fringes is: $d \sim 2\pi/\Delta Q_f$, where Q_f is the fringe spacing.) In *Fig. 4*, a schematic diagram of such an interface is shown, as well as calculated fringe patterns for two layer thicknesses.

Reflectometry can be used as a tool to study surface layers deposited or growing on substrates as well as layers lying beneath the surface. Roughness of the interface surfaces has the effect of blurring the specular reflection, as does instrumental resolution.

If regular structural order exists at the surface, grazing-incidence diffraction can occur. Usually specular reflection is measured using a single detector, setting α_i to a particular value and varying α_f in the plane of reflection. If, however, a two-dimensional detector array is used both the off-specular scattering due to surface roughness and any grazing-incidence diffraction peaks can be observed.

An example of GID is the study of the self-assembly of linear alkane-chain molecules on surfaces. These molecules may have many backbones comprising many CH_2 radicals, but the head and tail groups at the ends of the chain can be different. Museum conservators are interested in the suitability of these materials for use in the protection of metal surfaces. The long, aligned, molecular chains and the tight binding of the end groups to the surface effectively deny radicals that may cause corrosion access to the surface (Creagh et al. 1996). ❯ *Figure 5* shows a schematic diagram of such a self-assembly. The GID pattern shows strong hexagonal close packing of the alkane molecules (alkanes are saturated hydrocarbons, with a generic formula C_nH_{2n+2}. Many of the molecules of interest to conservators are salts of stearic acid $CH_3(CH_2)_{16}COOH$.).

Electrochemical cells have been devised to enable the study of the performance of self-organized systems exposed to different corroding environments (Price et al. 1996a,b).

3.1.3 Neutron Diffraction

The analysis of powder diffraction data from both X-ray and neutron experiments uses the Rietveld method (David and Jorgensen 1993). The specimens are assumed to be a mixture of crystalline phases, each phase contributing its own pattern to the overall pattern.

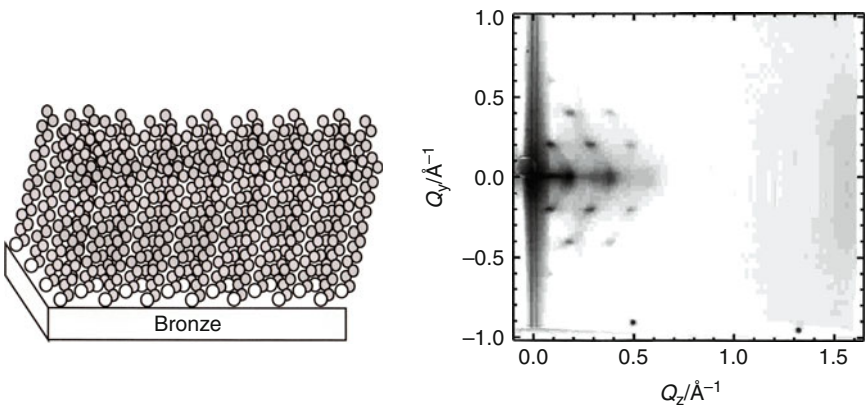


Fig. 5
***Left*: Pictorial representation of a self-assembled array of alkane molecules on a bronze substrate. *Right*: Grazing-incidence diffraction (GID) pattern of the manganese stearate molecules deposited on a silicon surface**

To determine the combination of phases present in the sample the Bragg equation is used to make a list of values of d_{hkl} corresponding to the observed peaks. Association of measured peak positions with calculated or observed positions of pure single-phase fingerprints can be made using database search–match routines. Once the phases are identified, the subsequent step is quantitative phase analysis, which assesses the amount of each phase in the sample material, either as volume or weight fraction assuming that: each phase exhibits a unique set of diffraction peaks and the intensities belonging to each phase fraction are proportional to the phase content in the mixture.

A full-pattern diffraction analysis can, in addition to the phase fraction determination, include the refinement of structure parameters of individual mineral phases, such as lattice parameters and/or atom positions in the unit cell.

The Rietveld method assumes that the structure of each component of a mixture is known, and its diffraction pattern can be modeled. It allows the refinement of phase-specific structure parameters along with experiment-specific profile parameters by fitting a calculated model pattern to the entire observed diffraction pattern using the least-squares algorithm which minimizes the quantity

$$D = \sum_i g_i \left(y_i^{\text{obs}} - y_i^{\text{calc}} \right)^2 .$$

The summation index i runs over all observed intensities y_i^{obs}. The weights g_i are taken from the counting statistics. y_i^{calc} are the calculated model intensities defined by instrumental and structural parameters, the latter including weight fractions in a multiphase refinement. By refinement of reflection profile parameters, crystallite size and micro-strain effects can be studied. The Rietveld routine calculates figures of merit (R-values) that indicate the quality of the fit of the entire model pattern to the entire observed diffraction pattern.

The weighted-profile R-value R_{wp} should converge to a minimum of < 5%:

$$R_{\text{wp}} = \left\{ \frac{\Sigma g_i \left(y_i^{\text{obs}} - y_i^{\text{calc}} \right)^2}{\Sigma g_i \left(y_i^{\text{obs}} \right)^2} \right\}^{1/2} .$$

Figure 6 shows a Rietveld fit to neutron data taken from the armor worn in the late nineteenth century by an Australian outlaw, Joe Byrnes. The lower curve shows the discrepancy between the experimental and the calculated values.

After fitting quantitative phase information can be obtained (a free downloadable Rietveld program that can be used with X-ray and neutron diffraction data is available from: http://www.ccp14.ac.uk/solution/gsas). As well, other parameters such as strain and structural disorder can be determined. These can be extracted from the data in the difference curve shown in *Fig. 6*. In this case the difference is due to small amounts of austenite and pearlite caused by reheating already forged steel at a low temperature (Reed-Hill and Abbaschian 1991).

The main advantages of quantitative phase analysis by the Rietveld method are as follows: no internal standard is required; crystal structure models are included explicitly; and structure parameters can be refined along with weight fractions of mineral phases; overlapping peaks and even peak clusters are handled without difficulty; preferred orientation of crystallites can be considered in the model.

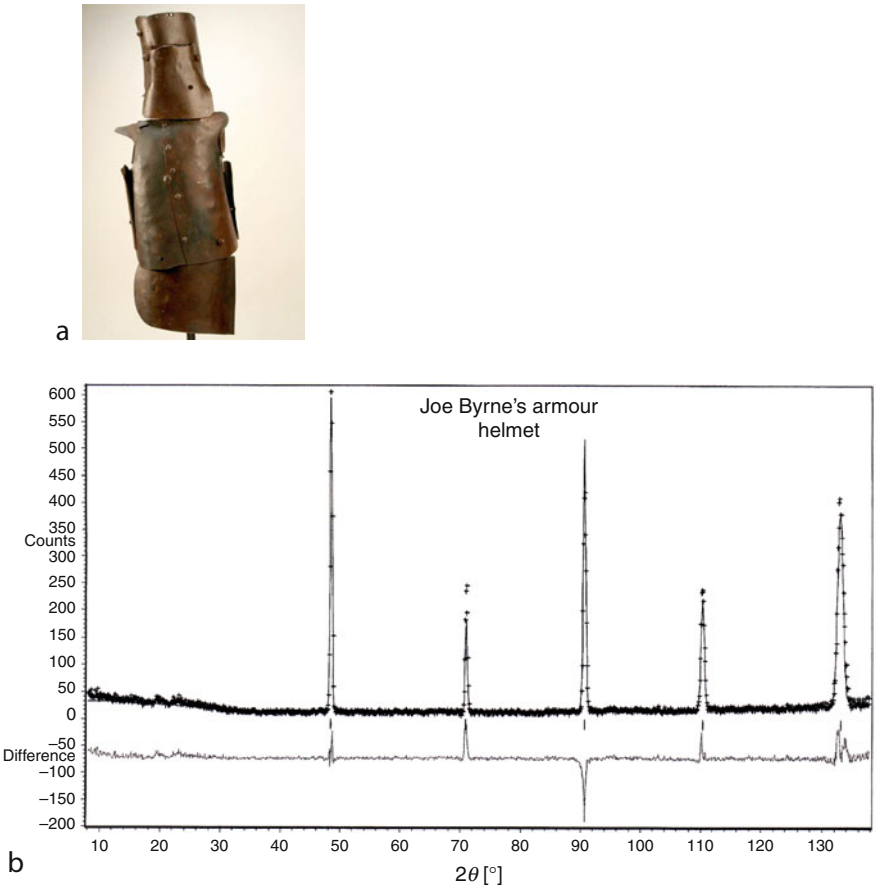


Fig. 6

(*Top* a) Joe Byrne's armor. (*Bottom* b) the upper curve shows the fitting of experimental and calculated results. The lower curve shows the difference between the two

3.2 Incoherent Scattering (Neutron Activation Analysis (NAA))

In NAA the specimen is irradiated by neutrons and artificial radioisotopes of the elements present are created. After irradiation decay occurs by the emission of particles, especially γ-rays. The radiations emitted are unique to the element from which they were emitted.

Modern NAA systems comprise either silicon or high-purity germanium solid state detectors, the pulses from which are processed and sent to a multichannel analyzer. For the NAA procedure to be successful the specimen or sample must be selected carefully. In many cases small objects can be irradiated and analyzed without the need of sampling. In conservation science, a small sample (~50 mg) is taken, usually by drilling in an inconspicuous place.

The sample is placed in a vial made of either high-purity polyethylene or glass and irradiated in a suitable reactor to a known thermal neutron flux ($\approx 10^{12}$ neutrons cm^{-2} s^{-1} with an average

energy of ~0.5 eV). Neutron capture (the efficiency of which varies from element to element) occurs. If the new nucleus is not stable it transmutes into another nucleus with the emission of particles (α- or β-) or γ-rays. Prompt and delayed γ-rays as well as α and β particles often accompany neutron capture. Of these only the delayed neutron activation is of interest here (DGNAA). The half-lives are dependent on the type of compound nucleus formed and can range from fractions of a second to several years.

Fast-neutron generators (14 MeV) are sometimes used for activation. Activation with fast neutrons is termed as Fast NAA (FNAA).

NAA can detect up to 74 elements (http://en.wikipedia.org/wiki/Neutron_Activation_Analysis) depending upon the experimental procedure used with minimum detection limits ranging from 0.1 to 1×10^6 ng g^{-1} depending on the elements under investigation.

NAA is used widely to study the composition of solids (in archaeology, botany, chemistry, conservation science, metallurgy, mineralogy), liquids (blood, fuels, waste materials, water), and slurries (sewerage, environmental water, paints). In the archaeological/conservation context, it has been used for studies of artifacts to determine trading patterns and ancient bones to determine population movements.

3.3 Total Neutron Scattering: Imaging

Imaging with neutrons, in its simplest sense, involves the passage of a divergent neutron beam through an object, and the detection and display of the absorption contrast on a screen. The experimental arrangement is similar to that used in medical radiography. The system resolution depends on the intensity of the beam and the pixel resolution of the detector.

3.3.1 Two-Dimensional Imaging

The simple system described above does not provide any possibility of discriminating between materials. Creagh (Chap. 25, "Technology for Border Security") described a system that uses a photon source (^{60}Co) to produce an image of the object as well as a sealed-tube fast-neutron generator (14 MeV) to produce the neutron image. In this system, the object is transported through fan beams of γ-rays and neutrons at a fixed rate, and two 2-D images are formed.

The ratio, *R*, of the X-ray and the neutron attenuations at a particular pixel is constant for a given material, and by comparing the measured value with a table of known values materials can be identified. The value of *R* can be read at any position, but to conform to the conventions used in X-ray baggage imaging, metals are color blue, inorganic materials red, and organic materials orange.

The resolution of the system is limited by the intensity of the neutron source and the size of the detectors. In *Fig. 7*, the *R*-value plot of a number of test objects is compared with a photograph of the test objects. Note the size of the polyethylene and steel step wedges used to test the penetration of the system.

For relatively thin samples and a relatively small degree of clutter, the system gives a good degree of materials discrimination. However, for thick samples, the ratio fails to be a good indicator of discrimination. This is apparent for the step wedges: each is a particular material, so each step in the step wedges should have the color characteristic of that material.



Fig. 7
***Top*: a rack of test objects and shapes and step wedges of polyethylene and steel. *Bottom*: *R*-values for these materials color-coded to the X-ray convention**

A recent modification to this system (using X-rays instead of γ-rays with upgraded software) provides much better materials discrimination.

3.3.2 Three-Dimensional Imaging: Computed Tomography

Many Computed Tomography configurations are possible. The position of the source and its collimating slits is fixed. Thereafter the configuration chosen is determined by the type of detector that is available. *Figure 8* shows schematically two such systems. The upper figure shows a system in which the neutron beam passes through the object onto a plate, which scintillates when struck by a neutron, and the image is reflected by a mirror onto a CCTV camera. (See *Table 1*.) This produces an image for one aspect of the object. Rotation of the object about an axis perpendicular to the beam by fixed amounts produces a series of images that can be reconstructed to form a 3-D image (a free textbook on reconstruction algorithms can be downloaded from: http://www.slaney.org/pct/pct-toc.html).

The middle image in *Fig. 8* is a more modern configuration using an array of neutron detectors. The spatial resolution is lower, but the speed at which measurements can be made is faster. In the bottom, a CT image of the head of a statue is shown, see F. Casali in Bradley and Creagh (2006).

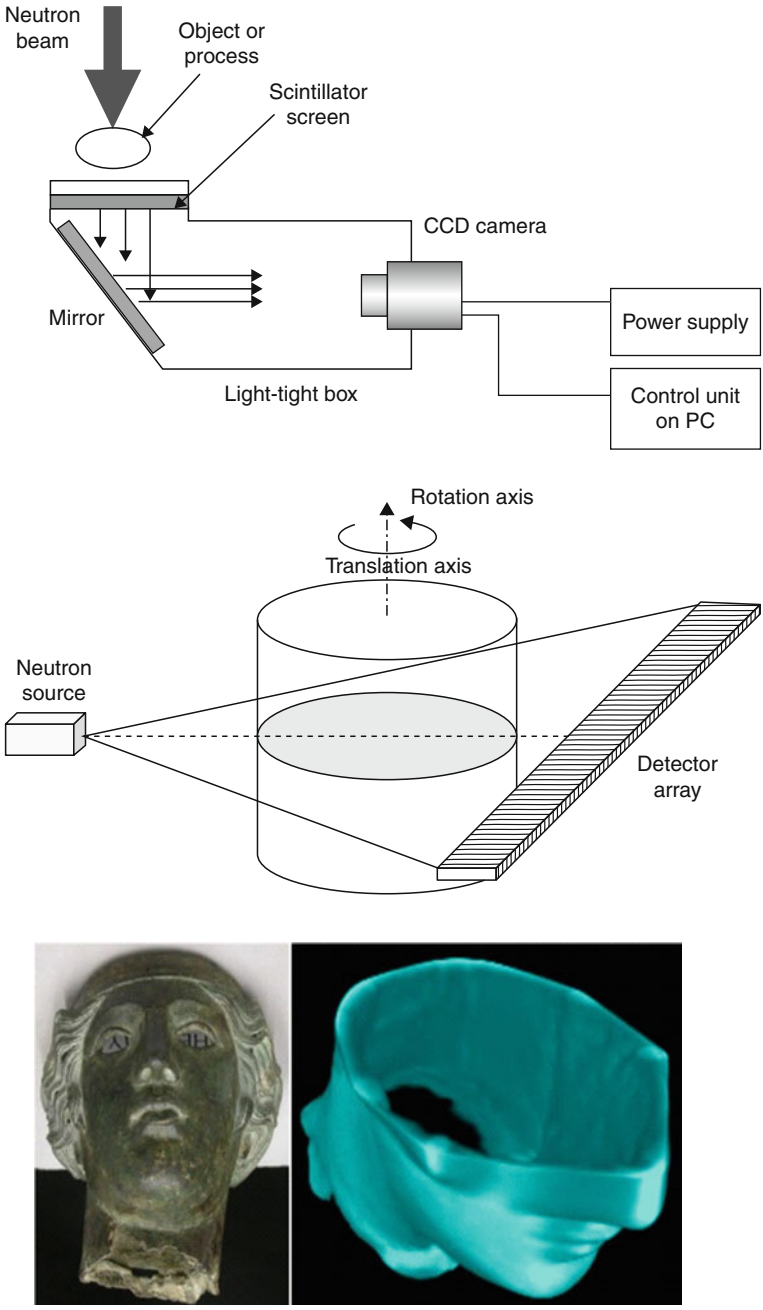


Fig. 8

***Top*: CT configuration using a scintillator/CCTV system. *Middle*: CT configuration using a linear array of neutron detectors. *Bottom*: CT image of the head of a statue**

4 Selected Applications

Neutron and X-ray techniques for the study of archaelological and cultural heritage materials are complementary: whatever X-rays can do in most cases neutrons can do as well. Each of the techniques has a research area in which it has an advantage, however. The use of neutron technology in cultural heritage science is not so developed as synchrotron radiation (SR) technologies. A rapid expansion of experiments using photons of energies from 0.01 eV (Infrared) to 200 keV ("hard" X-rays) has occurred at SR sources. A web-based archive of published papers has been created (E. Bertrand, D. Vantelon, E. Pantos: http://srs.dl.ac.uk/arch/publications.html). The number of published papers has increased from 8 in 1998 to more than 50 in 2010. Many neutron sources are now involved in conservation science. The neutron spallation source at the Rutherford–Appleton Laboratory (ISIS) has an active program in cultural heritage science (W. Kockelmann: http://srs.dl.ac.uk/arch/publications.html).

Like all major facilities many types of experiments are served by one spallation source. *Figure 9* shows the plan of the experimental hall at the spallation source at ISIS. At the bottom left the beamline carrying the protons from the accelerator to the spallation source is shown.

Around this are clustered groups of experimental stations. Stations relevant to potential users of these facilities for cultural heritage studies are the group associated with nuclear scattering by neutrons (SANS, Reflectometry, and GID) (surf, crisp, loq) and the group associated with neutron diffraction (gem, Engin-X, Hrpd, Nimrod, Osiris, Rotax).

Some examples will be given of the applications of these techniques. At the Saclay reactor *SANS* has been used to study particle and pore size in building stone taken from various archaeological sites in NE Sicily (Giodarno et al. 2007). Samples of quartzite, sandstone, calcarenite, and marble were shown to exhibit particle- and pore-size dimensions that are inversely proportional to the porosimetric values of the stones.

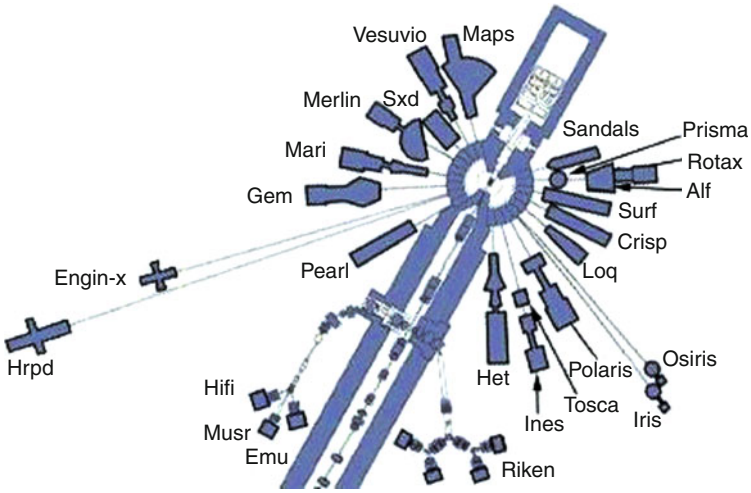


Fig. 9
Plan view of the experimental hall at the neutron spallation source, ISIS

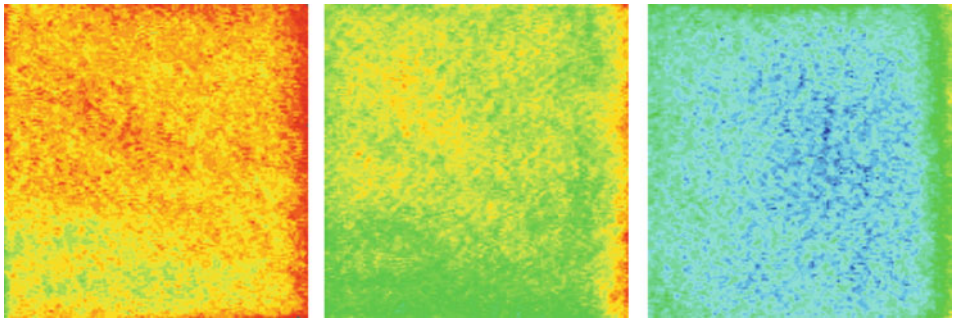


Fig. 10
Different stages in the uptake of water into a sandstone block. *Left*: dry. *Center*: exposure of 1 hour. *Right*: saturation

Some of the recent research using *neutron diffraction* has been undertaken on Gem, which has a multielement (7270) detector system and a range of scattering angles from 1 to 169°.

Any crystalline material (metal, pigments, rock, ceramics) can be analyzed by neutron diffraction. Large artifacts can be analyzed in air; without sampling or sample preparation being necessary. Analysis under vacuum or other environments (e.g., high or low temperature) is also possible. Measurements are made in live time, which can be useful for testing conservation materials and treatments.

Multiple data points can be collected across an object. A typical measurement time per point is around 30 min. Rietveld analysis is used to determine what metal, mineral, and inter-metallic compounds are present, and how much of each phase. In metals, grain orientation can be seen, indicating manufacturing techniques like casting and hammering. It is important to note that the neutron diffraction technique does not give data on atomic composition: this should be determined before the experiment begins using X-ray fluorescence spectrometry (XRF).

Reports of recent advances in the study of cultural heritage materials using neutrons can be found from the ISIS web site, for details see Kockelmann et al. (2003). This describes studies of sixteenth century silver/copper coins and an imperfectly repaired Greek bronze helmet.

Neutron imaging has been used to study the porosity of building materials. This may be applied to studies of rising damp in buildings of cultural heritage significance. *Figure 10* illustrates the change in the image (in this case one of a family of CT images) as a sandstone block is exposed to water. Analysis of the data reveals the uptake of water into pores within the sandstone (data from Frikkie de Beer, SAFARI research reactor, RSA).

Neutron Activation Analysis is widely used in archaeology, biology, botany, conservation science, geology, and many other research fields for the determination of atomic concentrations at the major, minor, and trace levels of concentration. It is a mature technique, and this is reflected in the abundance of papers in the literature. It has been used as a research tool by archaeologists for more than 50 years. A recent review of its use in archaelology had been published (Ashworth and Abeles 1957). An example of how NAA is applied in archaeology is the study of iron artifacts to establish the development of trade routes and to discriminate between different French iron-making regions (Desaulty et al. 2008).

5 Conclusions

Neutron and X-ray sources can be used in a complementary fashion for a variety of studies in conservation science. The underlying theory and the configuration of experiments are almost identical for small-angle scattering, reflectometry, and grazing-incidence diffraction, diffraction, and imaging. Differences exist due to the nature of the particle–atom interaction, however. And there is no X-ray equivalent of nuclear activation analysis.

Acknowledgments

The author is grateful to a number of colleagues from international research facilities for discussions on cultural heritage science over the past decade. Synchrotron radiation scientists who, over a period of time, have been very helpful include: Drs Loic Bertrand and Paul Dumas (Soleil); Eric Dooryhee (ESRF); and Manolis Pantos (Diamond). Scientists with neutron experience include: Drs Winfreid Kockelmann (ISIS); Frikkie de Beer (SAFARI-1); and Michael James and Stephen Holt (OPAL). The author acknowledges the active assistance of Australian national collecting institutions (Australian War Memorial, National Archives of Australia, National Film and Sound Archive, National Gallery of Australia, National Museum of Australia), and in particular David Hallam (NMA) and David Thurrowgood (National Gallery of Victoria).

References

Andersen IS, Brown PJ, Carpenter JM, Lander G, Pynn R, Rowe JM, Scharff O, Sears V, Willis BTM (2004) Neutron techniques. In: Wilson AJC, Prince E (eds) International tables for crystallography, vol C, 3rd edn. Kluwer, Dordrecht, Section 4.4, pp 452–467

Ashworth MJ, Abeles TP (1957) Neutron activation analysis and archaeology. Nature 210: 9–11

Bradley DA, Creagh DC (eds) (2006) Physical techniques in the study of art, archaeology and cultural heritage, vol 1. Elsevier BV, Amsterdam

Creagh DC (2007) Synchrotron radiation and its use in art, archaeometry, and cultural heritage studies. In: Creagh DC, Bradley DA (eds) Radiation in art and archaeometry. Elsevier Science BV, Amsterdam, pp 1–97

Creagh DC, Bradley DA (eds) (2000) Radiation in art and archaeometry. Elsevier Science BV, Amsterdam

Creagh DC, Bradley DA (eds) (2007) Physical techniques in the study of art, archaeology and cultural heritage, vol 2. Elsevier Science BV, Amsterdam

Creagh DC, Otieno-AlegoV, O'Neill PM (1996) X-ray reflectivity studies and grazing incidence X-ray diffraction studies of the adhesion of protective wax coatings on metal surfaces. In: JW Boldeman (ed), Report Australian Synchrotron Research Program, Ansto, Sydney

David WIF, Jorgensen JD (1993) Rietveld refinement with time-of-flight powder diffraction data from pulsed neutron sources. In: Young RA (ed) The Rietveld Method, Oxford University Press. pp 197–227

Desaulty AM, Mariet C, Dillman P, Joron JL, Fluzin P (2008) A provenance study of iron archaeological artefacts by ICPKS multi-element analysis. Spectrochim Acta part B At Spectrosc 63(11):1253–1262

Eberhardt JE, Rainey S, Stevens RJ, Sowerby DD, Tickner (2005) Fast neutron radiography scanner for the detection of contraband in air cargo. Appl Rad Isotopes 63:179–188

Fermon C, Ott F, Menelle A (2009) Neutron reflectometry. In: Dalaint J, Gibaud A (eds) X-ray and neutron reflectometry: principles and applications. Springer, pp 185–206

Giodarno R, Teixiera J, Tiscari M, Wanderlingh U (2007) Porosimetric and particle measurements by small angle neutron scattering. Eur J Miner 19:223–228

Glatter O, May R (2004) Small angle techniques. In: Wilson AJC, Prince E (eds) International tables

for crystallography, vol C, 3rd edn. Kluwer, Dordrecht, Section 2.6, pp 89–112

Katigiri M (2004) Scintillation materials for neutron imaging detectors. Nucl Instrum Methods A526:274–279

Kockelmann W, Pantos E, Kirfel A (2000) Neutron and synchrotron radiation studies of archaeological objects. In: Creagh DC, Bradley DA (eds) Radiation in Art and Archaeometry, Elsevier Science BV (2000) pp 347–378

Kockelmann W, Kirfel A, Linke R, Schreiner M, Traum R, Pantos E, Garner R, Prag, AJNW (2003) Genuine or fake? Neutron diffraction for non-destructive testing of museum objects. ISIS 2003 Annual Report, RAL-TR-2003-050, Science Highlights

Lewis EE (2008) Fundamentals of reactor physics. Academic, London

Porod G (1951) Die Rontgenklienwinklstreung von dichtgepackten kolloiden Systemen I. Kolloid Z 124:83–114

Price CL, Hallam DL, Ashton JA, Heath GA and Creagh DC (1996a) An electrochemical study of waxes for bronze sculpture. In "Metals 95". James and James Scientific Publishers, London, pp 233–241

Price CL, Hallam DL, Ashton JA, Heath GA and Creagh DC (1996b) Redefining the electrochemistry of corrosion processes. In "Metals95". James and James Scientific Publishers, London, pp 220–224

Reed-Hill R, Abbaschian S (1991) Physical metallurgy principles, 3rd edn. PWS-Kent, Boston

Sears VE (1989) Neutron optics. Oxford University Press, Oxford

34 Radiation Detectors and Art

Andrea Denker
Ion Beam Laboratory ISL, Helmholtz-Zentrum Berlin, Berlin, Germany

1 Introduction and Motivation 834

2 Proton Induced X-Ray Emission: PIXE 835
2.1 Basic Principles 835
2.2 High-Energy PIXE 840
2.3 PIGE 841
2.4 Application to Art Objects 842

3 Experimental Setup for Art Objects 843

4 Examples 844
4.1 Paintings 844
4.1.1 Flemish Painting 845
4.1.2 Modigliani Portrait 846
4.2 Metals 847
4.2.1 Silver Coins: Wiener Pfennig 848
4.2.2 Gold Scarab 850

5 Conclusions 852

6 Cross-References 852

References 852

C. Grupen, I. Buvat (eds.), *Handbook of Particle Detection and Imaging*,
DOI 10.1007/978-3-642-13271-1_34, © Springer-Verlag Berlin Heidelberg 2012

Abstract: The use of radiation detectors in the analysis of art objects represents a very special application in a true interdisciplinary field. Radiation detectors employed in this field detect, e.g., x-rays, γ-rays, β particles, and protons. Analyzed materials range from stones, metals, over porcelain to paintings. The available nondestructive and noninvasive analytical methods cover a broad range of techniques. Hence, for the sake of brevity, this chapter will concentrate on few techniques: Proton Induced X-ray Emission (PIXE) and Proton Induced γ-ray Emission (PIGE).

The basics of these techniques will be described together with tables and references with relevant information for this field. On selected examples, the potentials and the pitfalls of the applied methods will be described.

1 Introduction and Motivation

Worldwide, there is an increasing interest in matters related to our cultural heritage: among others, these issues comprise dating, provenance, manufacturing, origin, and conservation of an artifact. Organizations like UNESCO or the European Union are fostering networks of natural scientists and scholars for a better understanding of our cultural heritage.

Besides stylistic and historical considerations, combined with comprehensive studies of historical sources, analytical methods and techniques are essential tools as they provide the means to understand better the objects under investigation. These methods have their origin in the cutting edge of up-to-date science and comprise analytical techniques that have been developed for modern physics, chemistry, and biology. However, there is one essential difference between the analysis of ancient and modern materials: an art object or ancient artifact is unique; hence, in most cases sampling is prohibited. In the ideal case the analysis is nondestructive and noninvasive. Even in the case where sampling is allowed, nondestructive testing offers the possibility of obtaining more information on one specific sample as complementary techniques may be applied.

In 1888, Friedrich Rathgen was appointed head of the *Chemisches Labor der Königlichen Museen zu Berlin* (Chemical Laboratory of the Royal Museums in Berlin), the world's first natural science laboratory within a museum. Since then, the field has evolved rapidly: Besides all kinds of chemical analyses, there is today a vast abundance of many different noninvasive as well as nondestructive analytical techniques applied to art objects (Janssens and Van Grieken 2005; Bradley and Creagh 2006/2007), even when considering only those involving radiation detectors. Art objects are x-rayed or CT scanned (Lang 2005), or analyzed with neutron tomography (Kardjilov et al. 2007), only to name a few.

Several nondestructive and noninvasive methods rely on the emission of characteristic x-rays. They are characteristic for the emitting element, thus, allowing the identification of this element, and their intensity provides information about their concentration in the sample. To create excited atoms, which hereupon emit these x-rays, one can use either x-rays (X-Ray Fluorescence – XRF (Beckhoff et al. 2006), electrons (electron microprobe (Goldstein et al. 1992)), or ions. For the excitation with ions, commonly hydrogen ions are used. Therefore, this technique is called Proton Induced X-ray Emission (PIXE).

2 Proton Induced X-Ray Emission: PIXE

2.1 Basic Principles

When material is irradiated with ions having an energy of some MeV (Mega electron Volt), these can interact either with the atomic nuclei or with the electrons of the sample. The ions collide with the inner-shell electrons and eject these from the inner shell. This effect was observed for the first time by Chadwick (1912). The vacancies are filled with electrons from outer shells. The energy difference is released either by the emission of x-rays or the energy is transferred to another electron that leaves the atom as the so-called Auger electron. The probability that an x-ray is emitted instead of an Auger electron increases with the atomic number Z and is described by the fluorescence yield ω (see *Fig. 1*). It can be calculated by the semi-empirical formula of Bambynek et al. (1972):

$$\left(\frac{\omega}{1-\omega}\right)^{1/4} = \sum_{i=0}^{3} B_i Z^i. \tag{1}$$

The coefficients B_i are listed in *Table 1* (Johansson and Campbell 1988a).

The arrangement of the shells is different for each element; hence, the energy of the emitted x-rays is characteristic for the element in question. Moseley's law (Moseley et al. 1913) describes empirically the correlation between the frequency ν of the Kα line and Z:

$$\sqrt{\nu_{K\alpha}} = \kappa\,(Z-1) \tag{2}$$

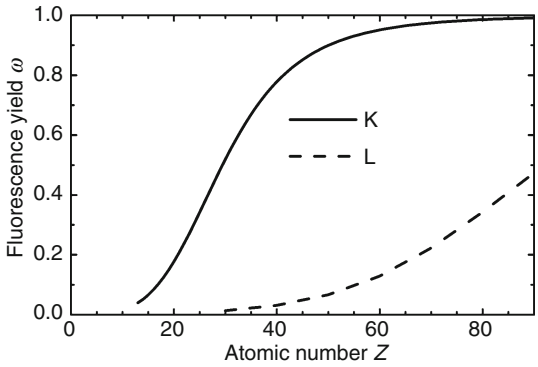


Fig. 1

Fluorescence yield as a function of the atomic number *Z*. The emission of a characteristic x-ray is the most probable process for heavy elements

Table 1

Coefficients for the calculation of the K- and L-shell fluorescence yields (Bambynek et al. 1972)

	K	L
B_0	3.7×10^{-2}	0.17765
B_1	3.112×10^{-2}	2.98937×10^{-3}
B_2	5.44×10^{-5}	8.91297×10^{-5}
B_3	-1.25×10^{-6}	-2.67184×10^{-7}

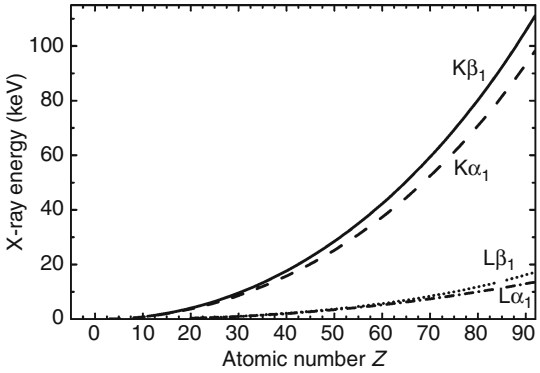


Fig. 2
Energy of the characteristic x-rays as a function of the atomic number

with $\kappa^2 = 2.48 \times 10^{15}$ Hz. The determination of the x-ray energy allows the identification of the emitting element. *Figure 2* shows the dependence of the x-ray energy as a function of Z (Bearden 1967).

The electrons around an atom are arranged in the main shells K, L, M . . . corresponding to the main quantum numbers $n = 1, 2, 3, \ldots$. The orbital angular momentum l and the spin s of the electrons interact, leading to the fine structure of the electron shells (see *Fig. 3*). The total angular momentum j is given by:

$$\vec{j} = \vec{l} + \vec{s}. \tag{3}$$

Electrons filling a vacancy in the K or L shell from outer shells have to fulfill the rules $\Delta l = \pm 1$ and $\Delta j = 0, \pm 1$. The Kα lines are the transitions from L to K, Kβ are the transitions from M to K, and so forth. For the K lines, the relative intensity of the various lines is quite well known (Scofield 1974; Perujo et al. 1987). For the L lines, the situation is more complicated due to the Coster–Kronig effect (Coster and Kronig 1935): a vacancy in the L shell may be converted via a non-radiation transfer to a higher shell before the emission of an x-ray or Auger electron.

The probability for the production of a vacancy in an inner shell is given by the cross section. The highest probability is achieved, when the velocity of the proton matches the velocity of the electron in that shell. Due to the different binding energies of the electrons in the different shells of atoms with different Z these velocities vary. Hence, the projectile energy with maximum production probability is a function of Z (see *Fig. 4*).

Cross sections were first calculated theoretically with the Plane Wave Born Approximation (PWBA), where perturbation theory is applied: In the start system, the projectile is described as a plane wave and a bound electron is included; the final state is a plane wave projectile and an electron in the continuum. This theory has been further developed by taking into account the energy loss (E), the deviation and the slowing down of the projectile in the Coulomb field of the atom (C), the perturbation of the stationary status of the atom (PSS), and relativistic effects (R). This is now known as ECPSSR theory (Brandt and Lapicki 1981). In addition to the theoretical calculations, a vast amount of experimental data is available. Paul and Sacher analyzed the data

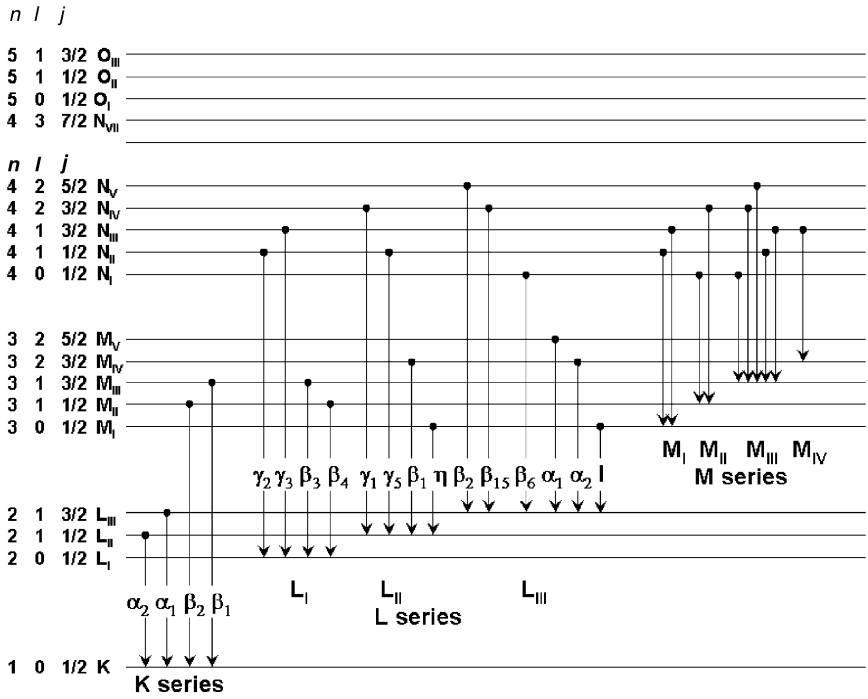


Fig. 3

Schematic diagram of the excited atomic states and the possible transitions leading to the emission of characteristic x-rays. Main quantum number *n*, angular momentum *l*, and total angular momentum *j* are indicated

for the K shell (Paul and Sacher 1989), resulting in so-called reference values. At 3 MeV proton energy, experimental values and theoretical values of the cross sections for K x-rays agree quite well (*Fig. 5*).

As the cross sections drop with increasing Z for a given proton energy, for proton energies up to 4 MeV usually the L lines of elements heavier than Sn are used for quantitative analysis. For the three L subshells, a compilation of experimental cross-section measurements, using the data of Krause (1979) for the L-subshell fluorescence as well as the Coster–Kronig probabilities is provided by Orlic et al. (1994). A recent survey of the Lα x-ray production cross sections from a number of empirical formulae published in the last two decades is reviewed by Lapicki (2009). Although the precision for the L shells is not as high as for the K shells, these data can be considered as sufficient for most analytical purposes.

The protons lose energy as they interact with the electrons and atomic nuclei. The x-rays are produced along the total flight path of the projectile ion through the sample. Therefore, the cross section for the production rate of the x-rays varies over the depth. This effect can be disregarded for very thin samples; however, for thick samples the energy loss of the ion has to be taken into account. Data for the energy loss can be found in the book of Ziegler et al. (1985). *Table 2* gives the ranges of protons in air, water, glass, and brass calculated by SRIM (Stopping and Ranges of Ions in Matter, www.srim.org).

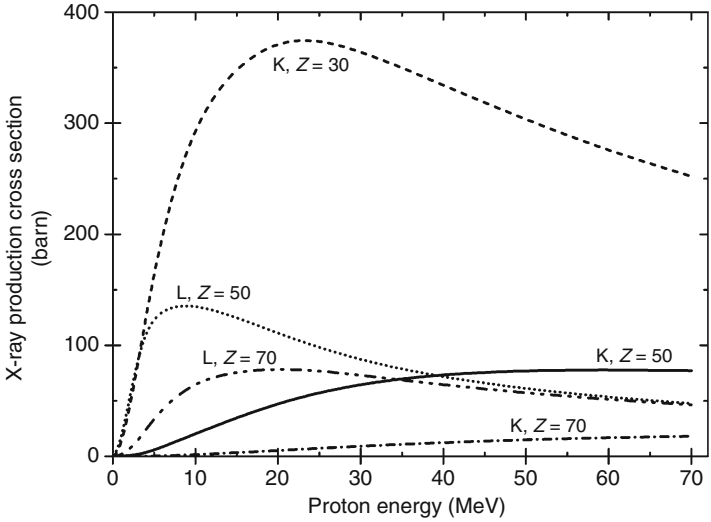


Fig. 4
Total x-ray production cross sections as a function of proton energy for various *Z*

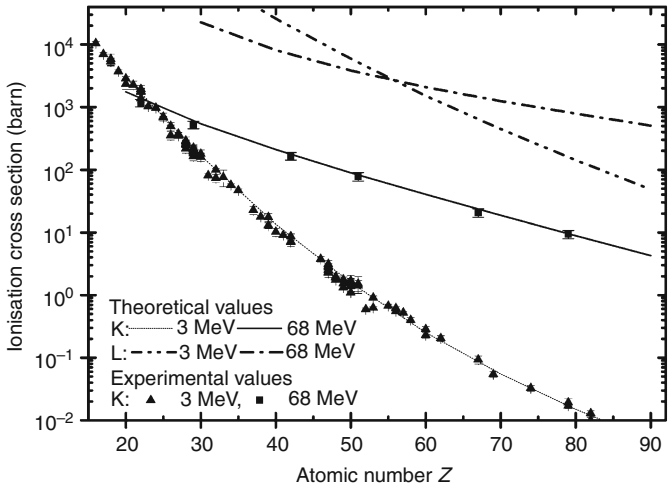


Fig. 5
Total ionization cross sections for 3 and 68 MeV protons as a function of *Z*. Experimental values are from Paul and Sacher (1989) for 3 MeV protons, and from Denker et al. (2005) for 68 MeV protons. At a proton energy of 68 MeV, the cross section for the K lines is several orders of magnitude larger than for 3 MeV protons for heavy elements

In addition, for the detection of the x-rays, the attenuation of x-rays in matter has to be considered. For very thin samples, the attenuation is negligible. In thick samples, the attenuation in the sample will determine the detection limits and the possible analytical depth. The intensity I from the original x-ray intensity I_0 after the transition of a material with a thickness

Table 2

Ranges of protons in different materials, calculated with SRIM2008

	1 MeV	2 MeV	3 MeV	4 MeV	68 MeV
Air	23 mm	72 mm	140 mm	230 mm	35 m
H_2O	26 μm	77 μm	149 μm	240 μm	38 mm
Glass (Pyrex)	15 μm	45 μm	86 μm	140 μm	21 mm
Brass (62% Cu, 35% Zn, 3% Pb)	7 μm	20 μm	37 μm	59 μm	7 mm

Table 3

Thickness of glass, silver, and gold that will attenuate x-rays of Cu and Pb by 90% (i.e., leaving 10% of the original intensity)

X-ray	Glass (Pyrex)	Silver	Gold
Si Kα (1.74 keV)	13.2 μm	1.08 μm	0.8 μm
Cu Kα (8.04 keV)	0.3 mm	0.01 mm	0.006 mm
Pb Lα (10.5 keV)	0.7 mm	0.02 mm	0.01 mm
Pb Kα (74.3 keV)	52 mm	0.63 mm	0.4 mm

d is given by

$$I = I_0 \exp(-\mu d), \tag{4}$$

where μ is the attenuation coefficient (Hubbell and Seltzer 1995). *Table 3* gives the thicknesses of different materials that reduce the intensity of x-rays to one tenth of the original intensity.

The number of atoms emitting characteristic x-rays in a sample is given by:

$$N_t = \frac{Y(Z)}{N_p \omega_Z b_Z \varepsilon_{abs} \int_0^{x_{max}} \sigma(x) \exp\left(\frac{-\mu x}{\sin\theta}\right) dx} \tag{5}$$

with

x: depth in the sample
N_t: number of atoms in the sample
σ: x-ray production cross section, which is a function of the depth as the protons lose energy along their flight path
θ: angle between the sample normal and the detector
$Y(Z)$: x-ray yield (in counts) from peak area of the correspondent line
N_p: number of protons
ω_Z: fluorescence yield
b_Z: fraction of x-rays in that line
ε_{abs}: absolute detector efficiency
μ: absorption of x-rays between target material and detector crystal

Besides the characteristic x-rays, there is also a continuous contribution to the spectrum due to various backgrounds:

- Atomic Bremsstrahlung (AB), created by the deceleration of bound target electrons in the Coulomb field of the projectile.

– Secondary Electron Bremsstrahlung (SEB), due to the deceleration of electrons emitted from atoms in the ionization processes. The maximum energy E_{max} that can be transferred to an electron in a central collision is given by ❯ Eq. 6, where m_e and M_P are the masses of electron and projectile and E_P is the projectile energy:

$$E_{max} = \frac{4m_e}{M_P} E_P. \tag{6}$$

If the collision is not head-on, a smaller energy than E_{max} will be transferred by Coulomb interaction. In the spectrum, we observe an intense contribution below E_{max}, decreasing rapidly with increasing energy.

– Quasi Free Electron Bremsstrahlung (QFEB) from the quasi-free electrons in the solid, and the maximum energy transfer to those electrons is increasing the low-energy part of the background,

$$T_r = \frac{m_e}{M_P} E_P. \tag{7}$$

– Compton background arises from the Compton scattering of γ-rays created in nuclear reactions. Comparing this background for protons and heavy ions at the same velocity, the Compton background is much larger for heavy ions.

Both AB and SEB have an anisotropic behavior and are maximal at 90° with respect to the incoming particle beam. Therefore, for most PIXE setups a backward geometry of the detector is preferred.

It is not possible to give a priori detection limits for a certain element in a particular sample: It depends on the sample composition, the thickness of the sample, and, therefore, the contribution of the background to the spectrum.

To solve ❯ Eq. 5, various computer codes are available, e.g., Geopixe (Ryan et al. 1990), Gupix (Campbell et al. 2005), Pixan (Clayton et al. 1986), Pixeklm (Szabo and Borbely-Kiss 1993), Sapix (Sera and Futatsugawa 1996), and Winaxil (Vekemans et al. 1994). The achievable precision for quantitative analysis is better than 5%, even for trace elements, if the sample is homogenous and has a flat, smooth surface.

2.2 High-Energy PIXE

Two factors define the achievable analytical depth: The range of the protons in the investigated material and the absorption of the x-rays.

The range of protons with a given energy depends on the composition and density of the material. The possible analytical depth is smaller than the proton range due to the energy loss of the protons in the sample. As mentioned above, the cross section reaches a maximum if the velocity of the protons matches the velocity of the electrons. Below that energy, the cross section for a given Z decreases with decreasing energy. At a proton energy of 3 MeV, the velocity of the protons matches the electron velocity in the K shell of fluorine. For higher Z, the cross section drops remarkably. Therefore, for $Z > 50$, commonly the L lines are used for analysis. For higher energies, the maximum of the cross section increases with higher Z. Hence, heavy elements may be detected via the higher-energetic K lines simultaneously with the L lines.

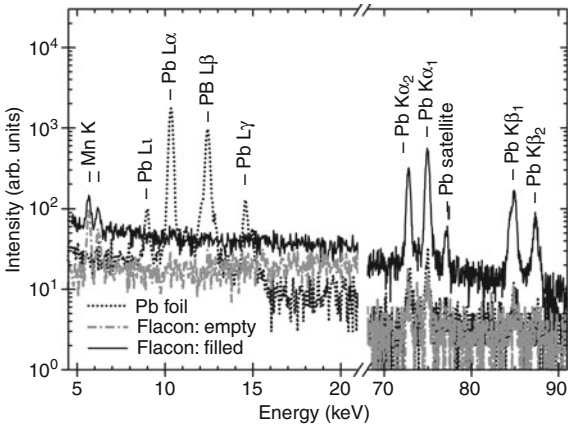


Fig. 6

High-energy PIXE spectra of an Egyptian flacon, measured on the empty part (*gray line*) and on the filled part (*black line*). On the filled part, the K lines of Pb are clearly visible, while there are no Pb L lines. Hence, the fill consists of a Pb-containing compound. For comparison, the spectrum of a thin Pb foil is shown (*dotted line*)

The other parameter limiting the analytical depth is the absorption of the x-rays in the above lying material. For proton energies above 50 MeV, the achievable analytical depth depends solely on the absorption. For low-Z elements, e.g., Na to Zn, only the K lines are visible in the spectrum. For heavier elements, the energy of the L lines is large enough to be detected. As the energy of the K lines is much higher compared to the L lines, they will suffer less absorption. An example is given in *Fig. 6*: The *Staatliche Museum Ägyptischer Kunst München* owns an Egyptian glass flacon, partly filled with a dark powder. The glass itself is brownish; hence, it is not possible to guess the real color of the powder. As the glass still has its original seal, the analytical task was to determine the elements of the powder through the glass, without the opening of the flacon. High-energy protons pass easily through a layer of 1 mm glass (*Table 2*), and the attenuation of the x-rays of heavy elements is small (*Table 3*). The flacon was measured on the empty part and on the filled part. The difference in the spectra comes from the powder. The Pb K lines have a high intensity, while the Pb L lines are completely absorbed by the glass. Thus, the powder contains a Pb compound.

The appearance of both K and L lines in the spectrum can be exploited for further information about the sample: Quantitative analysis can be done using either the K or the L lines. If the resulting concentrations are different, this indicates that the sample is not homogenous in depth.

However, the background in high-energy PIXE spectra is larger; especially, the contribution of the Compton background is much higher, as the probability for nuclear reactions is higher.

2.3 PIGE

At proton energies above some MeV, nuclear reactions may occur and prompt γ-rays can be produced. These γ-rays may be used to identify elements otherwise not detectable with the PIXE

Table 4

Nuclear reactions used for proton induced γ-ray emission and the energies of the measured γ-ray

Reaction	E_γ (keV)
$^{19}F(p, \alpha\gamma)^{16}O$	6,129
$^{23}Na(p, p\gamma)^{23}Na$	440
$^{23}Na(p, \alpha\gamma)^{20}Ne$	1,634
$^{28}Si(p, p\gamma)^{28}Si$	1,779

technique, especially light elements (*Table 3*). This technique is called Proton Induced γ-ray Emission – PIGE. *Table 4* lists possible reactions often used for the analysis of art objects. For instance, the reaction $^{19}F(p, \alpha\gamma)^{16}O$ has been used to determine the fluorine diffusion into archaeological teeth (Gaschen et al. 2005).

The advantage of PIGE is that the measurements can be performed simultaneously to the PIXE measurements by adding a detector for the more energetic γ-rays (see Chap. 17, "Gamma-Ray Detectors").

Hence, elements not detectable when the object remains in ambient atmosphere air due to the strong absorption of their low-energy x-rays in the air between sample and detector, e.g., Na or F, can be measured together with the x-rays of heavier elements. In addition, for some elements x-ray lines as well as γ-rays can be measured simultaneously. This is very useful, e.g., when measuring ancient glass: Si x-rays will provide information about the sample surface, whereas the γ-rays provide information from larger depths (Mäder and Neelmeijer 2004). As the glass surface is often altered due to aging processes, the two different information depths permit – in the same run – the determination of Si close to the surface and in the bulk. Thus, this technique complements the PIXE measurements.

2.4 Application to Art Objects

The first application of the PIXE technique was performed by Johansson et al. (1970). They irradiated samples consisting of carbon foils with subnanogram depositions of titanium and copper with proton beams of 2–3 MeV energy from a Van-de-Graaff accelerator. Since that time, PIXE has evolved tremendously, also thanks to the rapid development of suitable solid state x-ray detectors. Industry is offering today small accelerators with complete experimental setups, thus allowing the application of PIXE and PIGE in various fields, like archaeometry, biology, geosciences, medicine, and environmental studies (see Chap. 26, "Accelerator Mass Spectrometry and its Applications in Archaeology, Geology and Environmental Research"). Under favorable conditions, detection limits below 1 ppm can be achieved. The interested reader will find detailed descriptions of PIXE in the books of Campbell and Tesmer (Johansson and Campbell 1988b; Johansson et al. 1995; Tesmer and Nastasi 1995).

The possibility to extract the proton beam from the vacuum of the beam-line into ambient atmosphere has lead many laboratories worldwide into the analysis of art objects [e.g., Mandó 1995; Respaldiza et al. 1997; Demortier et al. 2000; Zucchiatti et al. 2002; Calligaro et al. 2002; Boutaine et al. 2006; Denker 2006]. The investigated materials comprise paintings and drawings, porcelain, ceramics and gems, as well as all kinds of metals. At the International Conference on PIXE and its applications held at the University of Surrey in 2010, laboratories from

Budapest, Catania, Cape Town, Debrecen, Ljubljana, Paris, Surrey, Rossendorf, and Vienna presented work on art and archaeometry (http://www.ionbeamcentre.co.uk/PIXE2010/). And this listing is by far not complete.

3 Experimental Setup for Art Objects

Most of the items needed for PIXE measurements have been described in detail in other chapters in this book. Hence, this section will give only a short overview.

The protons are produced in an ion source and accelerated to the desired energy (see ❯ Chap. 7, "Accelerators for Particle Physics"). The accelerator may be a Van-de-Graaff or Cockroft–Walton machine, or a cyclotron. A beam-line system guides the proton beam to the experiment with bending and focusing magnets. The beam-line is under vacuum, typically better than 10^{-6} mbar, in order to avoid interactions of the protons with residual gas.

The proton beam is extracted via a thin foil from the vacuum of the beam-line. As exit foils, thin aluminum, polymers, or silicon-nitrate foils are used. Thus, the art object remains in normal atmosphere. However, one has to keep in mind that the atmospheric environment in a physics laboratory may differ from the ideal conditions for the conservation of an art object.

The size of the proton beam spot on the sample depends on the beam-line: In most cases, the beam is focused to 0.5–1 mm, which is suitable for many purposes. The size is controlled by looking at a crystal emitting visible light when irradiated. At special, dedicated beam-lines a size of about 10 μm has been achieved. Fine structures, e.g., metal point drawings, need such excellent spatial resolution (Duval 2004).

The object under study is mounted on an x, y table that allows positioning of the object. As art objects rarely are solid, regular-shaped samples, customized sample holders are required. Many setups have lasers to mark the beam spot on the object. Commonly, a camera helps positioning the sample and documents the beam spot.

The x-rays are detected by solid state detectors made from single crystals. For the x-ray energy range of 1–25 keV, normally lithium-doped silicon (Si(Li)) detectors are employed. For x-ray energies above 25 keV, high-purity germanium (HPGe) detectors offer far better detection efficiency (see ❯ Chaps. 13, "Photon Detectors," ❯ 16, "Semiconductor Counters," ❯ 17, "Gamma-Ray Detectors," and ❯ 21, "New Solid State Detectors"). The signals of the detector are amplified, given to an analog/digital converter, and then processed by a multi-channel analyzer, usually combined with a computer to allow online display of the spectra (see ❯ Chaps. 2, "Electronics Part I," and ❯ 3, "Electronics Part II"). The identification of the elements present in the sample can be done during the measurements, whereas the final de-convolution of peak heights to concentrations is done off-line (see ❯ Chap. 4, "Data Analysis"). In addition, γ-rays can be measured in parallel using thick HPGe detectors with suitable detection efficiency for the expected γ-rays. In some setups, the volume between beam exit, object, and detector is flushed with helium: This does not only reduce the background originating from argon in the air, but also allows the use of a particle detector which detects backscattered protons from the sample (Rutherford Backscattering – RBS). The possibility to combine several analytical tools in the same setup is the great advantage of the PIXE technique, as additional, complementary information about the object may be obtained in one experimental run.

There are various ways to measure the intensity of the proton beam: for higher beam intensities the current on rotating metal fingers may be used or the x-rays from the exit window or from the argon in the air can be recorded in parallel to the normal PIXE measurements.

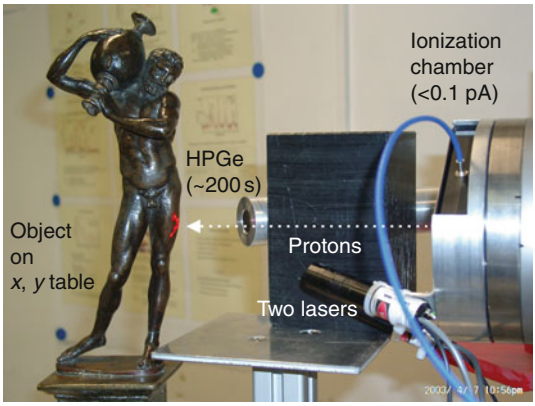


Fig. 7

Experimental setup for high-energy PIXE. The protons from the accelerator leave the vacuum of the beam-line via a thin Kapton foil. Their intensity is measured with a transmission ionization chamber. An HPGe detector is mounted under 135° with respect to the beam-line and shielded by a polyethylene block. The object under study, here an Italian Renaissance bronze statue is mounted on an *x*, *y* table with high positioning precision. Two laser cross-hairs mark the beam spot on the sample and the distance to the detector

For low beam intensities, backscattered protons from the exit foil or, at higher proton energies, transmission ionization chambers can be used. When measuring art objects the proton beam intensity is kept typically below 1 nA for a beam spot of 1 mm. This is a sufficient intensity to obtain good count rates in the x-ray detectors, and will not damage the object. For very radiation-sensitive material, like glass and porcelain, usually test irradiations are performed on modern material. If there are no observable changes in the modern material, the beam intensity is then reduced even further to ensure the safety of the art object. *Figure 7* shows a photograph of the high-energy PIXE setup at the former Hahn-Meitner-Institut (now Helmholtz-Zentrum Berlin), Germany, with an Italian Renaissance bronze statue.

4 Examples

4.1 Paintings

Today, there are portable XRF systems available that allow a detection of elements on-site in the museum. However, additional information can be obtained by the use of protons thus justifying the transport of a valuable to an accelerator laboratory. The combination of PIXE and PIGE permits the simultaneous measurement of typical x-ray elements such as Ca, Cu, and Pb as well as elements not visible in x-ray spectra like F, Na, and Al. An example for this is the analysis of the blue pigment in the"Madonna and Child Enthroned" from the Finnish National Gallery (Tuurnala and Hautojärvi 2000) by T. Tuurnala and A. Hautojärvi.

A further advantage is that the PIXE technique provides information about the depth-dependent distribution of the elements by varying the incident proton energy.

For instance, when increasing the proton energy the Pb Lα/Hg Lα ratio increases as well; this indicates a Hg-containing layer on top of a Pb-containing layer. The possibilities and constraints have been investigated by a European collaboration on various paint test samples (Neelmeijer et al. 1996). This technique was applied to Leonardo da Vinci's *Madonna dei fusi* by Grassi et al. (2005).

When the paint layers are very thick, above 100 μm, high-energy PIXE permits a nondestructive analysis. Due to the small energy loss of high-energy protons, the excitation conditions are about the same throughout the paint layers. Hence, changes in the intensity ratios Kα/Kβ of the lines of the same element yield an indication of the depth of this element. It could be shown on paint mock-ups that paint sequences, such as a white ground and on top of it layers consisting of pure pigments, can be clearly distinguished and identified. However, with increasing complexity limitations occur, e.g., if the same pigment is present in various layers (Griesser et al. 2000).

4.1.1 Flemish Painting

The painting shown in ❯ *Fig. 8* belongs to the Gemäldegalerie Berlin. It is a Flemish painting from an unknown artist, dating back to the seventeenth century and had been analyzed by neutron autoradiography at the research reactor of the Helmholtz-Zentrum Berlin. Neutron autoradiography is capable to provide two-dimensional distributions of pigments in a painting nondestructively. However, chalk, lead white, ochre, and lead–tin yellow will not cause distinct images (The Metropolitan Museum of Art 1982).

The goal of the additional PIXE analysis was therefore to look for the elements characteristic for those pigments: Ca, Fe, Pb, and Sn. In ❯ *Table 5*, the detected elements at the various spots and the corresponding pigments are listed. Pb and Fe have been detected on all analyzed spots.



Fig. 8
Flemish painting "Adoration of the shepherds", seventeenth century, unknown artist, *Gemäldegalerie Berlin*, during the measurement, as recorded by the video camera

Table 5
Measured spots, detected elements at this spot, and the corresponding pigment

Analyzed spot	Detected elements	Pigment
Blue: sky	Fe, Co, Cu, As, Pb	Smalt
Blue: window	Fe, Cu, Pb	Probably azurite
Red: coat and trouser	Fe, Hg, Pb	Cinnabar
Roof	Mn, Fe, Pb	Umber
Shepherd coat	Fe, Sn, Pb	Lead–tin yellow

For the dark parts of the painting, the intensity ratio of the various Pb lines indicates Pb in larger depths thus implying a lead-white ground. Earth pigments are based on Fe. In the blue sky, Co, together with small amounts of As, were found. As was detected by the Kβ line, as the Kα of As ($E_{K\alpha} = 10.53\,\text{keV}$) is overlapped by the Lα line of Pb ($E_{L\alpha} = 10.54\,\text{keV}$). The presence of Co and As points to smalt. The blue of the window, however, was painted with a Cu-containing pigment, probably azurite. In the red color of the trouser and the coat from one shepherd, Hg was found, suggesting the use of cinnabar. Mn, found in the spot measured on the roof, is typical for umber. As in the yellow of the coat of another shepherd, Sn was found together with Pb, most likely lead–tin yellow was used.

4.1.2 Modigliani Portrait

Whereas Old Masters could use only a limited palette, this changed in the nineteenth and twentieth century, when new, synthetic pigments became available. An example for this is the portrait, attributed to Modigliani, shown in *Fig. 9*. Below the now visible portrait of a man, a picture of a woman wearing a hat could be observed in the x-ray image of the painting. Thus, the artist reused the canvas. The question arises, if both paintings have been created by the same artist. Hence, high-energy PIXE was used to provide some clues, as the range of the protons is much larger than the thickness of the paint layers. Overall, seven points of the painting were measured.

Figure 10 shows the high-energy spectra of two points, hat and collar. Obviously, the artist (or artists) employed pigments containing a large variety of elements. On all seven measured spots, Fe, Zn, Cd, Ba, and Pb were found in varying amounts. On the gray waistcoat, Hg was found. Se could be detected on the lips. One problem arises from the omnipresence of Ba in the painting: It is thus impossible to determine Ti, because the energy of the Ti Kα line (4.51 keV) overlaps with the Ba Lα line (4.47 keV). This cannot be sidestepped by using the Kβ line of Ti – like in the example given above – as the Ti Kβ line (4.93 keV) and the Ba Lβ line (4.83 keV) overlap as well. In addition, if Ti in this painting has been used in the ground, the x-rays will be absorbed partly by the above lying paint layers. Besides the identification of the various elements at different spots on the painting, an estimation of their depth in the paint layer was made based upon the intensity ratio of the different lines from one element.

It is indeed possible that the two portraits were created by different artists, as they differ in the pigments used for the white color. The lower-lying portrait of a woman contains lead white, whereas the collar in the visible portrait of a man was executed in zinc white.



Fig. 9
Portrait attributed to Modigliani in the high-energy PIXE setup

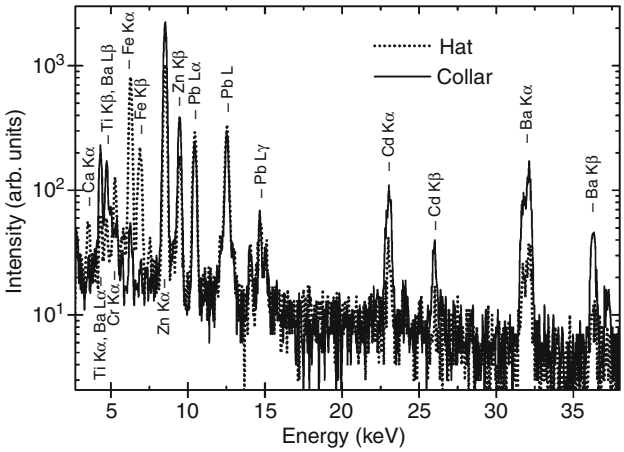


Fig. 10
High-energy PIXE spectra of the Modigliani portrait measured at the hat and at the collar

4.2 Metals

PIXE is widely applied to the analysis of metal artifacts: characterization of the alloy type, determination of the ore provenance, or gilding techniques are typical analytical tasks. One of the analytical challenges is that metals seldom have unaltered surfaces. Corrosion may occur due to long burial times in an unfavorable environment or the less noble metal can be depleted deliberately in the alloy. In addition, segregation processes may lead to inhomogeneous metal artifacts. In these cases, PIXE may provide additional information by varying the incident proton energy, thus changing the analytical depth, or by the combination with PIGE.

4.2.1 Silver Coins: Wiener Pfennig

A certain type of medieval silver/copper coins, the so-called Wiener Pfennige was minted around 1110–1395 in different parts of today's eastern Austria. It was minted by first cutting the blanks from a silver band and afterward striking them by a one-sided stroke with a hammer applying two different coining dies for the obverse and reverse sides of the coins. During their minting period (1110/20– ≈ 1395) the coins show a regular change of the coinage. These coins were classified at first by Bernhard Koch during the 1980s (Koch 1983) but some numismatic questions concerning the chronology and composition of these coins were still unanswered. In 1990, near Tulln, more than 10,000 of these coins were excavated. The major part of these coins belongs to the minting period assigned to "*Friedrich dem Schönen*" (1306/14–1330). As this finding is an authentic one and due to the high number of coins found, it seems to be perfectly suited for performing a survey. The composition was studied nondestructively on the obverse as well as the reverse with respect to:

- Enabling a new classification of the coinage and receiving a relative or even better absolute chronology for the different coin types
- Finding a possible relation between certain weight classes of Wiener Pfennige and their silver content
- Approving the assumption that coins minted in later years show a decreasing silver content in comparison to the coins minted at the beginning of the period of "Friedrich dem Schönen"

As the surface of the coins originating from the "Hoard of Tulln" had strongly been corroded due to the burial conditions, these coins were already restored, i.e., cleaned with EDTA and citric acid/ammonia solutions, to remove the corrosion products after the excavation. Therefore, the Cu content was visibly leached out in the near-surface regions of the coins (they appeared silvery and shiny) and a nondestructive investigation technique enabling the analysis of the coins' core (bulk) had to be applied.

The challenge was that a large number of coins had to be analyzed. An automatic device for positioning the coins was used enabling the investigation of more than 500 single objects within a time period of around 80 h.

In a first step, 330 coins from the hoard of Tulln and 43 coins from the same period (but from distinct finds) stored in the coin-cabinet of the *Kunsthistorisches Museum* Vienna were analyzed on the obverse as well as reverse. For the data evaluation of these coins, only x-rays of elements visible in the spectra were evaluated. With a few exceptions, the coins show the same composition on the obverse as well as on the reverse. Their average composition is 95% silver, 3% copper, and about 2% lead. As the calculated lead concentration was the same using either the Pb L line (low x-ray energy) or the lead K line (high x-ray energy) intensities for the evaluation, the Pb concentration has to be constant throughout the whole coin: Due to the tremendous difference in the x-ray attenuation and, therefore, also the analytical depth of the K and L lines, similar concentrations can only be calculated for both lines when the objects show homogeneous elemental distributions.

The composition of the "Wiener Pfennige" of the hoard of Tulln obtained during these measurements is in strong contrast to previous investigations from other findings, performed by melting a number of coins and analyzing them in a chemical way Koch 1983. The high-energy PIXE results of the 43 coins not belonging to the hoard of Tulln, however, show a composition

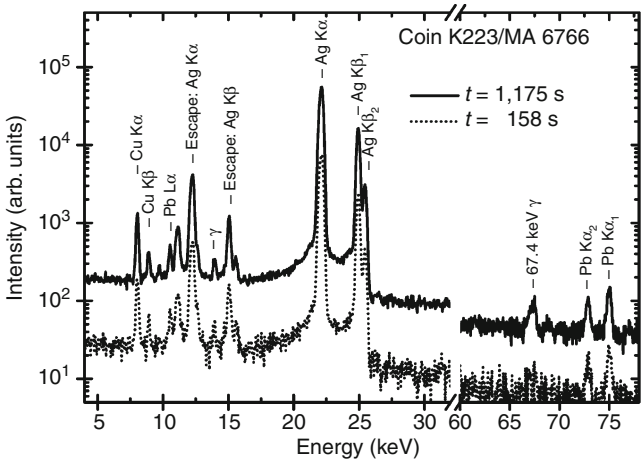


Fig. 11

High-energy PIXE spectra of coin K223/MA 6766, a *Wiener Pfennig* minted in Enns (today's Austria) about 1329/30. The x-ray lines of the main elements silver, copper, and lead can already be detected with sufficient statistics after a short measurement time of less than 200 s (*dotted line*). Increasing the counting time yields the detection of a γ line at 67.4 keV in the spectrum due to nuclear reactions of the ^{63}Cu (*solid line*)

of 10–35% copper, and 2% lead, the rest being silver – which agrees well to the chemical investigations and to the contemporary documents. However, keeping in mind that the measured Cu x-rays are mainly originating from the surface, the PIXE measurement results of the Cu content for the hoard of Tulln confirm the anticipated decrease of Cu in the near-surface layers due to the burial of the coins under detrimental conditions and presumably also due to the restoration treatment applied.

When using 68 MeV protons for the excitation of the Cu x-rays, there is the possibility of nuclear reactions. Besides the x-ray lines in Fig. 11, there is also a single line at 67.4 keV that is caused by γ radiation. This γ line is due to the formation of ^{61}Ni, created by the nuclear reaction ^{63}Cu(p,3n)^{61}Zn. The ^{61}Zn decays by two consecutive β^+ decays into the excited ^{61}Ni*. As the ^{61}Ni* atomic nucleus is still in an excited state, it relaxes to the ground state by emitting the surplus of energy as γ radiation with the well-defined energy of 67.4 keV. ^{61}Ni is a stable isotope, so no further reactions take place. The energy of this γ line is large enough to be detected from large depths, even a few hundred micrometers. The drawback is the transformation of Cu to Ni. However, only about 100,000 atoms undergo this reaction under these experimental conditions. Compared to the about 2×10^{21} Cu atoms in a 0.7 g silver coin with about 30% copper, this is negligible and the analysis can still be considered nondestructive (see Chap. 22, "Radiation Damage Effects"). For quantitative analysis, the peak intensity of the 67.4 keV γ line has to be correlated with the Cu concentration. This was done using Ag/Cu standards from ÖGUSSA (*Österreichische Gold- und Silberscheideanstalt*) as well as various Cu-containing materials from the BAM (*Bundesanstalt für Materialforschung und -prüfung*).

The probability for a nuclear reaction is much smaller than for the excitation of characteristic x-rays, hence, for sufficient statistics, longer measurement times are required.

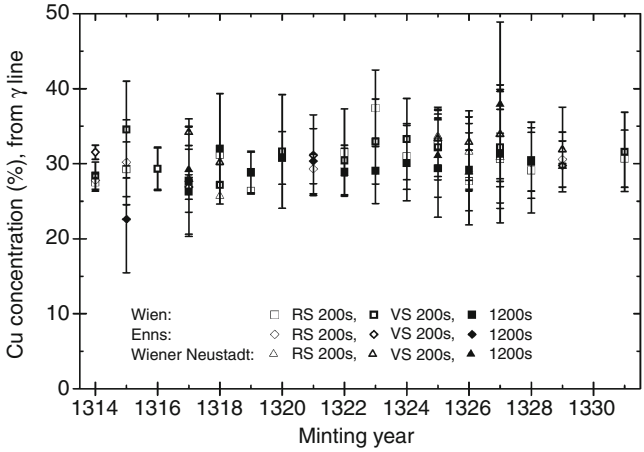


Fig. 12
Averaged copper concentrations for different minting places depending on the year of minting. No increase in the copper concentration can be observed

Therefore, in a second run, 180 coins were measured for about 1,200 s in order to increase the signal of the 67.4 keV γ line present in the spectra. Using this line for the determination of the Cu content, an analytical depth of nearly 400 μm can be achieved. The thickness of the coins is less than one millimeter; therefore, this line provides information about the inner part of the coins. The Cu content obtained by the γ line varied quite strongly from coin to coin: For one coin the intensity of the γ line was, even after 1,200s, not high enough, so we obtained a Cu concentration of zero. For another single coin, the copper concentration is 67%. For all other coins copper concentrations around 30% are obtained, which fits very well to the previous studies.

The large scattering of the Cu concentrations of single coins makes an analysis of the development of the Cu content as a function of minting time very difficult. As groups of ten coins per minting place and minting year have been analyzed, it was possible to average over these ten coins. *Figure 12* shows the average and standard deviation for each group of coins. For this figure, also included are those spectra of the first measurement campaign whose statistics for the γ line were good enough. The comparison gives an excellent agreement between long and short time measurements, as well as for measurements on the obverse and reverse. No differences between the different minting places could be observed. Looking at the time evolution of the copper concentration, no increase of the copper content is observed over the rather short minting period studied.

4.2.2 Gold Scarab

The *Ägyptische Museum und Papyrussammlung Berlin* owns a small Egyptian scarab (*Fig. 13*). The analytical questions were: What is the composition of the gold? Is the scarab solid or is it a gilded object? On regular-shaped objects, the latter question can be easily answered by determining the volume and the weight. High-energy protons excite K and L lines of Au in a sample. The absorption of the lines is very different, thus providing information from different depth

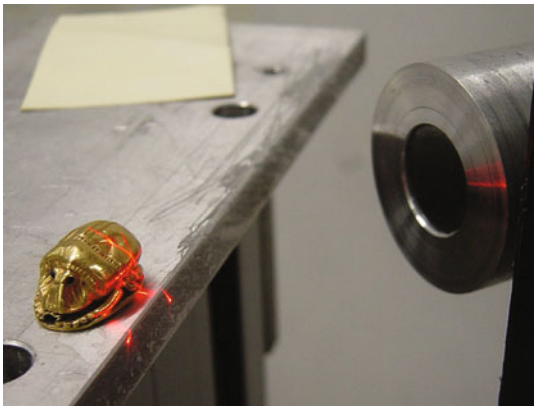


Fig. 13
Egyptian scarab during the high-energy PIXE measurements. The laser cross-hairs mark the analyzed spot and serve for precise positioning. On the right, the detector nozzle is visible

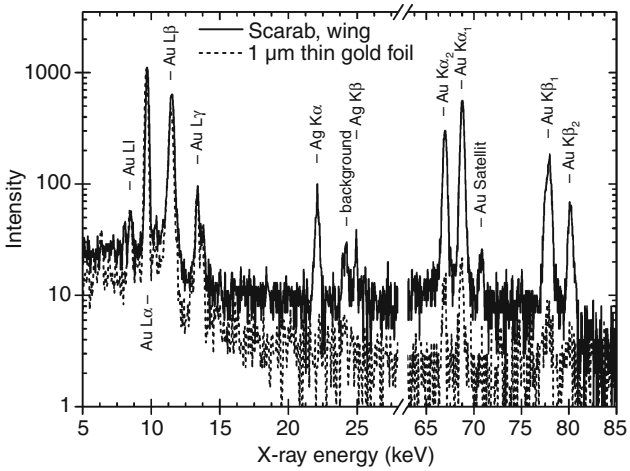


Fig. 14
High-energy PIXE spectra of an Egyptian scarab (*solid line*) in comparison with the spectra of a 1-µm thin gold foil (*dashed line*). The large intensity of the Au K lines shows that the scarab is made of massive gold

layers. Measurements were performed on the wing and the bellow of the scarab. The quantitative analysis yields a concentration of 95% Au and 5% Ag. The relative error in the concentration is around 15%, due to the irregular shape of the object leading to uncertainties in the angle of the incident proton beam as well as the angle toward the detector. The strong presence of the K lines of Au in the spectrum (*Fig. 14*) indicates the use of a massive gold. Hence, with a measurement of only a few minutes, not only the composition, but also information about the manufacturing could be obtained.

5 Conclusions

PIXE applied to objects from cultural heritage can provide, in a nondestructive way, information concerning age of the object, composition, manufacturing, or provenance. The combination with PIGE in the same experimental setup provides further, additional information. Depth-resolved analysis is possible when varying the incident proton energy.

The composition of the artifacts covers a broad range of materials. The objects are often unique, hence, simple solutions are neither available for positioning nor for data analysis. As ancient objects are scarcely homogenous or have a smooth surface, special care has to be taken of possible errors in the quantification.

As a final conclusion, I would like to state: Measurements on objects d'art are a fascinating and truly interdisciplinary field.

6 Cross-References

- Chapter 2, "Electronics Part I"
- Chapter 3, "Electronics Part II"
- Chapter 4, "Data Analysis"
- Chapter 17, "Gamma-Ray Detectors"
- Chapter 21, "New Solid State Detectors"
- Chapter 22, "Radiation Damage Effects"
- Chapter 26, "Accelerator Mass Spectrometry and its Applications in Archaeology, Geology and Environmental Research"

References

Bambynek W, Crasemann B, Fink RW, Freund HU, Mark H, Swift CD, Price RE, Veugopalo P (1972) X-ray fluorescence yields, Auger, and Coster-Kronig transition probabilities. Rev Mod Phys 44:716

Bearden JA (1967) X-ray wavelengths. Rev Mod Phys 39:78

Beckhoff B, Kanngießer B, Langhoff N, Wedell R, Wolff H (2006) Handbook of practical x-ray fluorescence analysis. Springer, New York

Boutaine JL (2006) The modern museum. In: Bradley J, Creagh D (eds) Physical techniques in the study of art, archaeology and cultural heritage, vol 1. Elsevier, Amsterdam

Bradley D, Creagh D (eds) (2006/2007) Physical techniques in the study of art. Archaeology and Cultural Heritage, vols 1 and 2. Elsevier Science & Technology, Amsterdam

Brandt W, Lapicki G (1981) Energy-loss effect in inner-shell Coulomb ionization by heavy charged particles. Phys Rev A23:1717

Calligaro T, S Colinart, Poirot JP, Sudres C (2002) Combined external-beam PIXE and μ-Raman characterisation of garnets used in Merovingian jewellery. Nucl Instrum Methods Phys Res B 189:320

Campbell JL, Maxwell JA, Teesdale WJ (2005) The guelph-pixe software package-II. Nucl Instrum Methods B 95:407–421

Chadwick J (1912) Philos Mag 24:594

Clayton E (1986) PIXAN: the Lucas heights PIXE analysis computer package, AAEC/M113

Coster D, De Kronig RL (1935) New type of Auger effect and its influence on the x-ray spectrum. Physica 2(1–12):13–24

Demortier G, Adriens A (eds) (2000) Ion beam study of art and archaeological objects. European Communities, Luxembourg

Denker A, Bohne W, Campbell JL, Heide P, Hopman T, Maxwell JA, Opitz-Coutureau J, Rauschenberg J, Röhrich J, Strub E (2005) High-energy PIXE using very energetic protons: quantitative

analysis and cross sections. X-Ray Spectrom B 34:376–380
Denker A, Adriaens A, Dowsett M, Giumlia-Mair A (eds) (2006) COST action G8: non-destructive testing and analysis of museum objects. Fraunhofer IRB Verlag, Stuttgart, ISBN 978-3-8167-7178-4
Duval A, Guicharnaud H, Dran JC (2004) Particle induced X-ray emission: a valuable tool for the analysis of metalpoint drawings. Nucl Instrum Methods Phys Res B 226:60–74
Gaschen A, Krähenbühl U (2005) PSI laboratory of radiochemistry and environmental chemistry, Annual Report. http://lch.web.psi.ch/files/anrep05/abstract05.html
Goldstein JI, Newbury DE, Echlin P, Joy DC, Romig AD, Lyman CE, Fiori C, Lifshin E (1992) Scanning electron microscopy and x-ray microanalysis. Plenum, New York
Grassi N, Migliori A, Mandò PA, Calvo del Castillo H (2005) Differential PIXE measurements for the stratigraphic analysis of the painting *Madonna dei fusi* by Leonardo da Vinci. X-Ray Spectrom 34:306–309
Griesser M, Denker A, Musner H, Maier KH (2000) Non-destructive investigation of paint layer sequences. In: Roy A, Smith P (eds) Tradition and innovation – advances in conservations, Contributions to IIC Melbourne Congress 82
Hubbell H, Seltzer SM (1995) Tables of x-ray mass attenuation coefficients and mass energy-absorption coefficients, originally published as NISTIR 5632. National Institute of Standards and Technology, Gaithersburg
Johansson SAE, Campbell JL (1988a) Pixe: a novel technique for elemental analysis. Wiley, New York, p 12
Johansson SAE, Campbell JL (1988b) Pixe: a novel technique for elemental analysis. Wiley, New York
Johansson TB, Akelsson KR, Johansson SAE (1970) X-ray analysis: elemental trace analysis at the 10-12 g level. Nucl Instrum Methods 84:141
Johansson SAE, Campbell JL, Malmqvist K (1995) Particle-induced X-ray emission spectrometry (Pixe). Wiley, New york
Janssens K, Van Grieken R (eds) (2005) Non-destructive micro analysis of cultural heritage materials. In: Comprehensive analytical chemistry, vol 42. Elsevier Science, Amsterdam
Kardjilov N, Lo Celso F, Donato DI et al (2007) Nuovo Cimento della Societa Italiana di fisica. C-Geophys Space Phys 30-1:79–83
Koch B (1983) Der Wiener Pfennig. Ein Kapitel aus der Periode der regionalen Pfennigmünze (Numismatische Zeitschrift 97), Wien
Krause MO (1979) Atomic radiative and radiationless yields for K and L shells. J Phys Chem Ref Data 8:307
Lang J (ed) (2005) Radiography of cultural material, 2nd edn. Oxford, Butterworth-Heinemann
Lapicki G (2009) Evaluation of cross sections for Lα x-ray production by up to 4 MeV protons in representative elements from silver to uranium. J Phys B Atom Mol Opt Phys 42:145204
Mäder M, Neelmeijer C (2004) Proton beam examination of glass – an analytical contribution for preventive conservation. Nucl Instrum Methods Phys Res Sect B Beam Interact Mater Atoms 226(1–2):110–118
Mandó PA (1995) Advantages and limitations of external beams in application to art & archaeology, geology and environmental problems. Nucl Instrum Methods B 85:815
Moseley HGJ (1913) The high-frequency spectra of the elements. Philos Mag 26:1024
Neelmeijer C, Wagner W, Schramm HP (1996) Depth resolved ion beam analysis of objects of art. Nucl Instrum Methods B 118:338
Orlic I, Sow CH, Tang SM (1994) Experimental L-Shell X-ray production and ionization cross sections for proton impact. Atom Data Nucl Data Tables 56:159
Paul H, Sacher J (1989) Fitted empirical reference cross sections for K-shell ionization by protons. Atom Data Nucl Data Tables 42:105
Perujo A, Maxwell JA, Teesdale WJ, Campbell JL (1987) Deviation of the Kβ/Kα intensity ratio from theory observed in proton-induced X-ray spectra in the 22<Z<32 region. J Phys B Atom Mol Phys 20:4973
Respaldiza MA, Gómez-Camacho J (1997) Applications of ion beam analysis techniques to arts and archaeometry. Universidad de Sevilla, Spain
Ryan CG, Cousens DR, Sie SH, Griffin WL (1990) Quantitative analysis of PIXE spectra in geoscience applications. Nucl Instrum Methods B 49:271
Scofield JH (1974) Exchange corrections of K x-ray emission rates. Phys Rev A9:1041
Sera K, Futatsugawa S (1996) Personal computer aided data handling and analysis for PIXE. Nucl Instrum Methods B 109–110:99
Szabo G, Borbely-Kiss I (1993) PIXYKLM computer package for PIXE analyses. Nucl Instrum Methods B 75:123
Tesmer JR, Nastasi MA (1995) Handbook of modern ion beam materials analysis. Materials Research Society Handbook, Amsterdam
The Metropolitan Museum of Art (1982) Art and autoradiography: insight into the genesis of

paintings by Rembrandt, Van Dyck and Vermeer. The Metropolitan Museum of Art, New York

Tuurnala T, Hautojärvi A (2000) Original or forgery – pigment analysis of paintings using ion beams an ionising radiation. In: Demortier G, Adriaens A (eds) Ion beam study of art and archaeological objects. European Commission EUR 19218 21. Office for Official Publications of the European Communities, Luxembourg

Vekemans B, Jensens K, Vincze L, Adams F, Van Espen P (1994) Analysis of X-ray spectra by iterative least squares (AXIL): new developments. X-Ray Spectrom 23:278

Ziegler JF, Biersack JP, Littmark U (1985) The stopping and range of ions in solids, stopping and range of ions in matter, vol 1. Pergamon, New York

Zucchiatti A, Bouquillon G, Lanterna F, Lucarelli PA, Mandò P, Prati J, Salomon MG (2002) Vaccari, PIXE and μ-PIXE analysis of glazes from terracotta sculptures of the della Robbia workshop. Nucl Instrum Methods B 189:358

Part 4

Applications of Particle Detectors in Medicine

35 Radiation-Based Medical Imaging Techniques: An Overview

John O. Prior[1] · *Paul Lecoq*[2]
[1]Centre Hospitalier Universitaire Vaudois and University of Lausanne, Lausanne, Switzerland
[2]CERN, Geneva, Switzerland

1 Introduction 859

2 Nuclear Medicine and Molecular Imaging 859
2.1 Sensitivity Versus Resolution in Imaging 860
2.2 SPECT Versus PET 861

3 Single-Photon Emission Computed Tomography (SPECT) 861
3.1 Conventional Gamma Camera and SPECT 861
3.2 SPECT/CT 863
3.3 Dedicated SPECT Systems 864
3.3.1 Ultrafast Dedicated Cardiac Camera 864
3.3.2 Breast Imaging with SPECT/CT 867
3.3.3 Single-Pinhole and Coded Aperture Collimation Systems 867
3.4 SPECT/MR 867
3.5 Other Detector Types 867

4 Positron Emission Tomography 868
4.1 Standalone PET Imaging 868
4.1.1 Sensitivity 868
4.1.2 Resolution 868
4.1.3 Quantitative Imaging 868
4.2 Hybrid/Multimodality PET Imaging 869
4.2.1 The Success Story of PET/CT 869
4.2.2 Time-of-Flight (TOF) PET 871
4.2.3 From PET/CT to PET/MR 872
4.3 Dedicated PET Imaging Devices 873
4.3.1 PET Mammography (PEM) 873
4.3.2 Endoscopic PET Imaging 874
4.4 Other PET Detectors 874

C. Grupen, I. Buvat (eds.), *Handbook of Particle Detection and Imaging*,
DOI 10.1007/978-3-642-13271-1_35, © Springer-Verlag Berlin Heidelberg 2012

5 Radiopharmaceuticals 875
5.1 SPECT Radiopharmaceuticals 875
5.2 PET Radiopharmaceuticals 877

6 Preclinical Imaging 879

7 Conclusions 880

References 880

Further Reading 881

Abstract: This chapter will present an overview of two radiation-based medical imaging techniques using radiopharmaceuticals used in nuclear medicine/molecular imaging, namely, single-photon emission computed tomography (SPECT) and positron emission tomography (PET). The relative merits in terms of radiation sensitivity and image resolution of SPECT and PET will be compared to the main conventional radiologic modalities that are computed tomography (CT) and magnetic resonance (MR) imaging. Differences in terms of temporal resolution will also be outlined, as well as the other similarities and dissimilarities of these two techniques, including their latest and upcoming multimodality combination. The main clinical applications are briefly described and examples of specific SPECT and PET radiopharmaceuticals are listed. SPECT and PET imaging will be then further detailed in the two subsequent chapters describing in greater depth the basics and future trends of each technique (see ❯ Chaps. 37, "SPECT Imaging: Basics and New Trends" and ❯ 38, "PET Imaging: Basics and New Trends").

1 Introduction

Over the past 2 decades, nuclear medicine and molecular imaging have been taking an increasing role in clinical medicine, mostly in the diagnosis, staging, and therapy of malignant disease. These radiation-based imaging techniques provide information on tissue and organ function that is complementary to the conventional anatomic imaging modalities of computed tomography (CT) and magnetic resonance (MR) imaging. This increasing role is due in part to the significant improvements in detector technology (scintillation crystals, photodetectors, detector electronics), scanner design including coupling to morphological imaging (also called hybrid imaging), efficient 3-D iterative reconstruction algorithms, and to the availability of new radiopharmaceuticals. Dramatic improvements have been witnessed with examination time decreasing from 1 h to 10–20 min for a whole-body oncologic PET scan and from 30 min to about 4 min for a myocardial perfusion SPECT scan. This has been paralleled with higher spatial resolution and increased sensitivity and signal-to-noise ratio.

This chapter gives an overview of radiation-based medical imaging first describing nuclear medicine and molecular imaging before outlining differences and specificities of both SPECT and PET imaging (including hybrid imaging) along with a few clinical applications and future trends. Each specific technique will be described in more details in the two subsequent chapters (❯ Chaps. 37, "SPECT Imaging: Basics and New Trends" and ❯ 38, "PET Imaging: Basics and New Trends").

2 Nuclear Medicine and Molecular Imaging

Nuclear medicine has been employed for more than half a century. It relies on using radioactive molecules administered to patients for diagnostic or therapeutic purposes. Radioactive molecules behave in vivo the same way as their nonradioactive "natural" equivalent and allow following a specific organ or bodily function. It is used daily in oncology, cardiology, neurology, pediatrics, rheumatology, or orthopedics for diagnosis and therapy.

Recently, the concept of molecular imaging has been introduced and is defined by the ability to visualize and quantitatively measure the function of biological and cellular processes in vivo (Pysz et al. 2010), as opposed to conventional radiology techniques based on the morphological alterations produced by disease. Molecular imaging is a fast-growing field that holds promises for improvement in specificity and early diagnosis, as well as in tailoring medical therapy to a specific patient ("personalized medicine") and prognostic information with early therapy follow-up (Pysz et al. 2010). Such personalized medicine approach is especially important in patients with advanced disease with poor prognosis, for whom the best care is not always known in advance and depends on tumor molecular characteristics specific to each patient. To illustrate this point, the HER2/neu-targeted therapy is only effective in patients with HER2-positive breast cancer (Pysz et al. 2010). Thus, a HER2-negative breast cancer patient treated with such drugs would receive ineffective therapy bringing unwanted side effects and diminishing remaining life quality. Another advantage of molecular imaging is the ability to measure earlier the effect of a drug before overt anatomical changes can be observed by conventional radiological techniques (CT, MR). Indeed, many modern drugs have specific molecular targets.

2.1 Sensitivity Versus Resolution in Imaging

To compare the SPECT and PET radiation-based medical imaging to conventional radiology techniques such as CT or MR, it is interesting to consider the number of agent molecules per

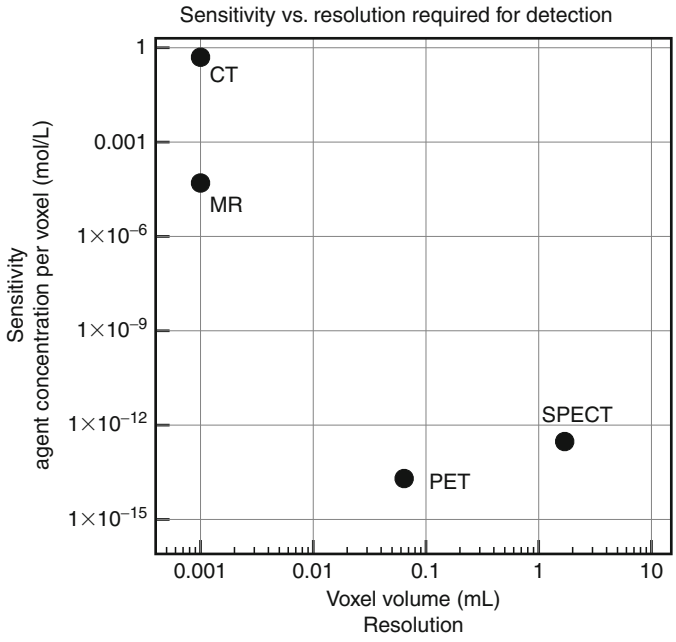


Fig. 1

Sensitivity versus resolution detection limits in clinical SPECT and PET as compared to conventional radiology techniques (CT, MR)

voxel necessary for detection (as a measure of sensitivity) versus the minimum detectable voxel volume (as a measure of spatial resolution). This has been plotted in ● *Fig. 1* from data extracted from the literature (Frangioni 2008). The respective advantages of each individual technique can be seen, with the excellent anatomic resolution of CT (millimetric resolution) and MRI and the very sensitive detection of SPECT and even better with PET (picomolar concentrations). This is the reason why integrating multimodality imaging by associating one of the anatomic imaging modality with nuclear medicine/molecular imaging modality will generally be beneficial for diagnosis, with improved image localization and functional/molecular information (Seo et al. 2008).

2.2 SPECT Versus PET

There are main differences in design and characteristics of SPECT and PET (● *Table 1*), but the most important one lies in the kind of radioisotope used. For SPECT, a single gamma ray is emitted from disintegrating nuclei, with a variable energy (from 69 keV for ^{201}Tl to 364 keV for ^{131}I). To form an image, the direction of the incoming gamma rays on the detector must be known, which is done by using a collimator. For PET, radionuclides decay by emitting a positron, a particle of the mass of the electron, but positively charged (e^+, also known as β^+, see ● Chaps. 1, "Interactions of Particles and Radiation with Matter" and ● 10, "Radiation Protection"), which annihilates rapidly by interaction with an electron to give two coincident gamma rays of opposite direction, each of 511 keV of energy. Using detectors placed around the patient working in coincidence, the direction of the two gamma rays is determined along the straight line passing by the two points of interactions.

3 Single-Photon Emission Computed Tomography (SPECT)

3.1 Conventional Gamma Camera and SPECT

SPECT has been used since the 1970s, first with 1-head gamma cameras, then with two or more heads. Using projection at different angles, SPECT allows reconstructing the 3-D distribution of the radiopharmaceutical activity. It was found to increase sensitivity and specificity as compared to simple planar projections, but the exact localization of the observed lesions remained somewhat questionable due to lack of anatomical details in SPECT images.

SPECT has a trade-off between resolution and sensitivity, which is given by the relation $\text{Sensitivity} \propto 1/(\text{FOV}/\text{Resolution})^2$, with FOV = field of view. Better performances are obtained by increasing the detector area or by limiting the FOV. As the collimator degrades resolution with the distance from the object, it is important to bring the object to image as close as possible to the detector. Thus, the collimator must be tailored for a specific purpose and optimized for a given imaging situation; many types of collimators exist (low energy general purpose LEGP, low energy high resolution LEHR, low energy high sensitivity LEHS, fan beam, pinhole, etc.), which are chosen in function of the clinical question to answer and the radioisotope energy.

Table 1

Main differences between SPECT and PET characteristics for clinical imaging with radiopharmaceuticals

Variable	SPECT	PET
Radiation type	• Emission of a single photon of variable energy (69–364 keV)	• Coincident emission of two 511-keV gamma rays
Sensitivity	• Geometric efficiency ~ 0.01%	• Geometric efficiency ~ 1%
	• Need of a physical collimation to reject photons not from a known incidence • Collimator designs with increased sensitivity (slit/slat, rotating slant hole) or increased resolution (pinhole) • Geometric design • Detector material stopping power for gamma rays	• Geometric design • Detector material stopping power for gamma rays • Signal-to-noise ratio improved by time-of-flight (TOF) capabilities
Spatial resolution in clinical images	• Resolution ~ 10–12 mm • Depends from the collimator–detector response function (CDRF), determined by: – Intrinsic detector resolution – Geometric response – Septal penetration – Septal scatter – It is distance dependent	• Resolution ~ 5–7 mm • Depends from the detector, as determined by: – Crystal composition – Crystal width – Inter-crystal scattering – Inter-crystal penetration • Photon non-collinearity • Positron range
Temporal resolution	• Low, but new 4-D reconstruction algorithms may improve observation of changing activity during tomography acquisition • Due to higher radionuclide half-life, biological phenomenon over hours to days can be observed	• Good, due to high sensitivity • List mode acquisition allows different phenomena to be observed (e.g., kinetic modeling and ventricular function in cardiac PET)
Attenuation correction	• More challenging than PET, but possible	• Correction is independent of the origin along the line of response
Detected scatter radiations	• 30–50% of emitted gamma rays	• 10–20% of coincidence events in 2-D mode (scanner with septa) • 40–60% of coincidence events in 3-D mode (latest scanner designs)
Iterative reconstruction	• Distance dependency of the CDRF can be modeled and taken into account to improve image resolution and signal-to-noise ratio	• The effects altering the spatial resolution can be modeled and taken into account to improve image resolution and signal-to-noise ratio (the so-called high-definition [HD] PET)

Table 1
(Continued)

Variable	SPECT	PET
Dual-tracers imaging	• Feasible based on multi-energy windows (e.g., ^{99m}Tc 140 keV + ^{201}Tl 75/167 keV)	• Although all photons have the same 511-keV energy, radioisotopes differentiation is feasible based on multidimensional analysis methods
Partial volume effect	• Correction possible with anatomic imaging, not routinely implemented	• Correction possible with anatomic imaging, not routinely implemented
Motion	• Correction possible	• Correction possible
Hybrid imaging	• SPECT/CT improves anatomic localization and specificity over SPECT-only • SPECT/MR has not yet been implemented in the clinics	• PET/CT improves diagnostic accuracy by 10–15% over all cancers • PET-MR hybrid scanners for sequential imaging start in clinical research • PET/MR simultaneous imaging is possible for the brain (PET insert in clinical MR scanner)

3.2 SPECT/CT

The first clinical prototype of SPECT/CT was developed in the mid-1990s (Seo et al. 2008), but it was only after the success of PET/CT that SPECT/CT became commercialized in 2004 (see Sect. 4.2.1 on PET/CT below).

In clinical applications, SPECT/CT has been found to be particularly useful in localizing disease in tumor imaging thanks to the supplementary anatomic information. This is of great help to the clinicians for oncologic applications who need to decide if a particular uptake could be from benign or tumoral processes (*Table 2*). Similar applications have been found in orthopedics, infection and inflammation, pulmonary function, endocrinology (Seo et al. 2008), as well as in musculoskeletal or infection imaging, where the better lesion localization increases specificity, as it allows distinguishing degenerative/physiologic changes from pathologic ones (*Figs. 2*, *3*, and *5*) (Bybel et al. 2008).

Nuclear cardiology has also benefited from CT-based attenuation correction allowing differentiating decreased inferior wall perfusion due to ischemia or infarct from a possible diaphragmatic attenuation of photons, thus improving specificity without any decrease in sensitivity (*Fig. 4*). However, great attention should be devoted to possible coregistration errors between the SPECT and the CT due to respiration or patient motion, as they can induce false-positive artifacts for decreased perfusion, starting with 1-pixel shift in the SPECT image.

A developing application is the use of SPECT/CT for calculating patient-specific radiation dose delivered by radioimmunotherapy or in local treatment of liver cancer and liver metastases with selective internal radiation therapy (SIRT). Indeed, CT allows greater accuracy with photon attenuation correction as well as using the patient's own anatomy rather than a generalized anatomic model from the Medical Internal Radionuclide Dosimetry (MIRD) methodology.

Table 2

Clinical applications improved by hybrid imaging with SPECT/CT over SPECT-only scanners

Application	Indication
Oncology: – Bone metastases – Endocrine tumors – Thyroid cancer – Adrenal tumors – Sentinel lymph node Benign bone pathology Infection or inflammation Parathyroid	Increased specificity and diagnostic confidence due to increased ability to distinguish pathologic from degenerative or physiologic changes (Fig. 2), as well as improved localization (Figs. 3 and 5)
Nuclear cardiology	Increase in specificity due to attenuation correction in myocardial perfusion scintigraphy (Fig. 4)
Radiation dose estimation for radioimmunotherapy or selective internal radiation therapy (SIRT) with microspheres	Use of the patient own anatomy and attenuation correction to estimate and optimize the radiation dose delivered to the tumor and to the dose-limiting critical organs (e.g., bone marrow, kidneys, or lung).

SPECT/CT with appropriate attenuation correction has the potential to give quantification that is more accurate. This would improve the ability to determine the response to therapeutic interventions. Practically, quantification of radiopharmaceutical activity by SPECT/CT is challenging and necessitates deriving a proportionality constant derived from the number of counts in a standard (Seo et al. 2008). In contrary, many factors can influence these measurements, such as attenuation, scatter, partial volume effect, although specific compensation methods have been developed. However, in contrary to PET/CT where attenuation correction is performed for each examination, SPECT/CT images are corrected for attenuation only in selected cases in clinical practice (Papathanassiou and Liehn 2008).

So far, SPECT/CT scanners have not achieved the same supremacy over SPECT-only as compared to PET/CT over PET-only scanners. This might be due to the larger spectrum of clinical applications in gamma cameras as compared to PET/CT (mainly in oncology). Major enhancements in sensitivity are expected from the newer solid-state detectors and novel radiopharmaceuticals currently in development. Thus, SPECT/CT clinical applications are likely to expand in the coming years, but long-term clinical and economic benefit of this new technology will have to be determined.

3.3 Dedicated SPECT Systems

3.3.1 Ultrafast Dedicated Cardiac Camera

Most clinical gamma cameras used today rely on the principle set by Hal Anger in the late 1950s, with a collimator placed in front of a detector constituted of a crystal coupled to photomultipliers. However, as myocardial perfusion scintigraphy is a very frequently used test in nuclear

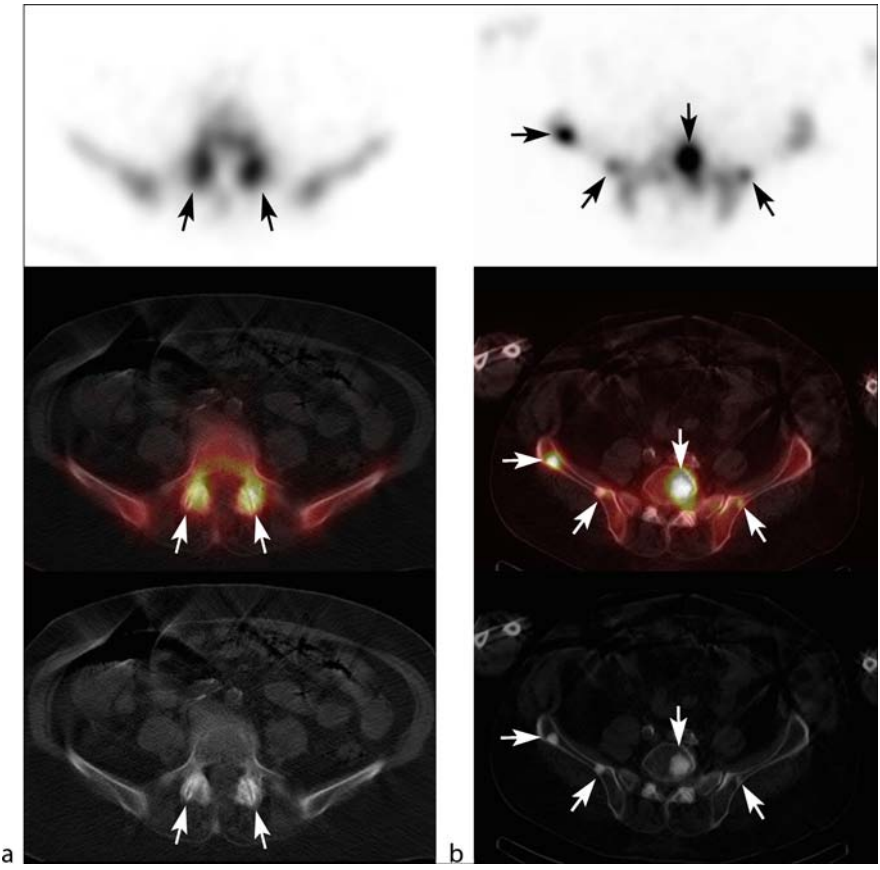


Fig. 2

Improved ability to distinguish degenerative from metastatic changes in bone lesions when using SPECT/CT: (a) 79-year-old woman with breast cancer and degenerative changes (*arrows*); (b) 78-year-old man with prostate cancer and metastases (*arrows*)

medicine, novel, ultrafast cardiac SPECT cameras have emerged as a viable imaging alternative with improved sensitivity and better resolution than conventional SPECT (Chap. 37, "SPECT Imaging: Basics and New Trends").

These cameras are based on cadmium zinc telluride (CZT) detectors based on direct conversion of incident radiation energy into charge carriers resulting in improved resolution and sensitivity. This allows decreasing injected activity (thus patient radiation dose) and study time (less patient motion), while improving image quality (*Fig. 4*). These cameras also benefit from improved, heart-centered collimator geometries, which further enhance sensitivity allowing study time reduction from 20–25 min down to 2–4 min. The improved energy discrimination makes dual-isotope protocols (injection of ^{201}Tl at rest followed by ^{99m}Tc perfusion agent at stress) feasible with only a single acquisition for both radiopharmaceuticals resulting

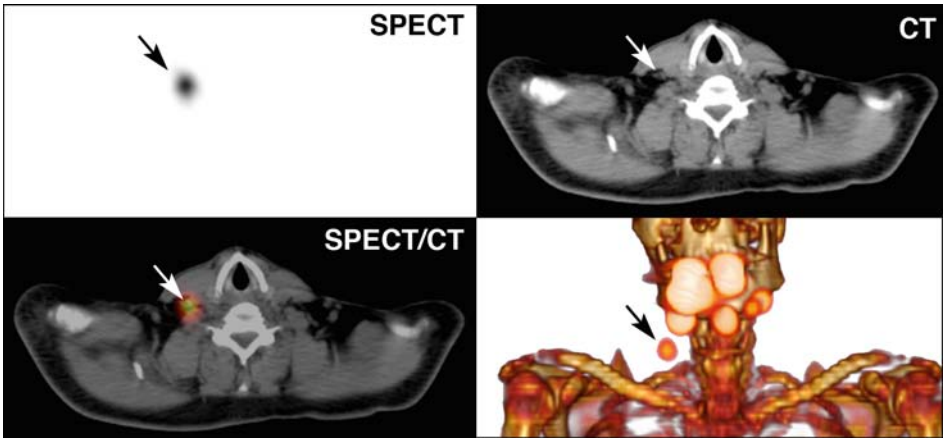


Fig. 3
Improved localization on the fused image (*lower left*) of a specific sentinel lymph node (*arrow*) by SPECT/CT over SPECT-only in a 54-year-old man with inferior lip melanoma

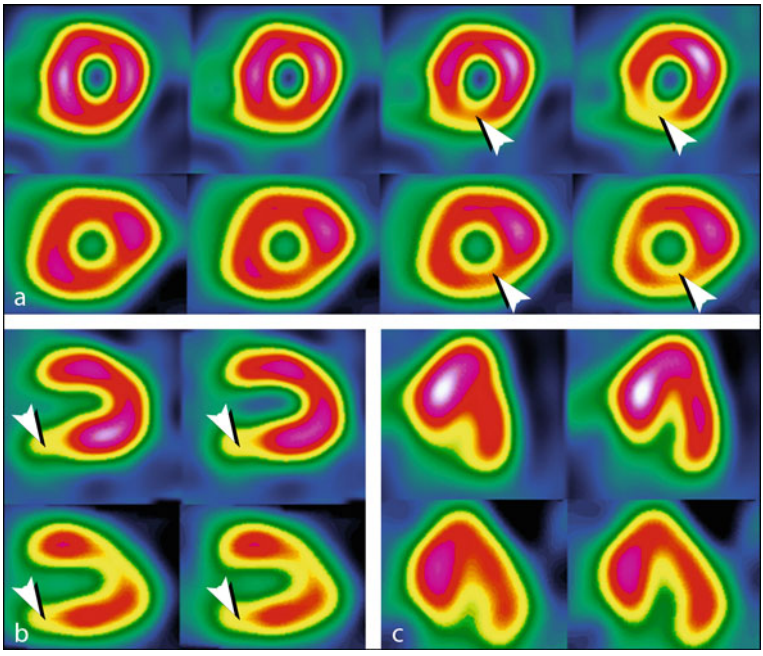


Fig. 4
Myocardial perfusion scintigraphy in the (a) short-axis, (b) vertical-axis, and (c) horizontal long-axis views in a 70-year-old woman with dedicated cardiac SPECT scanner (*upper row*) and conventional scanner (*lower row*) with a myocardial scar (*arrowhead*). Note the improved dedicated scanner image quality obtained with shorter acquisition time (8 min vs. 12 min) and lower activity (^{201}Tl 111 MBq vs. 62 MBq) (courtesy of B. Songy, M.D. [Centre Cardiologique du Nord, Saint-Denis, France])

in intrinsically better coregistration of rest and stress studies and increased patient comfort (shorter study).

3.3.2 Breast Imaging with SPECT/CT

There has been an attempt to develop multimodality dedicated scanners for the breast (SPECT/CT) using coupled CsI(Tl) phosphor and CZT detectors rotating around the uncompressed pendant organ, but this scanner has not entered the clinical arena (Tornai et al. 2003). Dedicated positron emission mammography (PEM) is emerging as an interesting alternative (see ❯ Sect. 4.3.1).

3.3.3 Single-Pinhole and Coded Aperture Collimation Systems

Pinhole imaging has been used since many years in conventional planar scintigraphy, providing better resolution images of small organs (e.g., thyroid, hip joint, etc.) when placed close to the pinhole. SPECT reconstruction with pinhole is possible and predominantly used in small-animal scanners (see ❯ Sect. 6 below). Promising multi-pinhole scanner designs are being considered for clinical stationary SPECT (Beekman and Have 2006). They offer a potential for significant improvement in performances, with a possible system resolution approaching 3 mm for a clinical machine.

3.4 SPECT/MR

The combination of SPECT and MR imaging has additional technical challenges as compared to PET/MR (see ❯ Sect. 4.2.3 below). In addition to high-magnetic-field-compatible detectors, the collimator material used for SPECT is likely to generate eddy currents, hence MR artifacts from collimator movements (Cherry 2009). Although attempts at designing small-animal SPECT/MR imaging systems were made, there is currently no available prototype of a clinical SPECT/MR scanner. An opportunity may arise with newer SPECT systems based on electronic collimation (also called Compton camera), which has been existing in theory for over 2 decades, but not built in practice because of the limited energy resolution of the available detectors (Kohara et al. 2008).

3.5 Other Detector Types

The new CdZnTe (CZT) semiconductor detectors allow better resolution with higher packing fraction and may allow to reach geometrical efficiency in the order of 0.3% (30 times better) with reconstructed resolution inferior to 5 mm for instance in a dedicated brain imager when this technology will be deployed on clinical head or whole-body clinical scanners (Jansen and Vanderheyden 2007).

4 Positron Emission Tomography

4.1 Standalone PET Imaging

Conventional PET imaging uses scanners designed as ring scintillation detectors with coincidence detection of 511-keV annihilation photons (Chap. 38, "PET Imaging: Basics and New Trends"). The detector typically uses a crystal with a high stopping power for the 511-keV gamma rays and produces light that a photomultiplier and its associated electronics can transform into a meaningful signal for image reconstruction (Lewellen 2008). Over the years, significant progresses have been made to scintillating materials (Chap. 15, "Scintillation Counters"), photodetectors and their readouts, and electronics in order to palliate factors that are known to degrade spatial resolution and to improve detection sensitivity (Chap. 38, "PET Imaging: Basics and New Trends").

4.1.1 Sensitivity

The sensitivity of a PET system is determined by its intrinsic and geometric efficiencies. High intrinsic sensitivities are already obtained with crystals with high stopping power such as BGO, LSO, and LYSO that are stopping 70–90% of annihilation photons. Better geometric efficiencies can be attained by increasing the axial field of view at the expense of increased system costs or by using new block detector concepts in order to decrease the amount of dead space between crystals. Future scanner design exists where the majority of the body is covered in one bed position. Such systems coupled with time-of-flight (TOF; see Sect. 4.2.2; see also Chap. 6, "Particle Identification") capability lead to expected improvement in signal-to-noise ratio of about one order of magnitude allowing to reduce study time and/or injected radiopharmaceutical activity.

4.1.2 Resolution

Resolution is mainly affected by detector size, positron range (*Table 3*), collinearity of the annihilation photons, and depth of interaction in the detector, among others. The main effect in clinical scanners is a collinearity, which is proportional to scanner diameter, reaching for instance 2.2 mm for a 100-cm ring for ^{18}F-based radiopharmaceuticals.

The signal-to-noise ratio is also influenced by the event type: only *true* events originating from the detection of both annihilation photons participate to the true signal formation. Other *single* events arriving within the coincidence window are also examined for energy and discarded if outside the energy acceptance window (typically 15–30% or 350–650 keV), but still a number of *random* and *scatter* counts are accepted leading to degrading signal-to-noise ratio (Lecomte 2009).

4.1.3 Quantitative Imaging

Modern image reconstruction algorithms use iterative methods such as ordered-subset expectation maximization together with a posteriori information and eventually the underlying physical model of the scanner to incorporate corrections for dead time, detector normalization, photon scatter, random events, and photon attenuation (Chap. 39, "Image Reconstruction").

Table 3

Positron range of major PET radioisotopes (listed by increasing positron energy)

Radioisotope	Positron energy (MeV)		Positron effective range[a] (mm)
	Maximum	Average	
^{18}F	0.64	0.24	0.54
^{64}Cu	0.65	0.28	0.55
^{11}C	0.96	0.39	0.92
^{13}N	1.20	0.49	1.4
^{15}O	1.73	0.74	2.4
^{68}Ga	1.90	0.84	2.8
^{82}Rb	3.36	1.52	6.1

[a]Defined as the average distance from the decaying nucleus to the end of the positron range measured perpendicular to the line defined by the two 511-keV annihilation photons

4.2 Hybrid/Multimodality PET Imaging

4.2.1 The Success Story of PET/CT

After a phase of introduction and clinical testing of a PET/CT prototype since 1998, the first commercial PET/CT was introduced in 2001 (Townsend 2008; Mawlawi and Townsend 2009). The advantages of hybrid PET/CT scanners over PET-only scanner led the major vendor to no-longer offer PET-only systems as early as 2006. Currently, commercial scanners are mostly using GSO, LSO, and LYSO, which offer better 3-D performance for whole-body imaging (Mawlawi and Townsend 2009).

The introduction of hybrid scanners combining a radiation detector with morphological information obtained from computed tomography (CT) has acted as a catalyst for the development of the nuclear medicine/molecular imaging market. PET/CT has become in a decade a standard diagnostic imaging modality and has led the way to other multimodality imaging modalities (first SPECT/CT and now PET/MR) (Blodgett et al. 2007).

Currently, PET/CT imaging has been proved more sensitive and specific than PET or CT alone for staging and restaging in a variety of cancer (*Table 4*), with significant improvement in accuracy ranging between 10% and 15%. PET/CT is considered as an essential imaging modality for cancer at present time (*Fig. 5*). Even for cancer not currently reimbursed in the United States, PET/CT led physicians to change 36.5% of their intended management after performing this examination (n = 22,975 PET or PET/CT studies in 1,178 centers) (Hillner et al. 2008).

Another developing application of PET/CT is the target volume definition of radiation therapy with the emphasis on conformal and intensity-modulated techniques necessitating more precise volume definitions. For the first time, tumor biology information can be integrated in radiation therapy planning possibly resulting in more efficient and effective radiation dose delivery to the tumor and sparing of normal tissue. However, the definition of the threshold selection for PET image volume contouring during tumor segmentation remains challenging and no method has gained wide acceptance so far, although a 40–50% fixed threshold is used in many settings, with a risk of overestimating tumor size in small lesions due to the partial volume effect. The clinical impact of targeting biologically active tumor volume remains to be determined.

Table 4

Clinical applications improved by hybrid imaging with PET/CT over PET-only scanners

Cancer	Indication
Head and neck (squamous cell carcinoma)	Initial diagnosis, staging, restaging, prognosis
Thyroid (aggressive form)	Initial diagnosis, staging, and restaging
Lung cancer	Initial diagnosis of single pulmonary nodule; staging and restaging
Esophageal cancer	Staging, restaging, and prognosis
Breast cancer	(Staging), restaging, and response to therapy
Colorectal cancer	Staging and restaging
Non-Hodgkin and Hodgkin lymphoma	Initial diagnosis, staging, restaging, and response to therapy
Melanoma	Staging and restaging
Unknown primary tumor	Initial diagnosis, staging, and restaging
Incidental, unsuspected cancer	Initial diagnosis

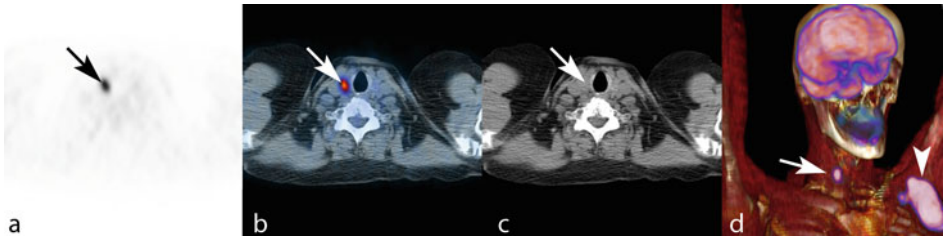


Fig. 5

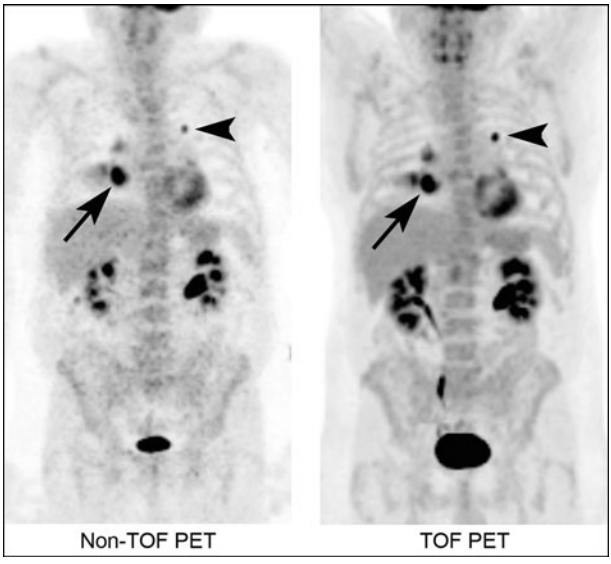
PET/CT in a 59-year-old woman with left shoulder melanoma showing incidental increased ^{18}F-fluorodeoxyglucose uptake in the right thyroid lobe (*arrow*); (a) PET, (b) PET/CT, and (c) CT image. Note the improved localization on fusion not obtainable from PET or CT image separately. (d) 3-D fused image with voluminous axillary metastases (*arrowhead*). The incidental thyroid lesion was a papillary thyroid carcinoma

The clinical adoption of PET/CT has been fast in many clinical applications in therapy management and follow-up. The cost-effectiveness of PET and PET/CT has been established for specific indications such as the investigation of a single pulmonary nodule, for non-small-cell lung cancers, for restaging colorectal cancers and lymphomas, for head and neck cancers, for pancreatic cancer, as well as in the clinical management of patients with suspected recurrent ovarian cancer (Buck et al. 2010; Lee et al. 2010). However, not all PET/CT applications are cost-effective, as for instance the surveillance by PET/CT of patients with Hodgkin lymphoma at the first relapse, where CT is generally adequate (Mansueto et al. 2009). More economic evaluation will need to be performed in the future to see if the PET/CT diagnostic supremacy will apply to cost-effectiveness.

4.2.2 Time-of-Flight (TOF) PET

There has been renewed interest in time-of-flight (TOF) PET imaging with the arrival of fast scintillators with high stopping power such as LSO and LYSO. The first commercial TOF PET/CT system was introduced in 2007 (Conti 2009). Nowadays, all major vendors offer scanners with TOF capabilities that improve the signal-to-noise ratio. This is done by measuring the difference between the two gammas' arrival time at the detectors to constrain the event in space (Mawlawi and Townsend 2009). For instance, with a coincidence timing resolution of 500 ps, the spatial uncertainty reaches 7.5 cm leading to a signal-to-noise ratio increase of about 2.3 for a 40-cm diameter uniform distribution. Key design parameters can be optimized specifically for TOF PET, including scintillators, crystal shape and surface, photodetectors, electronics, image reconstruction, and emission data correction (Conti 2009).

Benefits from TOF PET over conventional, non-TOF PET are decreased image noise and better image resolution. However, the improvement in the signal-to-noise ratio is correlated with patients' size, where better localization of annihilation events is obtained in larger patients where the reduction of photon count due to attenuation is most apparent (◉ *Fig. 6*) (Karp et al. 2008). Similarly, dynamic acquisitions with short frames substantially benefit from TOF PET, with better algorithm convergence due to less image noise. Finally, image quality may be kept at the same level as non-TOF PET and the signal-to-noise gain can be used for reducing acquisition time or injected activity for the patient comfort or study costs (Conti 2009).



◘ Fig. 6

Maximal intensity projection images of a 66-year-old patient with lung cancer (*arrow*) on a non-time-of-flight (TOF) PET scanner (*left*) as compared to the latest generation TOF PET scanner (*right*) 90 min after ^{18}F-fluorodeoxyglucose injection. Note the improved image quality of the TOF PET with decreased noise level and better visualization of the contralateral lesion (*arrowhead*) (courtesy of F. Corminboeuf, Ph.D. [Bern University Hospital, Switzerland])

The latest PET/CT designs draw upon improvements in sensitivity due to larger axial field of view, as well as in image quality due to the integration of the detector spatial response function (high-definition PET) and reduction of image blur due to respiratory motion. However, many challenges remain in improving the use of PET/CT in clinical imaging, such as the correction of the underestimation of activity concentration in small lesions (partial volume effect) affecting lesions size up to 2.5–3.0 times the PET resolution (15–20 mm) (Soret et al. 2007).

4.2.3 From PET/CT to PET/MR

Most commercial clinical PET systems still use photomultipliers as light detector, but the latest advances in detector technology is to replace them by novel semiconductor light detectors, especially for hybrid PET/MR scanners, as photomultipliers do not work in high-magnetic-field environment (Lewellen 2008; Pichler et al. 2008; Lecomte 2009). Among these new semiconductor-based light detectors, there are avalanche photodiodes (APDs) and silicon photomultipliers (siPMs), which are a new generation of Geiger-mode APDs; both are insensitive to magnetic fields (Pichler et al. 2008).

Several drawbacks from PET/CT can be listed: (1) PET and CT are not acquired simultaneously (misregistration due to organ motion); (2) CT adds radiation dose; and (3) CT only provides low soft-tissue contrast as compared to MR. Basically, three architectures of PET/MR can be envisioned at the clinical and preclinical, small-animal level (Delso and Ziegler 2009):

1. The sequential architecture is the easiest one with PET and MR scanner separated in space like the current PET/CT machines, with a common bed sliding from one machine to the other, with obvious advantages in electromagnetic shielding and costs, but with long total examination times, coregistration artifacts, and large room-size requirements. The first PET/MR machines were installed starting 2010 (*Fig. 7*).
2. The insert architecture where a removable PET detector is put inside a conventional MR scanner allowing simultaneous PET/MR imaging with reduced total costs when the MR system is already available, but with a reduced field of view allowing only brain studies.
3. The integrated architecture that is most technologically challenging with a PET detector inserted between the MR coils.

As of today, there is no clinical machine built for whole-body simultaneous PET/MR imaging. Many advantages are expected from such a simultaneous acquisition, notably a reduction of total study time, the possibility to correct for motion during PET acquisition using motion vectors extracted from MR motion-sensitive scans (Pichler et al. 2008), or to correct for partial volume effect using accurate morphologic information (Meltzer et al. 1990). In addition, a wealth of information can also be gained from simultaneous PET/MR acquisitions using at best the strength of each technique or to correlate the same parameter for cross-validation. Such whole-body simultaneous PET/MR is the subject of recently funded ambitious projects of the 7th European Framework Program such as "Hyper-Image" (www.hybrid-pet-mr.eu) or "Sublima" (http://www.sublima-pet-mr.eu).

It is still too early to tell whether PET/MR will replace PET/CT in clinical applications or remain reserved to specialized research centers. This will depend on the application that will be developed and the possible specific advantages of this technique, as well as cost-effectiveness issues.

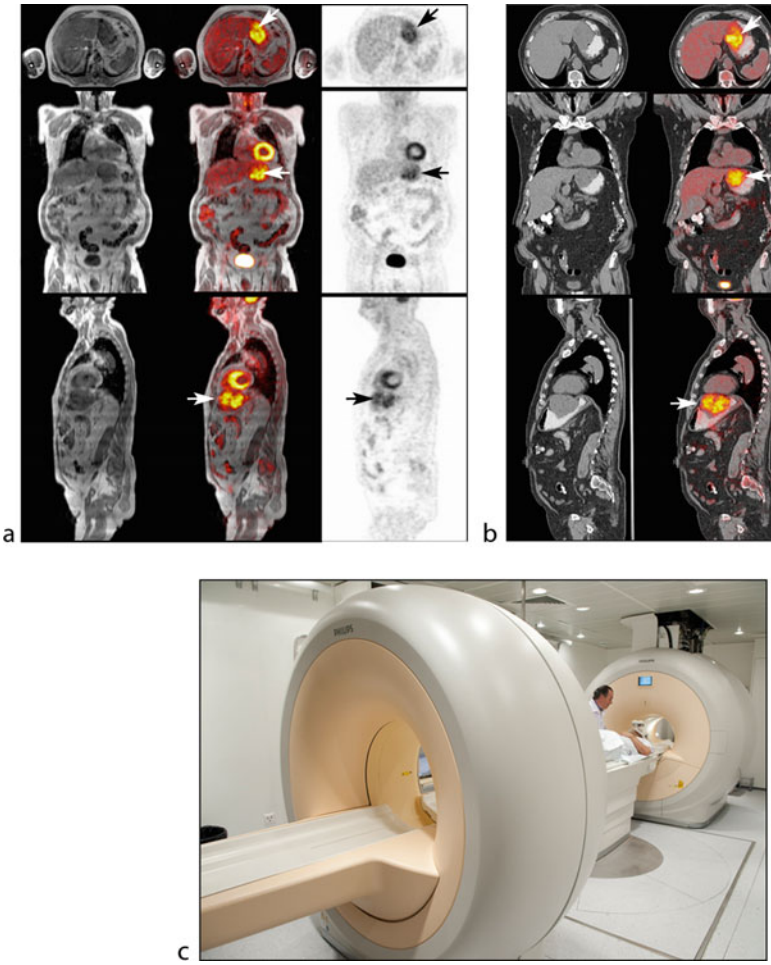


Fig. 7

Comparison of (a) PET/MR versus (b) PET/CT clinical imaging in a 77-year-old patient with gastrointestinal stromal tumor (*arrow*); Note the better soft-tissue contrast on MR that is acquired without any additional patient radiation dose. (c) Overview of the second clinical whole-body PET (*front*)/MR (*back*) scanner installed worldwide (courtesy of Prof. O. Ratib, M.D. Ph.D. [University Hospital of Geneva, Switzerland])

4.3 Dedicated PET Imaging Devices

4.3.1 PET Mammography (PEM)

There have been attempts to combine PET and mammography to reduce the false positive of mammography (Townsend 2008). An interesting device is the ClearPEM/Sonic device, which combines ultrasound/PET to benefit from including the elastographic tissue properties with

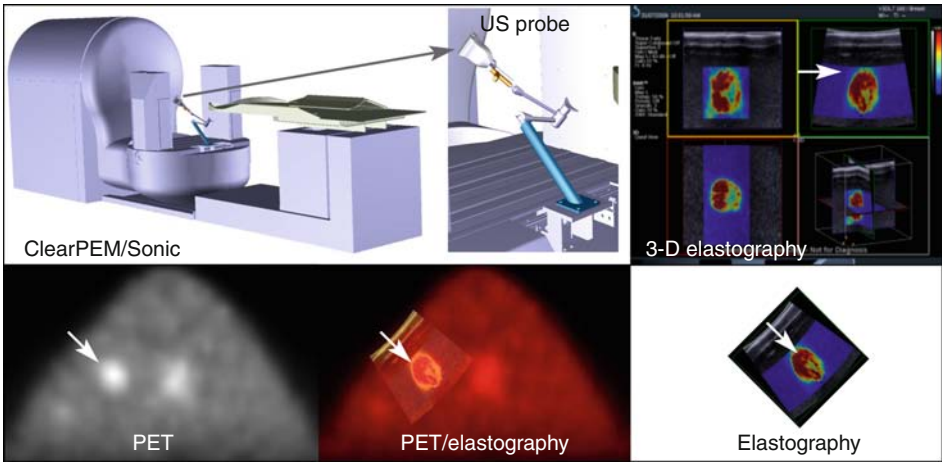


Fig. 8
ClearPEM/sonic scanner combining PET mammography with ultrasound-based elastography (*upper row*). PET and elastographic data (*lower row*) of a suspicious lesion (*arrow*)

the metabolic activity to differentiate benign from malignant breast lesions (Albuquerque et al. 2008) (*Fig. 8*). Clinical trials are under way to determine improvement in diagnostic performances over conventional mammography.

4.3.2 Endoscopic PET Imaging

The use of image-guided interventions (e.g., needle biopsy or placement of a needle for radiofrequency ablation) is growing and the introduction of molecular imaging information during these interventional procedures is desirable. This is the aim of the 4-year project entitled "EndoTOFPET-US" (http://endotofpet-us.web.cern.ch/) funded within the 7th European Framework Program and conducted by a large international and multidisciplinary consortium of 12 partners from six countries under the CERIMED guidance (http://cerimed.web.cern.ch/cerimed/php/default.php). This technically challenging project will start in 2011 and develop the first bimodal PET-US (PET and Ultrasound) endoscopic probe. It combines a miniaturized, fully digital, 200-ps time resolution TOF PET detector head coupled to a commercial endoscopic ultrasound system (*Fig. 9*). First clinical applications are envisioned in pancreatic cancer, after preclinical feasibility tests on pigs.

4.4 Other PET Detectors

Other detector technologies have been used or are currently in development, including multiwire proportional chamber, resistive plate chamber, and liquid xenon technologies. The discussion of their relative merits and possible future applications in medical imaging is outside the scope of this chapter, but related details can be found elsewhere (Lecomte 2009).

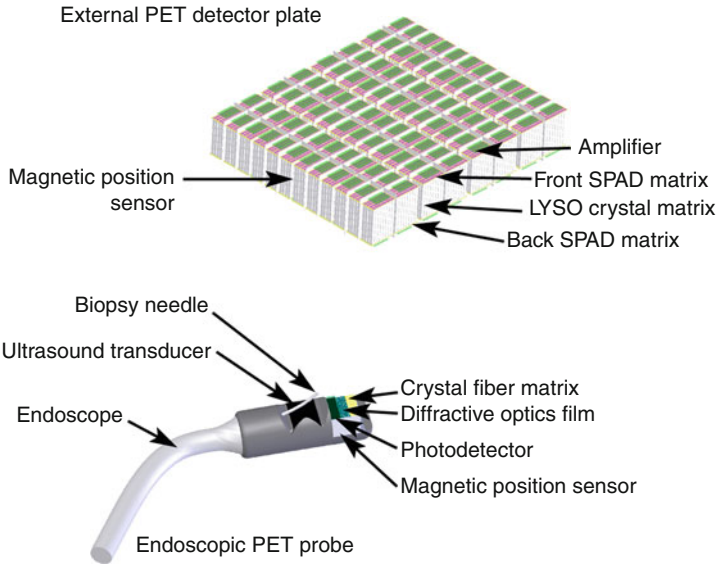


Fig. 9
Conceptual drawing of the time-of-flight (TOF) PET endoscopic probe for assisting biopsies with functional molecular imaging and anatomic landmarks from ultrasound (courtesy of B. Frisch [CERN, Geneva, Switzerland])

5 Radiopharmaceuticals

An overview of radiation-based medical imaging techniques would not be complete without describing, at least superficially, the wealth of existing and upcoming radiopharmaceuticals targeting different biological and molecular processes. Many textbooks and reviews are dedicated to this field (Ell and Gambhir 2004; Pysz et al. 2010). The number of SPECT and PET radiopharmaceuticals are growing every week and the online database MICAD (Molecular Imaging and Contrast Agent Database; http://micad.nih.gov) is developed at the United States National Institute of Health (NIH) and is currently containing 276 SPECT and 338 PET molecular imaging agents (NCBI 2010).

5.1 SPECT Radiopharmaceuticals

Examples of SPECT agents are given in *Table 5*. They are characterized by medium to long half-lives allowing observing biological processes over longer times than PET radiopharmaceuticals. They can also be produced locally at low cost from generators or delivered from remote production centers. A few of them allow the so-called theragnostic approach combining *thera*py with one radiotracer and dia*gnostic* with another one. Newer radiopharmaceuticals are in development, such as the expected ^{99m}Tc-labeled deoxyglucose, which will strengthen the role of SPECT for imaging hibernating myocardium (Zaman et al. 2010), or tracers for imaging apoptosis and angiogenesis.

Table 5

Example of common radiopharmaceuticals for SPECT

Name	Half-life	Molecular target or uptake process	Application
^{99m}Tc-phosphates (e.g., methylene diphosphate, hydroxymethylene diphosphate)	6 h	Hydroxyapatite crystals	Bone, musculoskeletal
^{153}Sm[a]-lexidronam (EDMTP)	1.9 days	Bone	Bone metastases
^{99m}Tc-bicisate (ECD)	6 h	Cerebral perfusion	Neurology
^{99m}Tc-exametazine (HMPAO)			
^{99m}Tc-sestamibi (Cardiolite®)	6 h	Perfusion	Cardiology
^{99m}Tc-tetrofosmin (Myoview®)			
^{99m}Tc-teboroxime (Cardiotec®)			
^{99m}Tc-disofenin (DISIDA)	6 h	Clearance	Hepatobiliary
^{99m}Tc-lidofenin (HIDA)			
^{99m}Tc-labeled red blood cells	6 h	Perfusion	Gastrointestinal bleeding
^{99m}Tc-sulfur colloids	6 h	Perfusion/clearance	Sentinel lymph node, splenosis
^{99m}Tc-mertiatide (MAG3)	6 h	Active tubular secretion and glomerular filtration	Renal
^{99m}Tc-succimer (DMSA)	6 h	Proximal tubular cells	Renal
^{123}I-, ^{131}I-, ^{131}I[a]-metaiodobenzylguanidine (MIBG)	13.3 h, 8.0 days	Specific, high-affinity Na uptake mechanism + energy-dependent type-I amine uptake mechanism	Pheochromocytoma
^{131}I-norcholesterol	8.0 days	Adrenocortical steroid hormones synthesis	Adrenocortical disorders
^{123}I-hippuran	13.3 h	Glomerular filtration + secretion	Renal
^{123}I, ^{131}I (lower activities)	13.3 h, 8.0 days	Sodium-iodide symporter	Thyroid
^{131}I[a] (higher activities)			
^{111}In-pentetreotide (Octreoscan®)	2.8 days	Somatostatin receptors	Endocrine tumors
^{111}In-pentetate (DTPA)	2.8 days	Perfusion	Cerebrospinal fluid
^{111}In/^{90}Y[a]-ibritumomab tiuxetan (Zevalin®)	2.8 days/64 h	CD20 leukocyte antigen	Non-Hodgkin lymphoma

[a]These radionuclides can be used for therapy

5.2 PET Radiopharmaceuticals

Many PET radiopharmaceuticals have been developed for research, but few are commonly used in clinical practice (Table 6). The workhorse used in oncologic PET is ^{18}F-fluorodeoxyglucose (FDG), a radioactive analogue of glucose whose uptake is increased in cancerous cells. In addition, other radiopharmaceuticals are currently used in a small, but increasing proportion, such as ^{18}F-NaF (for bone metabolism), ^{82}Rb (for myocardial perfusion, Fig. 11), or ^{68}Ga-based peptides (neuroendocrine tumors).

Table 6
Example of common radiopharmaceuticals for PET

Name[a]	Half-life	Molecular target or uptake process	Application
^{18}F-fluorodexyglucose (FDG)	110 min	Glucose transporter	Tumor glucose metabolism
^{18}F-NaF	110 min	Hydroxyapatite crystals	Bone turnover
^{18}F-fluoro-L-thymidine	110 min	Thymidine uptake in DNA/RNA synthesis	Tumor cell proliferation
^{18}F-galacto-arginine-glycine-aspartic acid (RGD)	110 min	$\alpha_V\beta_3$ integrin	Tumor angiogenesis
^{18}F-fluorocholine (FCH)	110 min	Choline transport	Tumor phospholipid synthesis
^{18}F-fluoro-L-ethyltyrosine (FET) (Fig. 10)	110 min	L-amino acid transporter	Brain tumor (glioma)
^{18}F-fluoromisonidazole (FMISO)	110 min	Hypoxia	Tumor hypoxia
^{11}C-methionine	20 min	Protein synthesis	Tumor protein synthesis
^{68}Ga-DOTA	68 min	Somatostatin receptor	Endocrine tumors
^{82}Rb	1.3 min	Na-K-ATPase pump in myocardium cells	Myocardial perfusion
^{11}C-acetate	20 min	Cell metabolism	Cardiac metabolism
^{11}C-palmitate	20 min	Fatty acid metabolism	Myocardial ischemia
^{11}C-metahydroxy-ephedrine	20 min	Adrenergic transmission	Heart failure
^{18}F-fluorodopamine (FDOPA)	110 min	Dopamine synthesis	Parkinson disease
^{11}C-raclopride	20 min	Dopamine postsynaptic receptor	Schizophrenia, addiction
^{11}C-Pittsburgh compound-B (PIB)	20 min	β-amyloid	Alzheimer disease
^{18}F-florbetapir, ^{18}F-florbetaben, ^{18}F-FDDNP	110 min 110 min		

[a] All radiopharmaceuticals are cyclotron-produced, except for ^{82}Rb and ^{68}Ga that are generator-produced

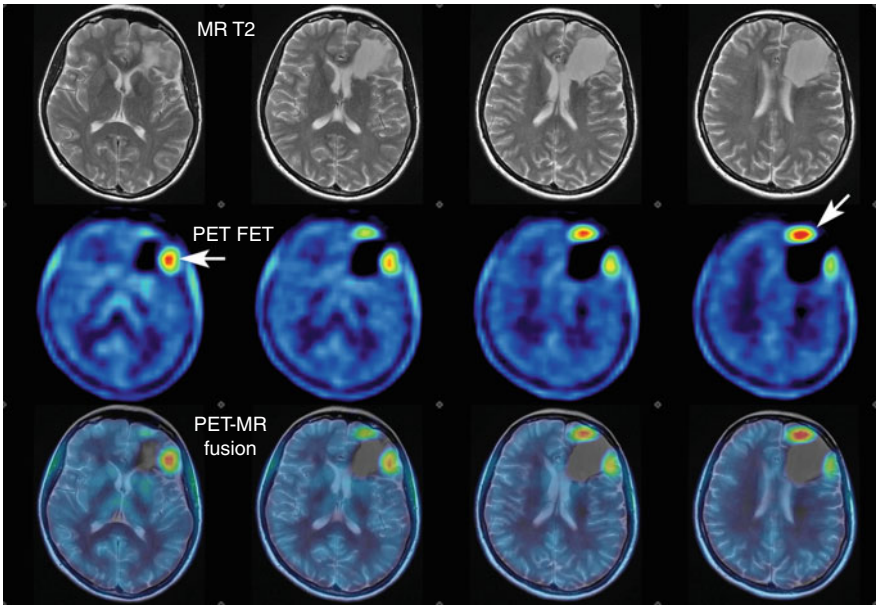


Fig. 10
Example of brain PET-MR fusion imaging with the amino acid analogue ^{18}F-fluoro-ethyl-tyrosine (FET). There is a tumoral relapse around the resection cavity (*arrows*) in a 37-year-old woman 5 years after removal of a glioma

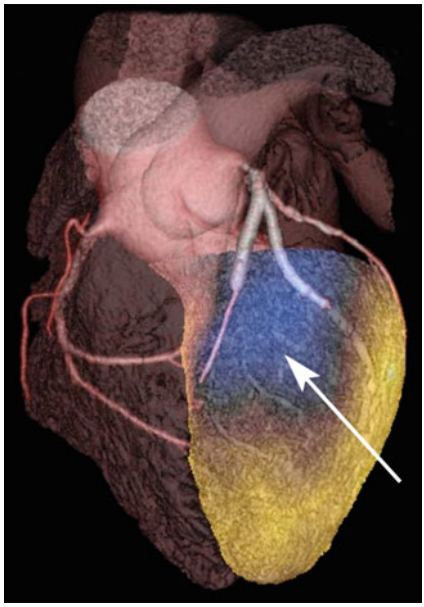


Fig. 11
Example of myocardial perfusion imaging with ^{82}Rb PET superimposed onto a coronary CT showing a region of diminished perfusion (*arrow*) in an ischemic region subtended by two cardiac stents

6 Preclinical Imaging

Preclinical imaging in small-animal is part of the development of new molecular radiopharmaceuticals, as well as drugs tested in animals before their application to man (translational research) (Pysz et al. 2010). Specific, dedicated scanners have been developed or are in development for small-animal imaging (SPECT, PET) as well as their hybrid imaging combination (PET/CT, PET/MR, SPECT/CT, SPECT/MR).

Many adaptations have been made to improve the resolution needed in rodents. For instance, in microPET, the spatial resolution should be 0.7 mm to provide whole-body mouse imaging at the same volume resolution as the current human machine. This is very challenging requiring long crystals with depth-of-interaction measurements. Nowadays, microPET resolutions are down to about 1.2 mm allowing to image small structures like mice hearts (*Fig. 12*).

On the other hand, microSPECT systems using pinhole imaging have currently reached submillimetric image resolutions of about 0.35 mm (0.04 μL) in the mouse, which can even

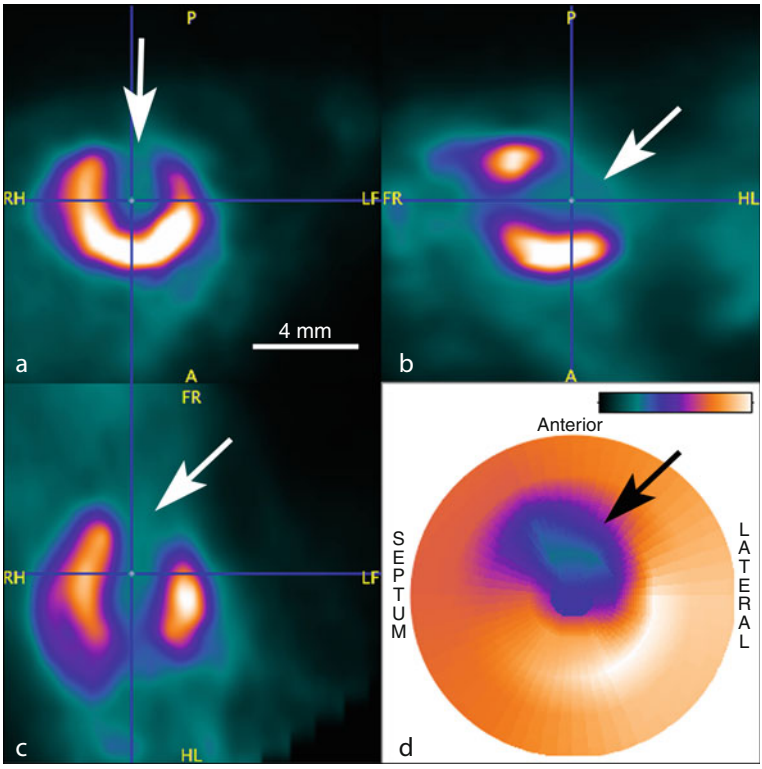


Fig. 12

Example of microPET imaging of myocardial glucose metabolism using ^{18}F-fluorodeoxyglucose in a mouse heart (6 mm long) with ligatured left anterior descending coronary artery. (a) Short-axis, (b) vertical-axis, and (c) horizontal long-axis views showing an infarct (*arrow*). (d) Polar map (courtesy of CIBM [CHUV–EPFL Lausanne, Switzerland])

be improved with strategies such as measuring the depth of interaction within the crystal. Thus SPECT imaging with pinhole has no intrinsic limitation such as in PET due to the positron energy, and therefore has the potential to obtain ultrahigh resolution down to subcompartments of mouse organs (Beekman and Have 2006).

7 Conclusions

Improvements in sensitivity and spatial resolution are expected thanks to the development of new detectors, in both SPECT and PET radiation-based imaging modalities. Multimodality imaging available today in the clinics (SPECT/CT and PET/CT) or upcoming (PET/MR, SPECT/MR, PET/US) increase lesion localization and specificity. New scanners dedicated to specific organs (breast, brain) or applications (biopsy or interventions) should complete the armamentarium of clinical applications in nuclear medicine and molecular imaging. Combined with the development of new radiopharmaceuticals, these techniques offer many opportunities to provide characterization of the biology and function of tissues noninvasively with the aim of personalizing therapy.

References

Albuquerque E, Almeida FG, Almeida P, Auffray E, Barbosa J, Bastos AL, Bexiga V, Bugalho R, Carmona S, Carrico B, Ferreira CS, Ferreira NC, Ferreira M, Frade M, Godinho J, Goncalves F, Guerreiro C, Lecoq P, Leong C, Lousa P, Machado P, Martins MV, Matela N, Moura R, Neves P, Oliveira N, Ortigao C, Piedade F, Pinheiro JF, Relvas P, Rivetti A, Rodrigues P, Rolo I, Sampaio J, Santos AI, Santos J, Silva MM, Tavernier S, Teixeira IC, Teixeira JP, Silva R, Silva JC, Trindade A, Varela J (2008) An overview of the clear-PEM breast imaging scanner. Nuclear Science Symposium Conference Record, 2008. NSS '08, Dresden, IEEE, Piscataway, 19–25 December 2008

Beekman F, Have F (2006) The pinhole: gateway to ultra-high-resolution three-dimensional radionuclide imaging. Eur J Nucl Med Molec Imaging 34(2):151–161

Blodgett TM, Meltzer CC, Townsend DW (2007) PET/CT: form and function. Radiology 242(2):360–385

Buck AK, Herrmann K, Stargardt T, Dechow T, Krause BJ, Schreyogg J (2010) Economic evaluation of PET and PET/CT in oncology: evidence and methodologic approaches. J Nucl Med 51(3):401–412

Bybel B, Brunken RC, DiFilippo FP, Neumann DR, Wu G, Cerqueira MD (2008) SPECT/CT imaging: clinical utility of an emerging technology. Radiographics 28(4):1097–1113

Cherry SR (2009) Multimodality Imaging: Beyond PET/CT and SPECT/CT. Semin Nucl Med 39(5):348–353

Conti M (2009) State of the art and challenges of time-of-flight PET. Phys Med 25(1):1–11

Delso G, Ziegler S (2009) PET/MRI system design. Eur J Nucl Med Mol Imaging 36(Suppl 1): S86–S92

Ell PJ, Gambhir S (2004) Nuclear medicine in clincal diagnosis and treatment. Elsevier, Edinburgh

Frangioni JV (2008) New technologies for human cancer imaging. J Clin Oncol 26(24):4012–4021

Hillner BE, Siegel BA, Liu D, Shields AF, Gareen IF, Hanna L, Stine SH, Coleman RE (2008) Impact of positron emission tomography/computed tomography and positron emission tomography (PET) alone on expected management of patients with cancer: initial results from the National Oncologic PET Registry. J Clin Oncol 26(13):2155–2161

Jansen FP, Vanderheyden JL (2007) The future of SPECT in a time of PET. Nucl Med Biol 34(7):733–735

Karp JS, Surti S, Daube-Witherspoon ME, Muehllehner G (2008) Benefit of time-of-flight in PET: experimental and clinical results. J Nucl Med 49(3):462–470

Kohara R, Shirahata T, Nakazawa T, Miyazaki O, Kabuki S, Kurosawa S, Miuchi K, Kubo H, Tanimori T, Nakahara T, Kunieda E, Kubo A, Fujii H (2008) Advanced Compton camera system for nuclear medicine: Prototype system study. In: Nuclear science symposium conference record, 2008, NSS '08, Dresden, IEEE, Piscataway, 19–25 December 2008

Lecomte R (2009) Novel detector technology for clinical PET. Eur J Nucl Med Mol Imaging 36(Suppl 1):S69–S85

Lee AI, Zuckerman DS, Van den Abbeele AD, Aquino SL, Crowley D, Toomey C, Lacasce AS, Feng Y, Neuberg DS, Hochberg EP (2010) Surveillance imaging of Hodgkin lymphoma patients in first remission: a clinical and economic analysis. Cancer 116(16):3835–3842

Lewellen TK (2008) Recent developments in PET detector technology. Phys Med Biol 53(17): R287–R317

Mansueto M, Grimaldi A, Mangili G, Picchio M, Giovacchini G, Vigano R, Messa C, Fazio F (2009) Positron emission tomography/computed tomography introduction in the clinical management of patients with suspected recurrence of ovarian cancer: a cost-effectiveness analysis. Eur J Cancer Care (Engl) 18(6):612–619

Mawlawi O, Townsend DW (2009) Multimodality imaging: an update on PET/CT technology. Eur J Nucl Med Mol Imaging 36(Suppl 1):S15–S29

Meltzer CC, Leal JP, Mayberg HS, Wagner HN Jr., Frost JJ (1990) Correction of PET data for partial volume effects in human cerebral cortex by MR imaging. J Comput Assist Tomogr 14(4):561–570

NCBI (2010) Molecular imaging and contrast agent database. National Library of Medicine (US). http://micad.nih.gov. Accessed 1 Aug 2010

Papathanassiou D, Liehn J (2008) The growing development of multimodality imaging in oncology. Crit Rev Oncol/Hematol 68(1):60–65

Pichler BJ, Wehrl HF, Judenhofer MS (2008) Latest advances in molecular imaging instrumentation. J Nucl Med 49(Suppl 2):5S–23S

Pysz MA, Gambhir SS, Willmann JK (2010) Molecular imaging: current status and emerging strategies. Clin Radiol 65(7):500–516

Seo Y, Mari C, Hasegawa B (2008) Technological development and advances in single-photon emission computed tomography/computed tomography. Semin Nucl Med 38(3):177–198

Soret M, Bacharach SL, Buvat I (2007) Partial-volume effect in PET tumor imaging. J Nucl Med 48(6):932–945

Tornai MP, Bowsher JE, Jaszczak RJ, Pieper BC, Greer KL, Hardenbergh PH, Coleman RE (2003) Mammotomography with pinhole incomplete circular orbit SPECT. J Nucl Med 44(4):583–593

Townsend DW (2008) Multimodality imaging of structure and function. Phys Med Biol 53(4): R1–R39

Zaman MU, Hashmi I, Fatima N (2010) Recent developments and future prospects of SPECT myocardial perfusion imaging. Ann Nucl Med 24(8):565–569

Further Reading

Cherry SR, Sorenson S, Phelps M (2003) Physics in nuclear medicine. Saunders, Philadelphia

Ell PJ, Gambhir S (2004) Nuclear medicine in clincal diagnosis and treatment. Elsevier, Edinburgh

Tavernier S, Gektin A, Grinyov B, Moses WW (2006) Radiation detectors for medical applications. Springer, Dordrecht

36 CT Imaging: Basics and New Trends

Françoise Peyrin[1] · *Klaus Engelke*[2]
[1]Inserm U1044; UMR CNRS 5220; INSA Lyon; Université de Lyon, Villeurbanne, France
[2]University of Erlangen, Erlangen, Germany

1 Introduction 884

2 Principles of X-ray CT 885
2.1 Physics of X-ray CT 885
2.2 Data Acquisition 887
2.3 Image Reconstruction 889
2.4 Image Quality and Artifacts 890

3 Historical and Current Concepts of CT Technology 895
3.1 Translation–Rotation CT: First and Second Generations 895
3.2 Fan-Beam CT: Third and Fourth Generations 896
3.3 Spiral or Helical CT 897
3.4 Multi-slice CT (MSCT) 898

4 New Developments in CT Technology and Applications 900
4.1 Perfusion CT 900
4.2 Quantitative CT (QCT) 902
4.3 Spectral CT 903

5 Radiation Exposure 905

6 From Clinical CT to Nano-CT 907
6.1 Micro-CT Using X-ray Tubes 907
6.2 Synchrotron Radiation Micro-CT 909
6.3 Nano-CT 911

7 Conclusion 911

Cross-References 912

References 912

C. Grupen, I. Buvat (eds.), *Handbook of Particle Detection and Imaging*,
DOI 10.1007/978-3-642-13271-1_36, © Springer-Verlag Berlin Heidelberg 2012

Abstract: This chapter presents the principle of X-ray CT and its evolution during the last 40 years. The first section describes the physical basis of X-ray CT, tomographic image reconstruction algorithms, and the source of artifacts in X-ray CT images. The second section is devoted to the evolution of CT technology from the first translation–rotation systems to multi-slice spiral CTs currently used today. The next section addresses specific developments of CT technology and applications, like perfusion CT, quantitative CT, and spectral CT. The fourth section introduces the problem of radiation exposure delivered to the patient and its evaluation. Finally the last section addresses the development in micro- and even nano-CT which is a rapidly evolving area in preclinical imaging and biology.

1 Introduction

In the 1970s, the introduction of X-ray Computed Tomography (CT) was a breakthrough in medicine since it was the first technique providing unbiased virtual sections of the human anatomy. The name *tomography* comes from the greek word, *tomos* meaning section or cut. X-ray CT relies on the combination of data acquisition and data processing that became possible only with the availability of the first computers. X-ray CT has considerably evolved since the first systems and is still the focus of active research. Today, despite of the development of alternative modalities providing slices, such as Magnetic Resonance Imaging (MRI), Positron Emission Tomography (PET) or Single-Photon Emission Tomography (SPECT), certainly X-ray CT remains the most widely spread imaging technique in medical practice and is used for diagnosis, follow up, surgery preparation, radiation therapy planning, and other applications. With the technological progresses of detectors, X-ray sources, acquisition time, image reconstruction algorithms, advanced CT image analysis, and reduction of radiation exposure, it is likely that clinical applications of this technique will further expand in future.

The discovery of X-rays by Wilhelm Roentgen in 1895 opened the way for radiological imaging in medicine. Since X-rays are able to penetrate through the human body, the first radiographic images allowed us to "see through" the patient. However, in a standard radiograph, all the structures along the X-ray beam are superimposed and thus depth information is lost. Without a good knowledge of anatomy, the radiologist cannot tell whether a structure is in front or behind another. Therefore, during a radiological examination, front and profiles views are often prescribed for better interpretation. This idea of combining information from different angles of views is the basic principle of CT.

X-ray CT removes the ambiguity of standard radiography by providing sections in the depth direction perpendicular to the radiographic image. X-ray CT has two distinct stages: 1) data acquisition during which measurements obtained when rotating the X-ray source around the patient are collected, and 2) image reconstruction, i.e., processing of all measured data in order to generate the digital image of the section. It has to be noticed that the measurements in CT do not directly provide the image but all necessary information to obtain it by solving an inverse problem. Thus, X-ray CT combines X-ray and computer technologies together with theoretical aspects in physics and mathematics.

The first prototype of X-ray CT was constructed by Godfrey N. Hounsfield (Hounsfield 1973), from EMI Laboratories, who proposed both a system for data acquisition and an algorithm for image reconstruction. In 1979, Hounsfield received the Nobel Prize for this major

discovery together with the physicist Allan Cormack who had independently contributed to the mathematical basis of CT (Cormack 1980). In addition to these pioneers, other researchers independently had similar ideas in other areas such as radio astronomy, electron microscopy, or positron emission (Herman 1980). The foundations on image reconstruction are due to the Austrian mathematician Johann Radon (Radon 1917) although in a purely theoretical context.

With the evolution of technology and new ideas brought by researchers, different generations of scanners succeeded one another. In the 1990s, the introduction of spiral or helical scanners was a significant improvement from single-slice toward three-dimensional imaging. Decrease of acquisition times was also a major advance to enable good-quality cardiac imaging. Today, the use of planar and X-ray-sensitive detectors or dual-X-ray sources still opens new perspectives. X-ray CT has evolved from a single-slice 2D technique to a multi-slice 3D imaging modality. With appropriate computer graphic software, it is thus not only possible to display virtually any section in any direction in the volume but to obtain 3D rendering on selected anatomical structures and to observe them under any incidence.

In this chapter, Sect. 2 introduces the principle of X-ray CT by presenting the X-ray physics, some basics in image reconstruction, and a discussion on image artifacts. Section 3 is devoted to the description of historical and standard concepts of CT technology. Section 4 discusses new developments and applications. Section 5 addresses the problem of X-ray exposure and dose. Finally, Sect. 6 introduces the developments in micro- and even nano-CT, which are rapidly evolving areas for applications in small-animal imaging and biology.

2 Principles of X-ray CT

The principle of image formation in X-ray CT illustrated in *Fig. 1* relies on two well-defined steps of data acquisition and data processing. The first step consists in recording the so-called projections, which are attenuation measures along a number of X-ray paths in the section of interest. The second step consists in processing all these measurements to reconstruct the digital image of the section. The theory of X-ray CT requires modeling data acquisition by taking into account the physics of X-rays and solving the inverse problem to reconstruct the digital image.

2.1 Physics of X-ray CT

X-rays are electromagnetic waves that can either be characterized by their wavelength or their energy. The range of energies used in medical imaging is between 20 and 100 keV. In this range, X-ray attenuation is the result of three main X-ray–matter interaction phenomena: photoelectric absorption, in which the X-ray photon transfers its energy to the absorbing material, Rayleigh (elastic) scattering, and Compton (inelastic) scattering in which the X-ray photon is deviated. Macroscopically, X-ray attenuation is modeled by the Beer–Lambert law. Let a monochromatic X-ray beam of given energy E travel through a homogeneous material of thickness L. If I_0 is the intensity of the incident X-ray radiation, its intensity I after passing through the material can be expressed as

$$I = I_0 \exp(-\mu\, L), \tag{1}$$

where μ is the energy-dependent linear attenuation coefficient of the material, expressed in reciprocal length, typically per centimeter (cm^{-1}). This relationship shows the exponential

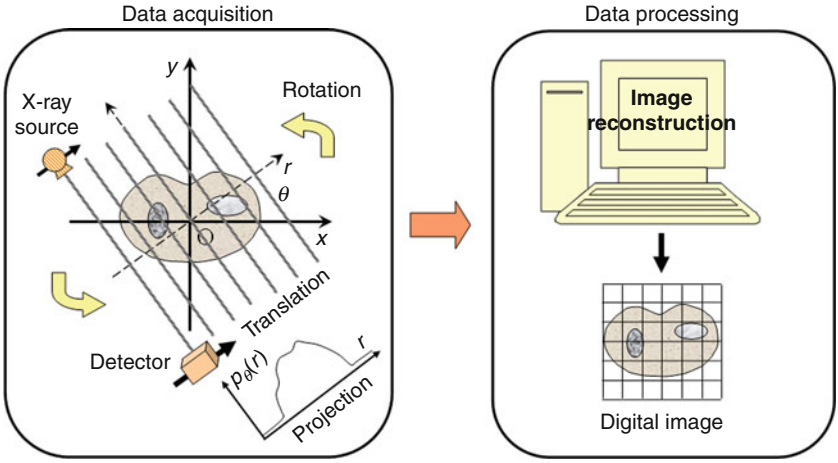


Fig. 1

The two steps of data acquisition and data processing in the first generation of X-ray CT. *Left*: During data acquisition, the attenuations of a pencil X-ray beam on parallel X-ray paths are measured on a detector, providing one projection. The X-ray source is then sequentially rotated and new measurements are collected. *Right*: When a complete rotation angle of 180° has been done, the data are processed through an image reconstruction algorithm providing a digital image

decay of the intensity as a function of thickness. The μ parameter describes the fraction of the beam absorbed or scattered per unit thickness. It includes the contributions of photoelectric absorption and Compton and Rayleigh scattering effects. When it is high (respectively low), the material is more opaque (respectively more transparent) to X-rays. μ depends on the composition and density of the materials. Typically, the linear attenuation coefficient decreases with energy and is lower for low-atomic-number materials (air, water, soft tissue) than for high-atomic-number ones (bone, metal). The mass attenuation coefficient expressed in square centimeters per gram (cm^2/g) is defined as μ/ρ, ρ being the material density. μ/ρ values for the various elements of the periodic table as a function of energy are available in different databases (Hubbell 2006) (NIST Standard Reference Database). The mass attenuation coefficient of a mixture or compound can be expressed as the linear combination of the mass attenuation coefficients of each element $(\mu/\rho)_i$ weighted by its weight fraction or concentration w_i as

$$\frac{\mu}{\rho} = \sum_i w_i \left(\frac{\mu}{\rho}\right)_i . \tag{2}$$

When X-rays travel through the human body, they encounter various materials. The Beer–Lambert law can be generalized to model the absorption in an inhomogeneous material. If $\mu(x, y)$ denotes the linear attenuation coefficient for energy E at each point (x, y), the relation between the incident and transmitted intensity becomes

$$I = I_0 \exp\left(-\int_D \mu(x, y)\, \mathrm{d}s\right), \tag{3}$$

where D is the X-ray path, modeled as the straight line joining the X-ray source to the detector.

Thus, in theory, the image provided in X-ray CT is a map of the linear attenuation coefficient in the section plane. In practice, X-ray CT scanners deliver the so-called HU for Hounsfield Unit, corresponding to the linear attenuation normalized by that of water:

$$HU(x,y) = 1{,}000\ \frac{(\mu(x,y) - \mu_{\text{water}}(x,y))}{\mu_{\text{water}}(x,y)}. \tag{4}$$

According to this definition, HU of water is zero, fat and lungs have negative HU values, bone or calcified tissues have high values (usually above 1,000), and other tissues fall between these extremes. The HU values are typically 12-bit coded providing 4,096 integer values ranging from −1,024 HU to 3,071 HU. Since the human eye is not sensitive enough to distinguish 4,096 gray levels simultaneously, the image is generally seen through a "window," which can be chosen interactively by the radiologist and determines the range of gray levels to be displayed. It is defined by its center and width, the center corresponding to the HU value of the structure of interest, and the width determining the contrast in the image. Typical couples of (center, width) for lung and bone are for instance (−600, 1,700) and (1,000, 2,500).

2.2 Data Acquisition

Data acquisition consists of recording the transmitted intensity along different X-ray paths as specified in ❯ Sect. 3 which, for a given energy can be equivalently rewritten:

$$\int_D \mu(x,y)\ \mathrm{d}s = \ln\left(\frac{I_0}{I}\right). \tag{5}$$

For a given X-ray path D, since the incident intensity I_0 is assumed to be known, and the transmitted intensity I is measured, the measurement gives access to the integral of the searched function on the straight line D. This line integral is known as a projection value.

In theory, the reconstruction of the image $\mu(x,y)$ requires to measure such attenuated intensities along all possible straight lines D of the planar section. In practice, the dataset is finite and the way it is parameterized defines the geometry of the acquisition system. The simplest geometry corresponds to the one initially proposed by Hounsfield known as parallel geometry, which is tailored to the first scanner generation also called translation–rotation scanners (see ❯ Sect. 3). The X-ray source and detectors are sequentially rotated and for each given angle of rotation, all X-ray paths are parallel to each other (❯ *Fig. 1*). This geometry is also the simplest to study the principle of image reconstruction.

By expressing the equation of the straight line defined by angle θ and detector position r, the parallel projection $p_\theta(r)$ (❯ *Fig. 1*) can be expressed as

$$p_\theta(r) = \int_{-\infty}^{+\infty} \mu(r\cos\theta - s\sin\theta,\ \ r\sin\theta + s\cos\theta)\ \mathrm{d}s. \tag{6}$$

The set of projections for which θ varies between 0 and π constitutes the Radon transform of the image, also called a sinogram in CT. This terminology comes from the fact that a point (x,y) in the original image is associated with a sinusoidal line in Radon space. ❯ *Figure 2* illustrates three 256 × 256 pixel digital images and their corresponding sinograms. The sinogram is displayed as

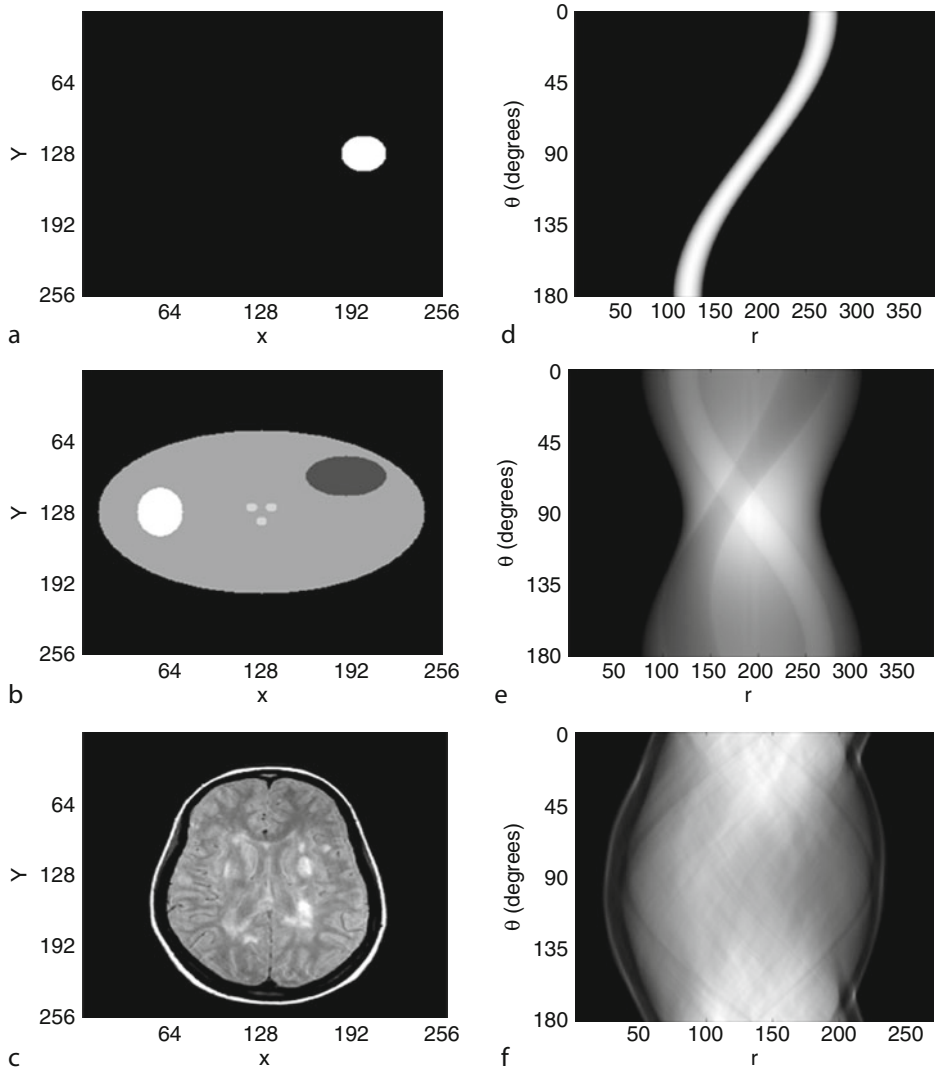


Fig. 2

Three 256 × 256 images (a–c) and their corresponding Radon transform or sinogram (d–f): (a, d) small circle, (b, e) ellipse image, (c, f) CT section. The Radon transform of the small circle illustrated the sinusoidal shape from which the sinogram takes its name. The Radon transform of the ellipse image can be calculated as the sum of the Radon transform of each ellipse

a gray-level image where black corresponds to the zero value and white to the highest value. The vertical axis corresponds to the rotation angle between 0 and π and the horizontal axis to the detector position. It has to be noted that the Radon transform is linear and then the Radon transform of a sum of images is the sum of each Radon transform.

2.3 Image Reconstruction

Image reconstruction is the estimation of the image from its sinogram. From a mathematical point of view, this problem is equivalent to invert the Radon transform of the image, i.e., to estimate the image from its line integrals over a large number of lines in the plane (Natterer 1986).

Historically, long before the actual realization of CT scanners, J. Radon had solved this problem in a general theoretical framework for n-dimensional functions (Radon 1917). The Radon inversion is directly applicable in X-ray CT for the reconstruction of images in dimension 2. However, other methods were developed and are preferred because of speed or simplicity. There are generally two broad classes of methods, Fourier and series expansion methods (Herman 1980; Kak and Slaney 1988; Grangeat 2002). These two approaches correspond respectively to a continuous and a discrete formulation of the problem. Fourier methods are based on analytic inversion formulas expressing the image as a function of its projections. In algebraic methods, the image is obtained through the resolution of a large linear system, which, given the size of the problem, is generally solved iteratively. The historical method proposed by Hounsfield was an algebraic method called Algebraic Reconstruction Technique (ART). Since more details can be found in ❯ Chap. 39, "Image Reconstruction", we only recall some basic notions useful in the scope of this chapter.

Let $f(x,\ y)$ be the image to be reconstructed. The fundamental result in the field of image reconstruction is the Fourier slice theorem. This theorem states that the 1D Fourier transform of a parallel projection of angle θ is a slice of the 2D Fourier transform of the image along the straight line in direction θ. Thus, when the angle θ varies between 0 and π, the entire 2D Fourier transform of the image is filled. This theorem is very important since it reveals why acquired CT data are necessary and sufficient to reconstruct the image. This theorem implies a reconstruction method called direct inversion. In practice this method requires 2D interpolation between a polar and Cartesian grid in the Fourier domain, a step which is very sensitive to errors.

The most popular reconstruction algorithm in X-ray CT is Filtered Backprojection (FBP). It relies on an exact analytical inversion formula that can easily be derived from the Fourier slice theorem (Ramachandran and Lakshminarayanan 1971; Shepp and Logan 1974):

$$f(x,y) = \frac{1}{2}\int_0^{\pi} \tilde{p}_\theta(x\cos\theta + y\sin\theta)\,\mathrm{d}\theta, \tag{7}$$

$$\tilde{p}_\theta(r) = (p_\theta * h)(r), \tag{8}$$

where $*$ denotes the convolution operator and h is the reconstruction filter. ❯ Equation 7 is called backprojection of the filtered projections $\tilde{p}_\theta(r)$. If the bandwidth of the projection is within the interval $[-W, W]$, then for the Fourier transform of the reconstruction filter the following relation must be valid:

$$F_1 h(R) = |R| \quad \text{for} \quad R \in [-W, W], \tag{9}$$

where F_1 denotes the 1D Fourier Transform operator.

From a geometrical point of view, the backprojection of one projection "spreads" the values of the projection along the projection direction. This is illustrated in ❯ *Fig. 3* showing the backprojections of the raw and filtered projections for $\theta = 0°$. In the FBP algorithm this operation is repeated for all projection angles and the reconstructed image is the sum of the backprojections of all the filtered projections. The first line of ❯ *Fig. 4* shows the result of FBP

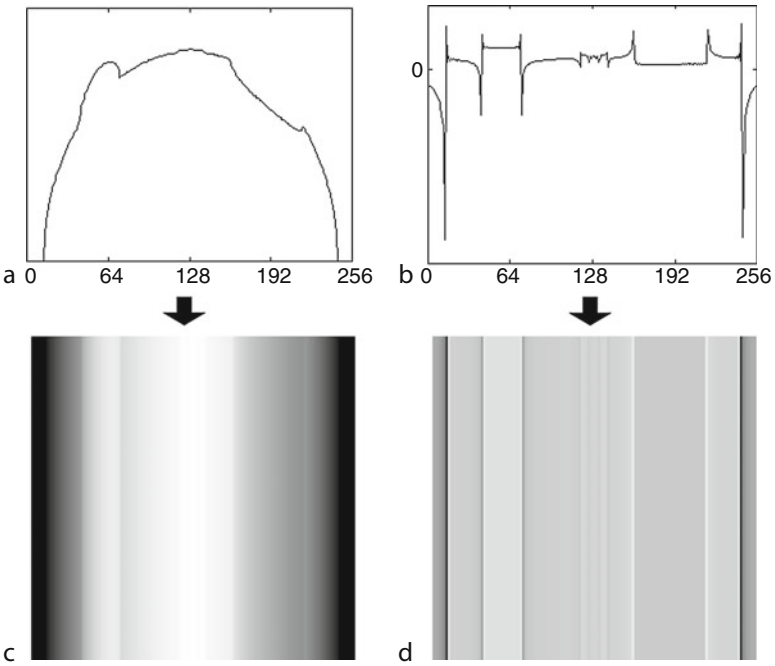


Fig. 3
Illustration of the backprojection of one single projection: (a) projection for angle $\theta = 0°$ and (c) its backprojection, (b) filtered projection and (d) its backprojection. The backprojection is displayed as a gray-level image

for respectively 2, 4, and all projections, the latter providing the final reconstruction, which is a good estimate of the original image. The second line of *Fig. 4* shows the backprojection effect directly applied to the projection without filtering. In this case, a blurred version of the image is obtained, a result that can be demonstrated mathematically. Thus, filtering is mandatory for reconstruction of the original image. Filtering introduces negative values in the projections which are compensated once all filtered backprojections are combined. Filtering can be implemented as a convolution in the spatial or a multiplication in the frequency domain. In this later case, the projection is first zero padded to avoid aliasing.

FBP is the method currently used in clinical scanners, due to its simplicity and speed. It is a sequential algorithm, only involving one-dimensional filtering operations, and the contribution of each projection can be computed before the end of data acquisition. It has been presented in the simple case of 2D parallel geometry, but the method can also be extended to fan- and cone-beam geometries found in modern of X-ray CT scanners.

2.4 Image Quality and Artifacts

As will also be discussed in Chap. 43, "Evaluation and Image Quality in Radiation-Based Medical Imaging", image quality in medical imaging is generally assessed in terms of spatial resolution (size of the smallest observable detail), contrast (smaller observable difference

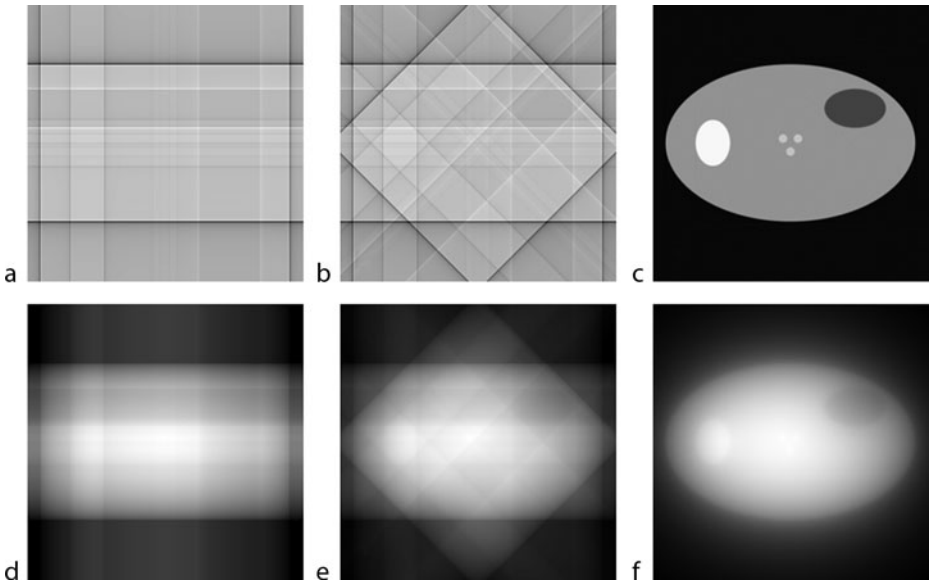


Fig. 4

(a–c): Filtered backprojection after summing, 2, 4 and all projections. (d–f) Backprojection without filtering after summing, 2, 4, and all projections. Image (c) is the reconstructed image while image (f) is a blurred version of the original image

of densities) and signal-to-noise ratio (SNR). In X-ray CT, image quality depends on a number of parameters, either related to data acquisition or reconstruction protocols or to the CT system itself. In this section we review the major parameters determining image quality and typical artifacts that can be found in tomographic images.

The FBP integral (Eq. 7) is an exact inversion formula for a continuous image assuming a continuous Radon transform defined for $(\theta, r) \in [0, \pi[\times]-\infty, \infty[$. In practice, the Radon transform is sampled since only a finite number of projection angles and a finite number of detector values are available. The choices of angular and spatial sampling have an important impact on image quality.

Spatial resolution: Spatial sampling on the detectors determines the spatial resolution of the reconstructed image. According to the Shannon–Nyquist theorem, if the sampling distance of the projection is Δr, the maximal available spatial frequency is $1/2\Delta r$. Using the Fourier slice theorem, this means that the support of the 2D Fourier transform of the image is a disk of radius $1/2\Delta r$. Thus, the sampling rate of the detector determines the sampling rate of the reconstructed image. If the original object effectively contains details smaller than Δr, the reconstructed image will be corrupted by aliasing artifacts, which result in smoothing and corruption of details. Thus, the spatial resolution of the projections must be chosen according to the desired spatial resolution in the image.

Partial volume effect: Another phenomenon closely related is the partial volume effect. This effect is related to the fact that the physical detector element which has a given area and thickness integrates the intensity measurement within a finite volume element. In direct imaging, such a problem causes smoothing of details and edges in the recorded image. However, in X-ray CT, an additional effect occurs, which is related to the logarithmic relation between measured

intensities and attenuation (Eq. 5). Since the logarithm is a nonlinear function, the logarithm of the averaged intensity values is not equal to the average of the logarithm of different intensity values. For example, a sharp edge in the object will introduce a nonlinearity visible in the reconstructed image as smoothing, underestimation of the CT values and possible streak lines. This effect can either originate from structures within the transverse section but also from structures slightly below or above the section due to the vertical extent of the detector. Thus, the slice thickness is also a parameter to take into account when interpreting the image.

Angular sampling: Angular sampling has to be appropriate to obtain an adequate reconstruction. Since in the FBP algorithm the integral in Eq. 7 is approximated, as many projections as possible should be taken in the interval $[0, \pi]$. Actually, the minimum number of projection angles M depends of the spatial resolution, i.e., may be estimated as a function of the number of detector resolution elements in one projection NP by the following formula (Kak and Slaney 1988):

$$M \approx \frac{\pi}{2} NP. \tag{10}$$

When the number of projections M is too small, typical streak artifacts along the projection rays are apparent. *Figure 5b* shows streak rays appearing in the reconstruction of the ellipse image in *Figure 5a* from only 32 projections. *Figure 5c* shows the reconstructed image

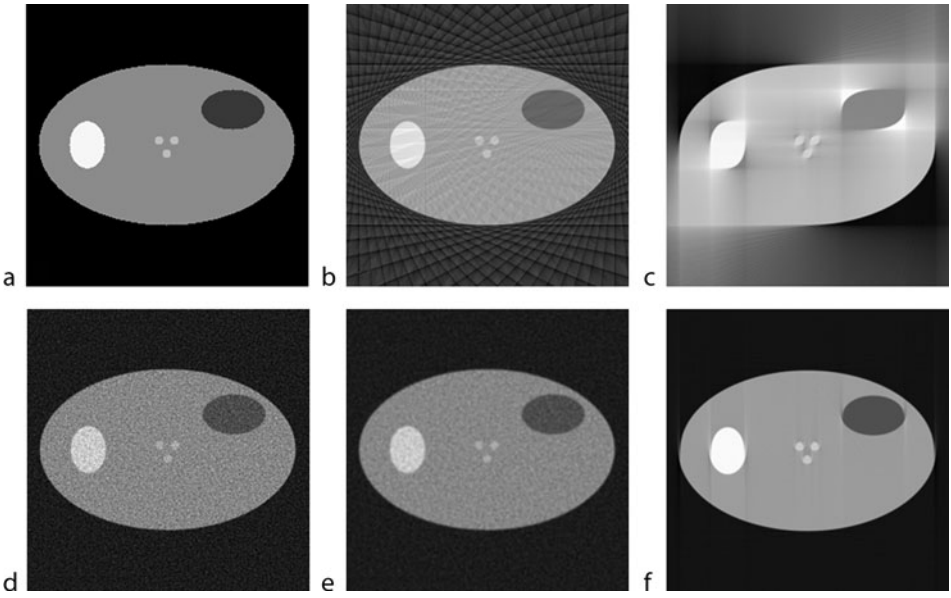


Fig. 5

Illustration of image quality and artifacts: (a) Original 256 × 256 ellipse image, (b) FBP reconstruction from 32 projections on [0, 180°], (c) FBP reconstruction from 192 projections on [0, 90°], (d) reconstruction from noisy projections with the ramp filter (standard deviation: 12.7), (e) reconstruction from noisy projections with the Cosine filter (standard deviation: 4.8), (f) reconstruction with off-centered rotation axis

from an angular coverage of only 90° instead of 180° showing the blurring introduced by the missing information.

Reconstruction filter: The reconstruction filter is another important parameter in FBP. The ideal reconstruction filter is defined by Eq. 9 and called the ramp or Ram–Lak filter in abbreviation of the authors who introduced it (Ramachandran and Lakshminarayanan 1971). However, this filter amplifies high frequencies and thus can also amplify noise in the filtered projections and therefore in the reconstructed image. That is why variants of this filter obtained by multiplication with a smoothing window in the frequency domain are often used. This operation is equivalent to a low-pass filtering of the reconstructed image and thus improves signal-to-noise ratio at the cost of a loss in spatial resolution. The Shepp–Logan filter uses a sinc window but other filters such as Hamming or Hanning have been used as well (Shepp and Logan 1974). The reconstruction filter is a parameter that CT operators can choose. *Figure 5d* and *e*, respectively, display the reconstruction of an ellipse image with data corrupted by Gaussian noise with the Ram–Lak and the Cosine filters. Although the differences are hardly visually apparent, the SNR in the two images differs by a factor of 2.7 (*Fig. 5e*). The improvement of the SNR has been obtained at the expense of the edges that appear more blurred. Thus, in choosing different reconstruction filters the radiologist has to choose a noisier textured image with sharp details versus a smoother image.

Acquisition geometry: Finally, image reconstruction also requires the precise knowledge of the acquisition geometry. *Figure 5f* illustrates the blurring introduced by a shift of one pixel of the center of rotation. Thus, the geometry of clinical X-ray CT scanners is accurately calibrated by using dedicated procedures, which should be repeated regularly.

X-ray CT images may be corrupted by other typical artifacts related to the X-ray source, the detection system or the patient itself.

Ring artifacts: Response differences among detector elements or the lack of stability of the intensity of the X-ray beam introduce typical circular or ring artifacts. These artifacts can be more or less eliminated by the specific preprocessing of the sinogram or post-processing of the image, but sometimes at a loss of resolution in the image.

Beam hardening artifacts: In standard CT systems, the X-ray beam is not monochromatic as is assumed in the theory of image reconstruction, but is polychromatic. Thus, the shape of the energy spectrum of the incident X-ray beam is modified during propagation in tissue. For example, photons with lower energies may be completely absorbed. In this situation, the physical problem becomes nonlinear and is no longer modeled by a simple Radon transform. If $I_0(E)$ is the incident energy spectrum, the total number of transmitted photons N should be

$$N = \int I_0(E) \ \exp\left(- \int_D \mu(x, y, E) \, \mathrm{d}s \right) \mathrm{d}E. \tag{11}$$

Since the reconstruction method does not take this nonlinearity into account, so-called beam hardening artifacts appear in the reconstructed image. These artifacts can for example be observed as ghost lines joining hard tissues in the CT image. To minimize this problem, X-ray beams used in commercial scanners are generally filtered with thin layers of aluminum or copper to remove low-energy photons. Also linear water-based correction algorithms are typically implemented on all current CT scanners that eliminate the so-called cupping artifact in pure water samples, which is indeed a beam hardening artifact. Metal artifacts caused for example by

metallic implant are also caused by beam hardening because an implant prevents X-ray transmission along certain directions. It should be noted that even in the absence of visible beam hardening artifacts, the use of a polychromatic X-ray beam does not allow interpreting the reconstructed image as a simple map of the linear attenuation coefficients for a given energy since it integrates the nonlinear contribution of the linear attenuation coefficients at different energies.

Motion artifact: A strong requirement for good image quality is data consistency, which means that all projections are from the same object. This assumption would always be fulfilled if data acquisition was instantaneous. However, since acquisition time is finite, recorded data may be inconsistent due to patient movement or breathing. This induces motion artifacts which appear in the form of fuzzy or distorted structures. Some methods have been proposed to correct motion artifacts but in general their assessment is nontrivial. Nowadays motion artifacts are significantly reduced due to the considerable reduction of acquisition time in multi-detector CT. Nevertheless, even if acquisition time is on the order of a few milliseconds, it remains too long with respect to cardiac motion; thus, in cardiac imaging prospective or retrospective EGC gating is used.

Truncation artifacts: The theory of image reconstruction assumes that the projections are completely known. In practice, it happens that the patient is larger than the field of view of the detectors for some or all incidences of the X-ray beam. This results in truncated projections generating truncation artifacts. Truncation can compromise the accuracy of the reconstructed image. Reconstruction from truncated projections is also known as local tomography (Natterer 1986). Theoretical works have shown that under some conditions, it is still possible to reconstruct the image (Clackdoyle et al. 2004).

Noise: Noise in X-ray CT is mainly quantum noise directly related to the total number of transmitted X-ray photons. It can be shown that the SNR in the reconstructed image varies as the square root of the product of tube current and exposure time (mA s product). Thus, quantum noise can be reduced by increasing mA s but the dose to the patient linearly scales with the mA s product. Thus, a compromise must be found to get an acceptable SNR while limiting the dose to the patient.

Evaluation of image quality: In practice, image quality is measured with various phantoms (Kalender 2005). For spatial resolution, test patterns containing objects of different sizes are offered by CT manufacturers. Resolution contrast can be assessed by phantoms made of objects of different known densities. Noise is often evaluated with a water phantom by measuring the deviation in a homogeneous region. Radiation exposure can be measured by dosimeters quantifying the energy deposited during a scan. In practice, CT operators have the possibility to act on some parameters such as: the voltage of the X-ray tube (kV) and the current of the X-ray tube (mA), the exposure time (s), the reconstructed slice thickness, and the reconstruction filter. For a given acquisition protocol an increase in slice thickness and the use of a low-pass filter improve the SNR at the cost of resolution in the image. In practice, the choice of spatial resolution, contrast, SNR, and dose are related and must meet a compromise.

In conclusion, FBP is a fast and efficient method which provides good image quality given the limitations and preconditions discussed above. Nevertheless it is worth noting that iterative algebraic methods currently regain interest. Although ART was initially proposed by Hounsfield, iterative methods were abandoned in clinical CT because they required longer computation times. However, these methods are also more flexible for incomplete datasets since they allow the inclusion of prior knowledge of the object and regularize the inverse reconstruction problem. With progress in computer technology, most manufacturers have recently introduced

advanced iterative reconstruction methods in their system. These methods are expected to allow for dose reduction since they can provide comparable image quality with a smaller number of projections.

3 Historical and Current Concepts of CT Technology

In this section we describe the concepts of CT technology and their evolution. Until the advent of multi-detector scanners, CT systems were traditionally classified into four generations. The subsequent evolutions were driven by the goal of achieving three-dimensional imaging and reducing acquisition time (Kalender 2005).

3.1 Translation–Rotation CT: First and Second Generations

The principle of the first generation of CT scanners corresponded to that proposed by Hounsfield and is illustrated in ❯ *Fig. 6a*. A collimated X-ray source emitting a pencil beam

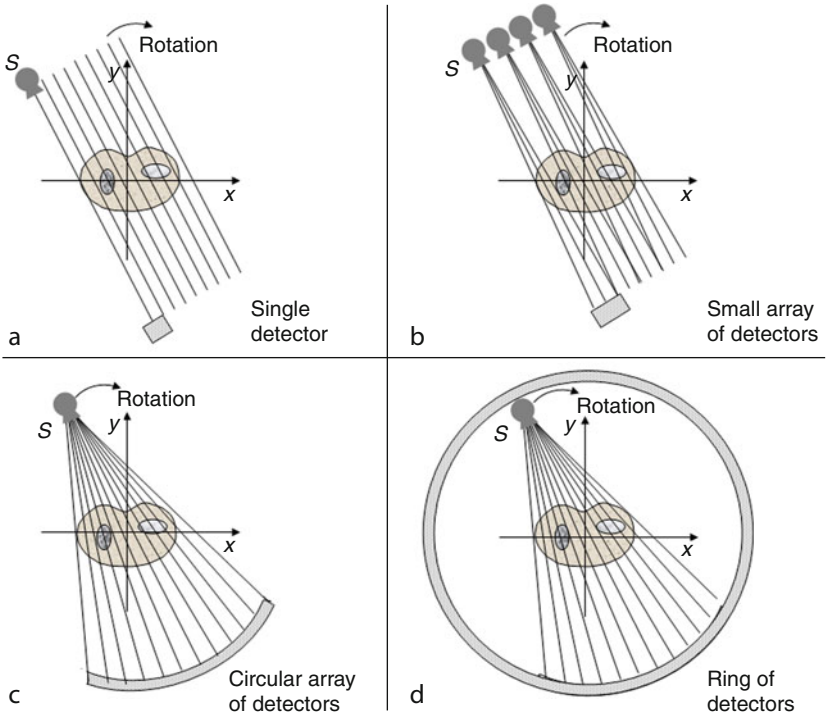


Fig. 6

The four historical generations of X-ray CT: (a–b) Translation–rotation CT, (a) first generation, (b) second generation. (c–d) Fan beam CT, (c) third generation, (d) fourth generation

was used to measure a transmitted intensity on a single sodium iodide (NaI) scintillation detector located on the opposite side. Source and detector were then linearly translated together in the same direction to record a new measurement. This operation was repeated sequentially to acquire a series of measurements of parallel rays covering the section and corresponding to a 1D projection. Then, source and detector were rotated and a new series of translations was performed. Data acquisition stopped when the rotation had covered a total angle of 180°. On Hounsfield's prototype, the data consisted of 180 projections taken at a step of 1° with 160 points each and the reconstructed image was made of 80 × 80 pixels. The first scans on phantoms took nine days with reconstruction times of more than 2 h.

The second generation was a slight variant of the first (◉ *Fig. 6b*). The systems were still based on the principle of translation and rotation, but the single detector was replaced by a small array of detectors (between 8 and 30). They included a narrow fan-beam with an aperture angle between 3° and 10°. Thanks to the simultaneous irradiation of multiple detector elements, acquisition time was reduced to a few minutes per section. In 1974, the first commercialized CT scanners built on this principle allowed to image sections of the head. One year later, whole-body CT scanners offering a wider field of view were introduced to image any section of the human anatomy. Nevertheless, the scan which took approximately 5–6 min could be corrupted by respiration and patient motion, thus leading to artifacts in the reconstruction image.

3.2 Fan-Beam CT: Third and Fourth Generations

A major goal for the next scanner generations was to reduce acquisition times to below 20 s per scan to avoid motion artifacts. This was achieved by abandoning the translation thanks to the use of large arrays of fan-beam detectors, which however necessitated the introduction of fan-beam geometry. In current CT scanners the fan angle varies between 30° and 60° to irradiate simultaneously an entire section of about 50 cm in diameter. Fan-beam geometry also better exploits the power of the X-ray source, because in pencil-beam acquisition most photons are lost due to collimation.

The third- and fourth-generation CT scanners differ by their detectors. In the third-generation systems (◉ *Fig. 6c*), the fan beam irradiates a large linear or circular array of detectors made of 400–1,000 elements and thanks to slip ring technology, the X-ray source can rotate continuously. About 1,000 projections are measured per rotation and total scan times of less than 20 s are typically achieved. The fourth-generation systems (◉ *Fig. 6d*) include a complete ring of detectors surrounding the patient. The X-ray source rotates continuously inside (or sometimes outside) the detector ring and the signal reaching the detector is sampled every few milliseconds. Two types of detectors were used: gas detectors (xenon ionization chambers) or scintillation detectors made of a crystal such as sodium iodide coupled to an electronic light sensor such as a photodiode.

Third- and fourth-generation systems have their own advantages and drawbacks. For example, if one detector element is defective, the reconstructed image will suffer from ring artifacts in a third-generation system while the defect will be less visible in a fourth-generation system since the erroneous values will not be present at the same location in all projections. Conversely, third-generation systems are less sensitive to scatter phenomena. Fourth-generation systems can be regarded as a side step in the technological development; they coexisted with third-generation systems in the 1980s but never really replaced them.

Fan-beam geometry requires the adaptation of the filtered back projection method described in the previous section for parallel beams. In fan-beam geometry all X-rays originate from a single source point. There are several methods to map the fan-beam projection into parallel projection data. One is rebinning, i.e., the reorganization of fan-beam projections into a parallel dataset, which is done before or after the filtering process. A more efficient procedure is the generalization of the FBP algorithm to fan geometry (Herman 1980). For this purpose some weights are introduced in the backprojection operator and the filter is modified if circular detectors are used. In parallel geometry projections must be acquired over a total angle of 180°, while in fan-beam geometry, an angle of at least 180° plus the half fan angle must be covered.

3.3 Spiral or Helical CT

The development of spiral CT or helical CT was driven by the goal to extend CT acquisition to several slices to move toward three-dimensional imaging. Actually the acquisition of a single slice became rapidly limited for diagnosis because it only provided partial information to the radiologist. For instance, it is not possible to determine the z-extension of a tumor or to see the surrounding organs. The first solution to this problem was to acquire a sequence of consecutive slices in "step-and-shoot mode" by moving the patient table between the slice acquisition. In theory, this should provide a stack of slices representative of a 3D volume. In practice, however, this procedure had severe limitations: the sections were not necessarily contiguous, the spatial resolution was lower in the z direction corresponding to displacement and the sections could be misaligned due to patient motion between each slice.

Finally, the technique was not reproducible and was abandoned and replaced by spiral or helical CT (Kalender 2005) which is now standard on all clinical whole-body scanners. The idea of spiral CT is to perform a continuous acquisition while the table moves slowly in the z direction (axis perpendicular to the cutting plane). Thus relative to the patient the X-ray source describes a spiral trajectory instead of a simple circle (see ❯ *Fig. 7*). Of course, spiral CT scanners can also be used in sequential mode.

An important dimensionless parameter in spiral CT is the *pitch* p defined as the ratio of the table displacement d in mm for a rotation of 360° and the slice thickness S:

$$p = \frac{d}{S}. \tag{12}$$

Typical settings for p vary between 1 and 2. A pitch of 1 corresponds to a table displacement equal to the collimation: if p is smaller than 1, imaged sections overlap and for p larger than 1, sections are disjoint. Larger pitch values enable faster acquisition times and reduce radiation exposure at the expense of spatial resolution in z direction.

For spiral CT, FBP techniques again had to be adapted since the data acquired during a 360° rotation do not come from the same section and the resulting sinogram is no longer comparable with the one acquired in single-slice scanning. In spiral CT the measured projections are interpolated in z direction to generate virtually consistent sinograms that can be used by FBP. These sinograms correspond to slices with arbitrary positions, i.e., in contrast to sequential scan modes the z-position of the reconstructed tomogram can be chosen retrospectively. Also the reconstructed slice thickness can be determined retrospectively and it may deviate from the collimation selected for the acquisition.

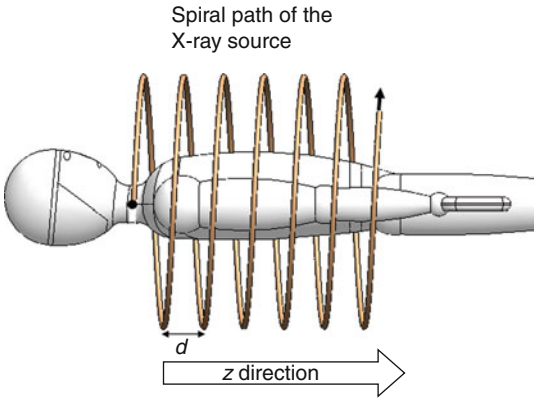


Fig. 7

Scheme of spiral or helical CT: the patient table is translated during X-ray acquisition. It is virtually equivalent to rotate the X-ray source on an helical trajectory

The first spiral interpolation method is known as 360° Linear Interpolation (LI) and estimates the projection value of an arbitrary slice as the weighted combination of the two data with an angular spacing of 360° that are closest to the section to be reconstructed. This method has the disadvantage of requiring an angular range of 2 × 360° to reconstruct a single slice (and introduces a loss of resolution in z direction). The second method, termed 180° LI, takes into account the data redundancy over 360°, since each projection ray is measured twice. By rebinning the data, it is thus possible to generate an additional virtual helix providing a denser angular sampling. Then the virtual sinograms for any slice can be interpolated by using a total rotation angle of two times 180° plus the fan angle. This algorithm, which is the most widely used today, allows choosing a pitch value up to 2.

A variant of this algorithm called 180° WI implements the interpolation directly in the backprojection by introducing a projection-dependent weighting factor. This process increases the computational cost of the backprojection but avoids the interpolation step. Higher-order interpolations (180° HI) have also been proposed but are hardly used in commercial CT scanners. One drawback of these methods is that noise in the reconstructed images depends on the position of the section and increases with the distance to the center of rotation. This property is visible for example in Maximum Intensity Projection (MIP) renderings in the form of horizontal stripes at the image periphery. More sophisticated reconstruction techniques such as 180° Additional filtering and Interpolation(AI) can reduce this problem by including an additional filtering such as Wiener filtering after interpolation.

3.4 Multi-slice CT (MSCT)

Spiral CT scanners have evolved to multi-row detector or multi-slice CT scanners (MSCT). The principle remains the same but the acquisition is performed with a multi-slice instead of a single-slice detector array. Acquisition time considerably decreases since it is reduced by a

factor equal to the number of detector rows. Due to their high quantum efficiency, scintillation detectors for instance built of Gd_2O_2S gadoliniumoxysulfide crystals coupled to photodiodes are used. In addition, X-ray tube heating is also reduced and longitudinal spatial resolution is improved. Since the 2000s, the number of rows has increased to 8, 16, and then to 64 with rotation time down to 0.3 s. With 16-slice CT systems sub-millimeter volumes can be acquired at a speed of 48 mm/s corresponding to short breath-hold times.

Despite the better exploitation of the X-ray output in multi-slice detectors, the increasing use of volumetric scanning places high demands on the power of the X-ray tubes. New concepts that replace the standard rotating anode with a rotating tube design allow for much more efficient cooling and power values of up to 100 kW. Today, in CT scanners, X-ray tubes are typically operated with voltages between 80 and 140 kV and currents between 10 and 500 mA. X-rays are usually filtered with 1–2.5 mm of aluminum or 0.5–1 mm of copper to eliminate as much as possible low-energy photons to reduce beam hardening artifacts discussed in Sect. 2.4. The maximum tube current is generally lower than in the sequential scans to avoid overheating of the anode for long scans. Recent CT scanners allow acquisition times up to 100 s with currents of 500 mA.

In MSCT, if M is the number of rows of the detector array, the concept of pitch is generalized to

$$p = \frac{d}{MS}. \tag{13}$$

For example, for a four slice CT ($M = 4$), a slice thickness of 1 mm and a table feed of 6 mm/360°, the pitch is 1.5. Conversely, the pitch, the slice thickness and the number of detector rows determine the speed of the table.

MSCTs also require some adaptations of the image reconstruction algorithms. The 180° Multi-slice Linear Interpolation (MLI) algorithm is a generalization of the 180° LI algorithm presented in the previous section. This algorithm is suitable as long as the number of detector rows remains small (< 8). For a higher number of rows between 8 and 64 slices, usually the Advanced Single-Slice Rebinning algorithm (ASSR) is used (Kachelriess et al. 2000). It estimates parallel sinograms corresponding to oblique sections better suited to the spiral trajectory. In a post-processing step, the stack of reconstructed oblique slices is interpolated to generate a new stack of parallel sections in the usual coordinate system.

For MSCT systems with 128 or 256 detector rows, the ASSR is no longer adequate because of the large X-ray cone angle. Reconstruction from cone-beam projections is more complex. Tuy defined sufficiency conditions for exact cone-beam reconstruction (Tuy 1983) which are satisfied by a spiral but not by a circular source trajectory. In cone-beam geometry an important concept is the so-called PI-line, which is defined as a line joining two points S_1 and S_2 on the trajectory located at an interval of less than 2π. Any point on this line can then be reconstructed accurately from a set of source points located in the segment $[S_1, S_2]$. By using this idea, Katsevich proposed various inversion formulas, which can be expressed in terms of filtered backprojection (Katsevich 2004a, b). In the case of divergent helical scanners, it is theoretically possible to use these exact algorithms. However, approximated algorithms either based on a reorganization into parallel data (Kachelriess et al. 2000) derived from the Feldkamp algorithm (Kachelriess et al. 2004) are most currently used. Further details on cone-beam reconstruction are given in Sect. 6.1.

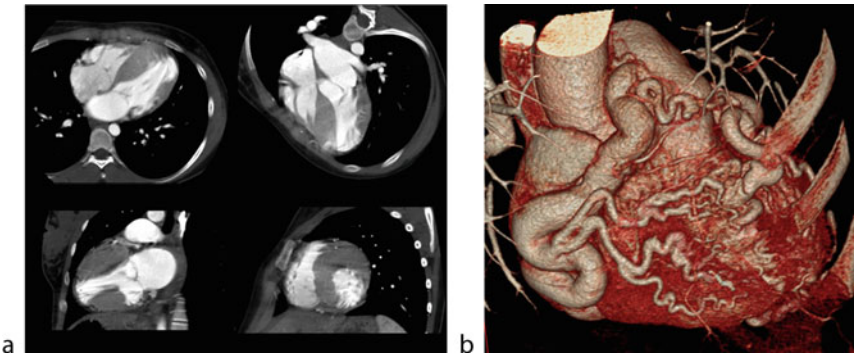


Fig. 8
(a) Different CT slices of the heart acquired in MSCT, and (b) rendering of the heart wall (courtesy of Dr. L. Boussel, CREATIS, Université de Lyon)

4 New Developments in CT Technology and Applications

MSCTs are routinely used in the clinic today. Scans times are typically 1 s or less per image with a total scan time lower than 1 min. They can produce realistic images of the whole body with almost isotropic spatial resolution of about 0.5 mm; thus, partial volume artifacts are significantly reduced. The slice thickness can be chosen between 0.5 and 10 mm, scan times vary between 0.3 and 2 s for a 360° rotation, and it is possible to achieve 10–60 slices per rotation (Kalender 2005). Large areas can be covered allowing rapid and even dynamic investigations of whole organs. State-of-the-art display techniques include multi-planar reformations (MPR), maximum intensity projections (MIP), and a variety of 3D rendering methods. As an example, *Fig. 8a* illustrates different CT slices of the heart acquired in MSCT and *Fig. 8b*, a 3D rendering of the heart. CT technology is further evolving aiming at still faster scan times, higher resolution, and lower radiation exposure coupled with advanced quantitative image analysis. Here, we can only give a limited glimpse into these applications, which already demonstrates the immense value of CT in medicine.

4.1 Perfusion CT

Dynamic CT or perfusion CT (PCT) is used to determine time-dependent attenuation curves of contrast agents such as xenon or iodine from which functional instead of morphological parameters can be quantified. In dynamic CT repetitive or permanent gantry rotation without table feed is used. With increasing number of detector rows, larger volumes of interest (VOIs) can be covered but 4–5 cm typically covered with 64-row scanners still do not suffice to cover whole organs. So-called continuous periodic (Haberland et al. 2010) or toggling table movement techniques (Roberts et al. 2001) have been developed to address this deficiency but this may only be an intermediate step toward the preferred use of CT scanners with larger-area detectors (Gupta et al. 2008; Salomon et al. 2009) which are currently introduced in the market.

The main application of PCT is the support of stroke diagnosis using hemodynamic parameters such as cerebral blood flow (CBF), cerebral blood volume (CBV), and mean transit time (MTT) (Wintermark et al. 2005). Depending on the properties of the contrast agent several theoretical models have been developed to derive these parameters from the time-dependent attenuation curves (Wintermark et al. 2001) – the Kety–Schmidt model for xenon, a dissolvable agent or the maximal slope model and the central volume principle for iodine, which to first order is assumed to be a nondissolvable contrast agent. The Kety–Schmidt model is based on the Fick principle stating that the change in the amount of a substance in tissue (Q) is given by the arterio (C_a) – venous (C_v) concentration difference of the tracer times the flow:

$$Q(t) = \mathrm{CBF} \int_0^T (C_a(t) - C_v(t))\,\mathrm{d}t. \tag{14}$$

From this equation, the Kety–Schmidt model assuming a single-tissue compartment can be derived as (Wintermark et al. 2001)

$$C(t) = \mathrm{CBF} \int_0^T C_a(t)\,\exp\left(-k(T-t)\right)\,\mathrm{d}t. \tag{15}$$

where $C(t)$ is the concentration of the tracer in cerebral parenchyma and $C_a(t)$, the arterial blood concentration at time t. $k = \mathrm{CBF}/\lambda$, where k is the enhancement parameter and λ so-called partition coefficient, which is the ratio of the tissue to venous blood concentration. If the tracer is freely diffusible then k is time independent, which means that there is equilibrium of the xenon concentration between parenchyma and local capillary blood. This is one important assumption in xenon CT. The single-tissue compartment assumption means that for each voxel the parenchyma is assumed to be homogeneous. The determination of arterial and parenchymal attenuation, i.e., concentration curves from CT measurements at different time points thus allow the calculation of CBF for each voxel. Typical protocols include six repetitive scans of the complete head within a period of about 5 min (Wintermark et al. 2005).

Unlike in xenon CT, in dynamic perfusion using iodine, a continuous CT scanning protocol is selected because iodine passes the vascular system much more rapidly than xenon. After bolus injection iodine equilibrium in blood is reached within a few seconds before it is eventually slowly washed out in the kidney. The maximum slope model, which was developed for the description of the distribution of microspheres in capillary networks, makes the assumption that in ❯ Eq. 14 $C_v = 0$ i.e., that the venous concentration is zero. Then (Konstas et al. 2009)

$$\mathrm{CBF} = \frac{[\mathrm{d}Q(t)/\mathrm{d}t]_{\mathrm{Max}}}{[C(t)]_{\mathrm{Max}}}. \tag{16}$$

The flow is the ratio of the maximum slope of the tissue times density curve to the peak arterial concentration. The assumption $C_v = 0$ is only valid within the first 4–6 s after injection (Konstas et al. 2009), which requires a high injection rate typically not achievable in vivo. As a consequence, so-called deconvolution-based models are used, which are more adequate for lower injection rates (Wintermark et al. 2001; Konstas et al. 2009; Kudo et al. 2010). A more detailed description is beyond the scope of this section.

While stroke is one of the major applications, PCT is also used for the myocardial perfusion imaging (MPI) (Ho et al. 2010; Stanton et al. 2010), for the lung perfusion studies, for example

to diagnose embolism (Thieme et al. 2010a, b; Hoffman and Chon 2005), and increasingly for tumor perfusion studies in various other organs such as liver or kidney. One major goal in perfusion studies is the reduction of radiation exposure. One approach is to use intermittent instead of permanent radiation.

4.2 Quantitative CT (QCT)

The term quantitative CT (QCT) is not well defined as principally all analyses performed in a CT dataset beyond a pure visual interpretation include quantitative aspects. Historically the acronym QCT has been used to denote the quantification of tissue concentration, for example of hydroxyapatite (HA), which is the major constituent of bone mineral, of lung nodules (Zerhouni et al. 1982), or of lung parenchyma (Kalender et al. 1990; Gould et al. 1991; Heremans et al. 1992). Obviously PCT is also a quantitative technique. The renewed interest in dual-energy CT (see next section) will certainly increase the variety of QCT applications because it allows for the quantification of any base material concentration.

The determination of bone mineral density (BMD) is one of the early applications of QCT. Single- and dual-energy approaches exist and a detailed description of the techniques applied has recently been published in a report on bone densitometry of the International Commission on Radiation Units (ICRU 2009). In single-energy QCT, the CT values must be converted to BMD values. For this purpose a calibration phantom containing different (HA) concentrations is used, which is typically measured together with the patient. The phantom consists of water equivalent material without HA ($HA = 0\,mg/cm^3$) and a compartment with a HA concentration of $200\,mg/cm^3$. The BMD value of a given pixel, e.g., in the spine, BMD_{tiss}, can be determined from the CT value of the pixel CT_{tiss} by

$$BMD_{tiss} = CT_{tiss}\frac{200}{CT_{200} - CT_0}, \tag{17}$$

where CT_{200} and CT_0 are the measured CT values in the two compartments of the calibration phantom. In the dual-energy approach also, a phantom would be included in the measurement to determine the mass attenuation coefficients of the two base materials, which in the case of bone densitometry would be water and bone HA equivalent materials.

Traditional BMD measurement sites have been the lumbar spine (Genant et al. 1982; Kalender et al. 1987) and the distal forearm. With the introduction of spiral CT, which allows for the rapid acquisition of a large volume, the application has been extended to the proximal femur (Engelke et al. 2008). For the spine and the femur, clinical whole-body CT scanners are used (*Fig. 9a, b*). Special dedicated smaller peripheral QCT scanners (pQCT) have been developed specifically for the forearm, which typically operate with tube voltages between 45 and 60 kV which is more adequate for the forearm compared to 80 or 120 kV typically available on whole-body CT scanners. More recently a new high-resolution pQCT scanner with the aim of quantifying trabecular and cortical bone structure of the distal radius and tibia has extended the diagnostic instrumentation of BMD measurements. This high-resolution scanner is equipped with an area detector covering a width of 9 mm. It offers excellent spatial resolution down to a voxel size of $(40\,\mu m)^3$ at the expense of a spiral mode and relative long scan times of around 3 min. *Figure 9c* demonstrates the spatial resolution of a high-resolution pQCT scanner.

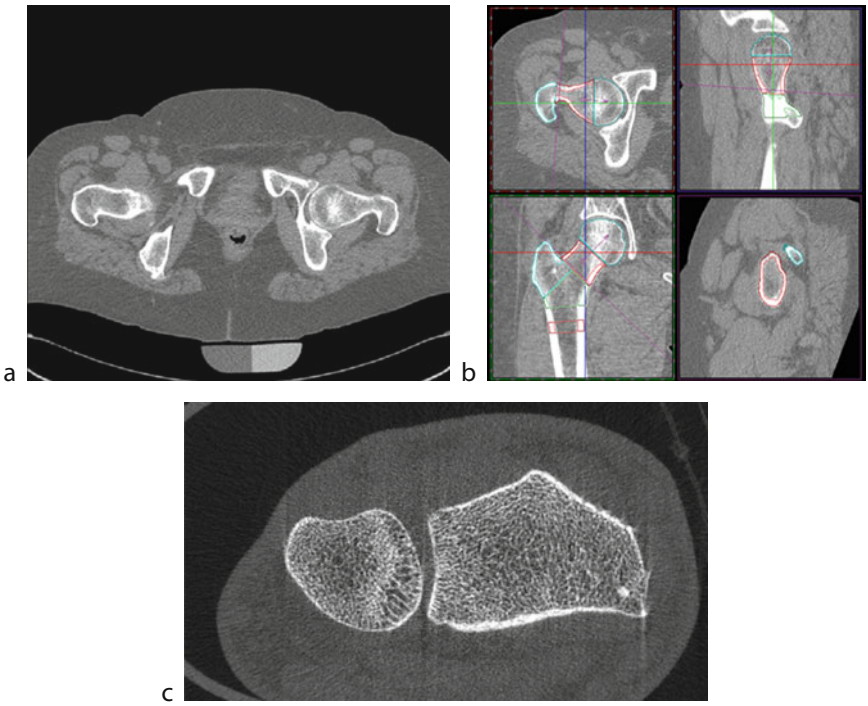


Fig. 9
(a–b) QCT of the femur: (a) one slice of the 3D stack with calibration phantom, (b) MPR with segmented femur, (c) high-resolution peripheral QCT of the distal radius using dedicated peripheral CT scanner

4.3 Spectral CT

The aim of spectral CT imaging is the isolation and quantification of chemical elements or compounds such as iodine or calcium hydroxyapatite. Another purpose is the reduction or elimination of beam hardening artifacts. Element selective imaging by CT can be performed in several manners. In the K-edge approach the object is scanned at two energies below and above the edge of the element of interest and a weighted subtraction image is calculated (Kruger et al. 1977; Riederer and Mistretta 1977). Dual-energy imaging exploits the difference of the energy dependence of the Compton and photoabsorption cross sections. In this case a single element can be distinguished from its matrix if it differs largely in Z. In the same manner a group of heavier elements can be separated from a group of lighter elements. Other element selective techniques such as fluorescence imaging require the use of monochromatic radiation. However, when using X-ray tubes as in clinical CT, monochromatization of the emitted energy spectra would reduce the usable intensity below practical needs; therefore, here we assume that always a broad energy spectrum is used.

It was shown in a series of theoretical papers (Alvarez and Macovski 1976; Lehmann et al. 1981; Hawkes et al. 1986) that in the range of diagnostic X-ray energies ($30\,\text{keV} < E < 200\,\text{keV}$),

the energy dependence of the mass attenuation coefficient μ/ρ of most biological tissues can be parameterized with sufficient accuracy by a linear combination of two energy functions:

$$\left(\frac{\mu}{\rho}\right)(E) = a_1 f_1(E) + a_2 f_2(E). \tag{18}$$

The above equation is only valid as long as no absorption edges are present, which is the case for biological tissues in the energy range given above. The functions $f_1(E)$ and $f_2(E)$ can be represented by the mass attenuation coefficients of two so-called base materials. The parameterization above implies that a measurement of μ at two different energies suffices to describe its energy dependence in the diagnostic range, in other words the measurement of μ at a third energy results in redundant information. More importantly, the equation implies that a given tissue mixture can be separated into two components or conversely that if a tissue mixture contains just two components, then their mass concentrations can be determined. Thus, we obtain a linear system of two equations for the two X-ray energies E_L and E_H and the two base materials 1 and 2:

$$\left(\frac{\mu}{\rho}\right)^L = c_1\left(\frac{\mu}{\rho}\right)_1^L + c_2\left(\frac{\mu}{\rho}\right)_2^L \quad \text{and} \quad \left(\frac{\mu}{\rho}\right)^H = c_1\left(\frac{\mu}{\rho}\right)_1^H + c_2\left(\frac{\mu}{\rho}\right)_2^H. \tag{19}$$

The constants c_i are the mass concentrations ρ_i' of the base material i, divided by the mass density ρ of the investigated volume: $c_i = \rho_i'/\rho$. As the mass attenuation coefficients of the two base materials $(\mu/\rho)_{1,2}$ are known for both energies, by elimination of c_2 the set of equations can be solved for the mass concentrations ρ_1'.

If more than two materials are present, only an equivalent base material concentration is obtained. Obviously the choice of base materials 1 and 2 is arbitrary and the equation system above can be solved even if the selected base materials are not even contained in the sample. Mathematically the change of base materials is a base transformation in a vector space, i.e., a combination can be selected so that for one of the materials present in the sample $\rho_i' = 0$, which means than this material can be "removed" from the image.

In CT significant efforts were already made during the 1980s to develop dual-energy CT for quantifying bone mineral density (BMD) either by post-processing two images taken at 80 and 120 kV, respectively (Brooks et al. 1977; Genant and Boyd 1977), or by preprocessing two sets of projections acquired by rapidly switching the tube voltage between each projection (Kalender et al. 1986; Vetter et al. 1986). The latter technique requires special hardware but is less sensitive to movement artifacts and corrects for beam hardening artifacts. With respect to BMD measurements, the diagnostic relevance of dual energy was limited because the increased accuracy was accompanied by poorer precision and as a consequence the approach was abandoned.

The goal of reducing scan times in particular in cardiac applications along with the difficulties to further increase the X-ray tube output power led to the development of dual-tube/detector CT (DSCT) scanners (Flohr et al. 2006), which again sparked interest in multiple-energy CT. Recent applications of dual-energy CT include bone (Lell et al. 2007; Kemmling et al. 2010; Zhang et al. 2010) or plaque removal for the separation versus iodine in angiography applications (Thomas et al. 2010), mapping of the iodine distribution of the myocardium (Ruzsics et al. 2009; Bauer et al. 2010), lung perfusion (Kang et al. 2010; Thieme et al. 2010a, 2010b), or the discrimination of the elemental composition of urinary stones (Zilberman et al. 2010).

The application of preprocessing dual-energy algorithms is not possible in current dual-source CT scanners because the scan geometries of the two raw datasets differ. As both energy raw datasets are sampled already at the same time, motion artifacts are no longer a problem for post-processing techniques but differences in beam hardening of the two absorption images

remain. One promising approach to this problem is the linear combination of so-called generalized precorrected projections to calculate material selective images instead of the combination of projections including water-based linear beam hardening correction (Maass et al. 2009).

5 Radiation Exposure

The rapid progress of CT technology in the last two decades has resulted in an increasing number of CT installations and an increasing frequency and type of examinations (IAEA 2009). While this reflects the diagnostic relevance of CT it also raises some concerns from the perspective of radiation exposure. The dose per CT examination is increasing as often a large part of the body is scanned, spatial resolution is improved, or dynamic scan protocols are used. For example, in the US in 2006 CT accounted for 17% of all radiological contributions but for almost 50% of the collective dose (Mettler et al. 2008). To address these concerns several techniques that will be described in more detail below have been developed to reduce dose.

The standard physical measurement in CT to quantify dose is the Computed Tomography Dose Index (CTDI), which was introduced to also account for out-of-plane scattered radiation that exposes tissue not illuminated by the direct X-ray beam. As a consequence the radiation exposure of a series of slices is larger than the sum of the exposure of the individual slices. The CTDI is determined using a 100-mm ionization chamber inserted in little openings of Lucite phantoms (see ● *Fig. 10*) with a diameter of 16 cm (head) or 32 cm (body) (Kalender 2005):

$$\mathrm{CTDI}_{100}\,[\mathrm{mGy}] = \frac{1}{S}\int_{-50\,\mathrm{mm}}^{50\,\mathrm{mm}} D(z)\,\mathrm{d}z. \tag{20}$$

where the index 100 refers to the length of the radiation chamber, D is the dose, and S the slice thickness. The average dose in the scanning plane can be described by the weighted CTDI_w:

$$\mathrm{CTDI}_w\,[\mathrm{mGy}] = \frac{1}{3}\mathrm{CTDI}_{100}(\mathrm{center}) + \frac{2}{3}\mathrm{CTDI}_{100}(\mathrm{periphery}). \tag{21}$$



● Fig. 10
CDTI body (Ø 32 cm) and head (Ø 16 cm) phantoms with circular inserts of a 100 mm radiation chamber

$CTDI_{100}$ is proportional to the mA s product and approximately independent of the slice thickness. In spiral acquisition modes depending on the pitch p a spiral overlap or gap may occur. Thus for volumetric scan modes a volume $CTDI_{vol}$ has been introduced:

$$CTDI_{vol}\,[\mathrm{mGy}] = \frac{CTDI_w}{p} \quad \text{with} \quad p = \frac{d}{M\,S}, \tag{22}$$

where d is the table feed in millimeter and M the number of simultaneously scanned slices. In order to compare the radiation exposure among different scanners or among different acquisition protocols of a given scanner the CTDI is normalized to the product of tube current and rotation time (typically $I_{tube}\,t_{rot} = 100$ mA s):

$${}_{n}CTDI_w\,[\mathrm{mGy}] = \frac{CTDI_w}{I_{tube}\,t_{rot}}. \tag{23}$$

It is important to note that CTDI parameters quantify CT-scanner dose characteristics, not patient doses. For the latter purpose the effective dose values in mSv must be used. Effective dose values E for typical CT examinations are for example summarized by the United Nations Scientific Committee on the Effects of Atomic Radiation (UNSCEAR 2008). Effective doses are often estimated from the dose length product (DLP)

$$DLP\,[\mathrm{mGy \cdot cm}] = CTDI_{vol} \cdot \text{irradiate length} \quad E = k\,DLP \tag{24}$$

using dedicated conversion factors (Jessen et al. 1999). However, recently major inaccuracies of this estimation have been reported (Christner et al. 2010). Higher accuracy can be achieved by calculating effective dose values based on tissue weighting factors published by the International Commission on Radiation Protection (ICRP 2007) using Monte Carlo simulations with virtual whole-body phantoms (Deak et al. 2008).

For patient exposure the ALARA principle (as low as reasonable achievable) is always valid. This means that the exposure to the patient ultimately has to be determined by the requirements of the diagnostic purpose and not by the desire to obtain pretty images. Dose values for typical X-ray and CT examinations are defined as so-called diagnostic reference values and are published by the national and international authorities. The main parameters that determine image quality are noise, spatial resolution, and contrast resolution. Their dependency can be conveniently graphed in a contrast detail dose diagram. At a given radiation exposure the relationship between image noise and spatial resolution can be altered by using different reconstruction kernels.

The radiation exposure delivered to the patient depends on several acquisition parameters. Most important the dose increases linearly with the $I_{tube}\,t_{rot}$ (mA s) product. Dose also increases approximately linearly with an increase of the scan volume; details depend on the illuminated organs. In spiral CT dose scales with $1/p$, i.e., relative to a pitch of 1 a pitch of 2 decreases the dose by 50% and a pitch of 0.7 increases it by 42%. The slice thickness is of lesser importance because in volumetric scanning it can be set during the reconstruction process. However, at a given exposure the reduction of slice thickness increases the noise in the images. If the noise should be constant then the dose must be increased linearly with decreasing slice thickness. Another important parameter is patient size. Image noise approximately scales linearly with patient weight (IAEA 2009). Thus, in order to achieve the same noise level a weight difference by a factor of 2 would require a dose difference by a factor of 4.

Automatic adaptation of exposure to patient size is one component of automatic exposure control (AEC) (Kalra et al. 2005; McCollough et al. 2006, 2009) which aims to optimally adapt

exposure once the desired image quality in terms of noise and spatial resolution has been specified. One of the first AEC components was anatomy-adapted tube current modulation (TCM), which exploits the fact that in patients the X-ray absorption is higher in lateral than in anterior posterior (ap) direction and therefore the tube current can be significantly reduced in ap direction without increasing noise (Kalender et al. 1999). Indeed the tube current in lateral direction can even be slightly increased so that despite lower dose noise is also reduced. The achievable dose reduction depends on the body location and is largest in the shoulder and the pelvis. The same principle, dose adaptation to the actual absorption can be applied along the z-axis of the CT scanner, which is another component of AEC. Homogeneous noise distribution in the slice of an image stack is another goal of AEC. There are different technical approaches to AEC, and their implementations vary by manufacturer. For modern 16 and 64 detector row CT scanners a dose reduction between 35% and 60% has been reported with a more homogenous image noise distribution compared to fixed mA s acquisition protocols (Söderberg and Gunnarsson 2010).

Another approach to reduce radiation exposure is to decrease noise in the reconstructed images by applying advanced tomographic reconstruction methods such as multi-slice-adapted filtering (Kachelriess et al. 2001; Baum et al. 2004). Also it has recently been shown that the use of iterative reconstruction methods can significantly reduce noise (Hara et al. 2009; Flicek et al. 2010; Prakash et al. 2010), although from a practical perspective these algorithms are still too slow to be used in clinical routine.

6 From Clinical CT to Nano-CT

Current clinical CT scanners can produce images with a maximum spatial resolution of about 0.3 mm, and a slice thickness of about 0.5 mm. Despite all the technological advances, the fundamental limitation to increase spatial resolution is radiation exposure, which at a given object size and constant SNR is inversely proportional to the fourth power of the detector resolution. As a consequence, dose increases by a factor of 16 each time the pixel size is halved. Thus, it is difficult to increase spatial resolution of in vivo imaging in humans much below the values given above unless the size of the object to be scanned is reduced. This is for example the case in the periphery and one reason why trabecular structure in humans is assessable in the forearm (Boutroy et al. 2005) but not in the spine. However, micro-CT and nano-CT are receiving increasing interest in various domains going from preclinical imaging and biology to materials science. This is also the topic of Chap. 45, "High-Resolution and Animal Imaging Instrumentation and Techniques".

6.1 Micro-CT Using X-ray Tubes

Micro-CT relies on the same principle as CT but can achieve spatial resolutions up to the micrometer level. The investigation of bone architecture has been a strong motivation for development of these devices in the late 1990s (Feldkamp et al. 1989; Rüegsegger et al. 1996). Although the first micro-CTs were laboratory prototypes, the interest in this type of imaging has led many manufacturers to provide micro-CT systems at the end of the 1990s. Micro-CT can be used in vivo for small-animal imaging but is restricted to ex vivo examination of human biological samples due to dose limitations.

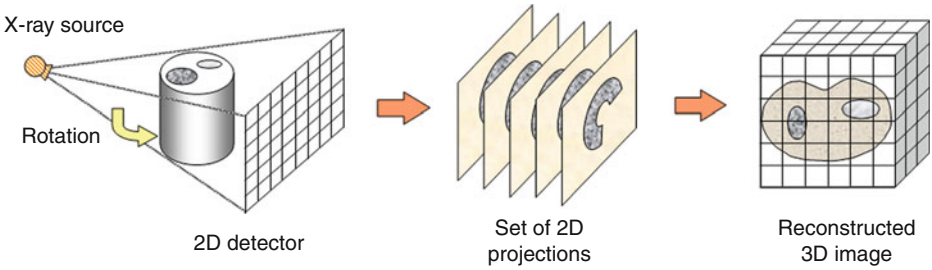


Fig. 11

Principle of 3D cone-beam CT: (*left*) data acquisition with a X-ray source following a circular path providing (*middle*) a set of radiographs which are processed through a cone-beam reconstruction algorithm to obtain (*right*) a 3D digital image

Micro-CT systems using X-ray tubes are mainly based on the concept of *truly* 3D CT using cone-beam geometry. The idea is to acquire 2D projections of a 3D image from different angles as illustrated in ❯ *Fig. 11*. For each X-ray source position, an X-ray radiograph of the object is acquired, that can be regarded as a 2D projection of the object. The 3D digital image is then estimated from this set of 2D radiographs by a specific cone-beam reconstruction algorithm.

The first cone-beam CT prototype was the Dynamic Spatial Reconstructor (DSR) with 14 X-ray sources at the Mayo Clinic used for 3D dynamic cardiac imaging (Ritman et al. 1980). In France, the "Morphometre" was a prototype with two X-ray sources developed by General Electric and the CEA-LETI for angiography (Saint-Félix et al. 1994). However, these first implementations had limited image quality mainly because the detection system consisted of image intensifiers. The introduction of 2D scintillation detectors allowed a fast progression of cone-beam micro-CT. In these systems, the 2D radiographs are acquired with a 2D detector, usually consisting of a scintillator screen, an optic, and a CCD camera.

As already mentioned in ❯ Sect. 3.4, cone-beam CT reconstruction algorithms are more complex because the information relative to a single slice is distributed along different detector lines. Thus, it is not possible to perform a sequence of 2D tomographic reconstructions. This problem is known in the literature as reconstruction from cone-beam projections (Herman 1980; Tuy 1983; Peyrin 1985; Grangeat 1991). It was shown theoretically that a 3D image is completely determined by its cone-beam projections if any plane that intersects the image also intersects the source trajectory (Tuy 1983). Thus, a cone-beam acquisition using a circular trajectory does not uniquely determine the 3D image. This nonuniqueness can be interpreted in terms of missing data in the 3D Radon space. As in 2D, there is an inversion formula expressing the 3D image as a function of its 3D Radon transform defined as integrals on planes. For a circular X-ray path, the 3D Radon transform cannot be estimated on its entire support thus leading to missing data. Hence, reconstructed images from circular cone-beam data are always approximated whatever method is used.

Despite this limitation, several cone-beam reconstruction methods have been proposed. Because of its simplicity, the most popular approach is the Feldkamp algorithm (FDK) (Feldkamp et al. 1984). It is a generalization of the FBP algorithm for cone-beam geometry consisting of three steps: weighting the cone-beam projections, filtering each row of the 2D projections with the usual ramp filter, and cone-beam backprojection of the 2D filtered projection. Unlike the 2D FBP, this formula is not exact, except in the central plane ($z = 0$).

The reconstruction errors increase with the distance to the central plane. Cone-beam artifacts manifest themselves in the form of geometric distortions and intensity attenuations in the vertical direction. They are more visible on high-contrast objects. Many heuristic methods have been proposed to correct cone-beam artifacts in circular cone-beam geometry (Valton et al. 2006).

Cone-beam micro-CTs using a micro-focus X-ray source and a 2D high-resolution detector can achieve spatial resolutions down to a few micrometers. Even if the 2D detector resolution is about 10 μm, the system benefits from magnification brought by the divergent X-ray beam. However, the cone-beam angle increases with magnification, resulting in more pronounced cone-beam artifacts in FDK reconstructions. For this purpose, the useful part of the data is sometimes limited to the central area of the detector.

Manufacturers generally provide different micro-CT systems with different specifications. The characteristics include the range of spatial resolutions and fields of view that they offer, but also the properties of the X-ray tube in terms of energy spectrum. The choice of the system has to consider the size and type of objects to be investigated and the desired level of detail. It can be noted that the higher the spatial resolution, the smaller the field of view. If for example the pixel size in the reconstructed image is 1 μm and the detector has 2,000 × 2,000 pixels then the maximum sample size is limited to diameter and a height of 2 mm each.

For instance, specific systems with rotating gantry are designed for the in vivo examination of small anesthetized animals. Similar to a whole-body clinical CT they include a small table to position the animal in the field of view. These scanners typically offer spatial resolutions between 40 and 100 μm, a FOV between 5 and 15 cm and scan times of up to a few minutes. It is difficult to achieve considerably higher spatial resolutions in animals in vivo because this would require much longer scan times, which could yield motion artifacts. Conversely, systems optimized for the examination of samples and biopsies use smaller FOVs but achieve higher spatial resolution down to the micrometer level. Typically in these systems the sample rotates and a selectable source detector distance provides the possibility to adjust the spatial resolution to the sample requirements.

Through the dissemination of the technique, 3D micro-CT has become a routine technique for studying 3D bone micro-architecture or other porous natural or industrial materials. This success is due to the possibility of acquiring nondestructively a three-dimensional image and not a single section, which offers many advantages in the quantification of morphological or topological properties of materials.

6.2 Synchrotron Radiation Micro-CT

Coupling X-ray micro-CT to synchrotron sources presents a number of advantages in terms of image quality that were first highlighted in the 1980s (Grodzins 1983). Synchrotron radiation is produced when high-energy electrons are forced on a circular trajectory. Synchrotron sources are characterized by their brilliance which is several orders of magnitude higher than that of X-ray tubes. They can be used to provide quasi-monochromatic beams with very high fluxes. The use of a monochromatic beam avoids beam hardening artifacts in the reconstructed images (see ❯ Sect. 2.4). The high photon flux provides the possibility to obtain very high spatial resolution with excellent SNR and reduced acquisition times. Micro-CT experiments have been developed on different synchrotron sources in the world (Engelke et al. 1993).

A micro-CT setup has been developed in the late 1990s on beamline ID19 at the European Synchrotron Radiation Facility (ESRF) in Grenoble (Salome et al. 1999). It consists in recording

2D projections of a sample with a high-resolution detector composed of a scintillator converting X-ray photons into visible light, optic lenses for magnification and a 2,048 × 2,048 CCD camera. The sample is sequentially rotated around the vertical axis to cover a total angle of 180°. Since this system is installed at a distance of 145 m from the X-ray source, it can be considered a truly 3D parallel geometry CT. Thus, a straightforward FBP algorithm can be applied to the different parallel transverse sections of the 3D image. *Figure 12* shows a transverse and sagittal slice extracted from the 3D reconstructed image of the mice femur as well as a 3D display of bone surfaces. The raw slices show the high contrast and high SNR of the image although some ring artifacts can be seen. 3D displays allow visualization of the complex organization of the trabecular network. With appropriate 3D display software, it is possible to rotate the 3D image under any angle and obtain sections in arbitrary planes or even to virtually navigate through these 3D images.

Due to its properties, synchrotron micro-CT images can be seen as quantitative in the sense that it provides a reliable map of the linear attenuation coefficient for the given energy. This property is particularly interesting in bone research since, in addition to studying the structure, the tissue mineralization can also be quantified (Nuzzo et al. 2002). Data acquisition is also fast, for example it is about 10 min at a micrometer spatial resolution while it would be more than 10 h to achieve the same SNR in standard micro-CT.

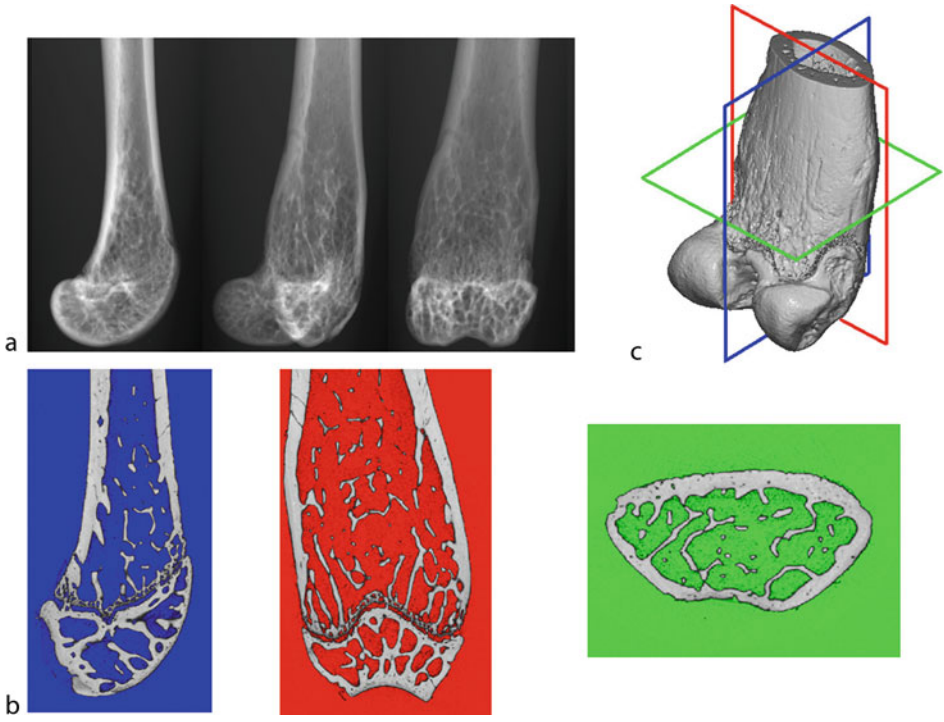


Fig. 12

Illustration of synchrotron radiation micro-CT of a mice femur at 7 μm resolution: (a) 2D projections between 0° and 90°, (b) transverse and sagittal slices, and (c) 3D rendering of bone surfaces

While synchrotron micro-CT is an excellent tool that offers unparalleled spatial resolution and image quality, its availability is limited to a few synchrotron facilities around the world. Thus, these systems are reserved to applications requiring quantitative imaging and are mainly devoted to research.

6.3 Nano-CT

Recently, technical progresses permitted to push the limit to nano-CT, this nomenclature being understood by manufacturers as an imaging technique achieving sub-micrometer spatial resolution. For instance, Skyscan proposed a nano-CT based on an X-ray source having an extremely small focal spot size (< 400 nm), a voltage between 20 and 80 kV, and a 12-bit CCD-based detector. The announced spatial resolution is 400 nm at 10% of the modulation transfer function. The system can image 0.2 mm objects with a voxel size of 150 nm to 11 mm objects with a voxel size of 9 μm and provide 1,280 × 1,280 × 900 digital images. The feasibility of this system was recently demonstrated to examine osteocyte lacunae in bone tissue (van Hove et al. 2009). Nano-CT is also available with synchrotron micro-CT with the same advantages as presented above. The ESRF nano-CT setup has already been used for the analysis of micro-/nano-porosities in bone tissue such as osteocyte lacunae and micro-cracks (Peyrin 2009). Without any doubts, nano-CT systems open attractive perspectives for the three-dimensional investigation of biological samples.

7 Conclusion

X-ray CT, which has been a pioneer technique in medical imaging has evolved from a single-slice to a fully three-dimensional imaging technique. Throughout the years, image quality and acquisition times have been considerably improved by the introduction of new theoretical and technological concepts. Many improvements have been motivated by important problems in public health such as cardiac, cerebral, or bone diseases requiring progresses in cardiac, brain, or bone imaging. X-ray CT is a perfect example in which technical advances have resulted from the synergy of progresses in different domains.

Multi-slice spiral CTs provide in a few seconds sub-millimeter slices covering a sufficiently large region of interest for most clinical applications. Currently 16–64 slices systems are most common but new generation systems can already use up to 128 slices. These systems provide multi-planar slices in the investigated volume and three-dimensional displays of structures of interest. Flat panel detectors coupled to CT scanners, although not yet commercially available, is a promising technique for volumetric imaging on larger fields of view with isotropic spatial resolution. With fast volume scanning and improved detectability, these systems would open new applications in functional imaging.

By using specific calibration, QCT provides not only images but quantitative measurements. With the new developments in spectral CT, exploiting the polychromaticity of the X-ray beam, it is even expected to improve the sensitivity of CT scanners and the discrimination of different tissues.

The improvement of spatial resolution has also been noticeable although clinical CT suffers from a fundamental limitation which is the acceptable dose delivered to the patient.

Specific high-resolution systems have been developed for the investigation of bone microarchitecture but they are limited to peripheral bone. New image reconstruction methods based on iterative reconstruction and image models have recently been implemented on clinical CT to reduce the dose. They actually provide a similar image quality from a smaller number of projections, thus yielding dose reduction.

Radiation exposure is less crucial when imaging small animals or biological samples. In this domain, micro-CT is a considerable success and is becoming a standard tool for biologists. The recently introduced nano-CT technique can be seen as a new three-dimensional microscopic imaging tool expected to offer new perspectives in biology.

Cross-References

- Chapter 8, "Synchrotron Radiation and FEL Instrumentation"
- Chapter 10, "Radiation Protection"
- Chapter 22, "Radiation Damage Effects"
- Chapter 34, "Radiation Detectors and Art"
- Chapter 35, "Radiation-Based Medical Imaging Techniques: An Overview"
- Chapter 39, "Image Reconstruction"
- Chapter 41, "Quantitative Image Analysis in Tomography"
- Chapter 44, "Simulation of Medical Imaging Systems: Emission and Transmission Tomography"
- Chapter 45, "High-Resolution and Animal Imaging Instrumentation and Techniques"

References

Alvarez RE, Macovski A (1976) Energy-selective reconstructions in X-ray computerized tomography. Phys Med Biol 21(5):733–744

Bauer RW et al (2010) Dual-energy CT for the assessment of chronic myocardial infarction in patients with chronic coronary artery disease: comparison with 3-T MRI. Am J Roentgenol 195(3):639–646

Baum U et al (2004) Improvement of image quality of multislice spiral CT scans of the head and neck region using a raw data-based multidimensional adaptive filtering (MAF) technique. Eur Radiol 14(10):1873–1881

Boutroy S et al (2005) In vivo assessment of trabecular bone microarchitecture by high-resolution peripheral quantitative computed tomography. J Clin Endocrinol Metab 90(12):6508–6515

Brooks RA (1977) A quantitative theory of the Hounsfield unit and its application to dual energy scanning. J Comput Assist Tomogr 1(4):487–493

Christner JA et al (2010) Estimating effective dose for CT using dose-length product compared with using organ doses: consequences of adopting International Commission on Radiological Protection publication 103 or dual-energy scanning. Am J Roentgenol 194(4):881–889

Clackdoyle R et al (2004) Quantitative reconstruction from truncated projections in classical tomography. IEEE Trans Nucl Sci 51(5): 2570–2578

Cormack AM (1980) Nobel Award address. Early two-dimensional reconstruction and recent topics stemming from it. Med Phys 7(4):277–282

Deak P et al (2008) Validation of a Monte Carlo tool for patient-specific dose simulations in multi-slice computed tomography. Eur Radiol 18(4):759–772

Engelke K et al (1993) High spatial resolution imaging of bone mineral using computed microtomography. Comparison with microradiography and undecalcified histologic sections. Investig Radiol 28(4):341–349

Engelke K et al (2008) Clinical use of quantitative computed tomography and peripheral quantitative computed tomography in the management

of osteoporosis in adults: the 2007 ISCD Official Positions. J Clin Densitom Off J Int Soc Clin Densitom 11(1):123–162

Feldkamp LA et al (1984) Practical cone-beam algorithm. J Opt Soc Am 1(6):612–619

Feldkamp LA et al (1989) The direct examination of three-dimensional bone architecture in vitro by computed tomography. J Bone Miner Res Off J Am Soc Bone Miner Res 4(1):3–11

Flicek KT et al (2010) Reducing the radiation dose for CT colonography using adaptive statistical iterative reconstruction: a pilot study. Am J Roentgenol 195(1):126–131

Flohr TG et al (2006) First performance evaluation of a dual-source CT (DSCT) system. Eur Radiol 16(2):256–268

Genant HK, Boyd D (1977) Quantitative bone mineral analysis using dual energy computed tomography. Investig Radiol 12(6):545–551

Genant HK et al (1982) Quantitative computed tomography of vertebral spongiosa: a sensitive method for detecting early bone loss after oophorectomy. Ann Intern Med 97(5):699–705

Gould GA et al (1991) Lung CT density correlates with measurements of airflow limitation and the diffusing capacity. Eur Respir J Off J Eur Soc Clin Respir Physiol 4(2):141–146

Grangeat P (1991) Mathematical framework of Cone Beam 3D reconstruction via the first derivative of the radon transform. In: Herman GT, Louis AK, Natterer (eds) Mathematical Methods in Tomography. Lecture notes in mathematics, vol 1497. Springer, Berlin, pp 66–97

Grangeat P (2002) La tomographie: fondements mathématiques, imagerie microscopique et imagerie industrielle (Traité IC2, série traitement du signal et de l'image). Hermès, Paris

Grodzins L (1983) Optimum energy for X-ray transmission tomography of small sample. Nucl Instrum Methods 206:541–545

Gupta R et al (2008) Flat-panel volume CT: fundamental principles, technology, and applications. Radiographics 28(7):2009–2022

Haberland U et al (2010) Performance assessment of dynamic spiral scan modes with variable pitch for quantitative perfusion computed tomography. Invest Radiol 45(7):378–386

Hara AK et al (2009) Iterative reconstruction technique for reducing body radiation dose at CT: feasibility study. Am J Roentgenol 193(3): 764–771

Hawkes DJ et al (1986) Tissue analysis by dual-energy computed tomography. Br J Radiol 59(702): 537–542

Heremans A et al (1992) Measurement of lung density by means of quantitative CT scanning. A study of correlations with pulmonary function tests. Chest 102(3):805–811

Herman GT (1980) Image reconstruction from projections: the fundamentals of computerized tomography. Academic, New York

Ho KT et al (2010) Stress and rest dynamic myocardial perfusion imaging by evaluation of complete time-attenuation curves with dual-source CT. JACC Cardiovasc Imaging 3(8):811–820

Hoffman EA, Chon D (2005) Computed tomography studies of lung ventilation and perfusion. Proc Am Thorac Soc 2(6):492–498, 506

Hounsfield GN (1973) Computerized transverse axial scanning (tomography). 1. Description of system. Br J Radiol 46(552):1016–1022

van Hove RP et al (2009) Osteocyte morphology in human tibiae of different bone pathologies with different bone mineral density – is there a role for mechanosensing? Bone 45(2):321–329

Hubbell JH (2006) Review and history of photon cross section calculations. Phys Med Biol 51(13):R245–R262

IAEA (2009) Dose reduction in CT while maintaining diagnostic confidence: a feasibility/demonstration study

Jessen KA et al (1999) Dosimetry for optimisation of patient protection in computed tomography. Appl Radiat Isot Incl Data Instrum Methods Agric Ind Med 50(1):165–172

Kachelriess M et al (2000) Advanced single-slice rebinning in cone-beam spiral CT. Med Phys 27(4):754–772

Kachelriess M et al (2001) Generalized multi-dimensional adaptive filtering for conventional and spiral single-slice, multi-slice, and cone-beam CT. Med Phys 28(4):475–490

Kachelriess M et al (2004) Extended parallel backprojection for standard three-dimensional and phase-correlated four-dimensional axial and spiral cone-beam CT with arbitrary pitch, arbitrary cone-angle, and 100% dose usage. Med Phys 31(6):1623–1641

Kak AC, Slaney M (1988) Principles of computerized tomographic imaging. IEEE, New York

Kalender WA (2005) Computed tomography: fundamentals, system technology, image quality, applications, 2nd edn. Publicis Corporate, Erlangen. Publicis MCD Werbeagentur Verlag

Kalender WA et al (1986) Evaluation of a prototype dual-energy computed tomographic apparatus. I. Phantom studies. Med Phys 13(3):334–339

Kalender WA et al (1987) Vertebral bone mineral analysis: an integrated approach with CT. Radiology 164(2):419–423

Kalender WA et al (1990) Measurement of pulmonary parenchymal attenuation: use of spirometric

gating with quantitative CT. Radiology 175(1): 265–268
Kalender WA et al (1999) Dose reduction in CT by anatomically adapted tube current modulation. II. Phantom measurements. Med Phys 26(11):2248–2253
Kalender W, Engelke K, Fuerst TP, Glüer C-C, Laugier P, Shepherd J (2009) Quantitative aspects of bone densitometry. J ICRU 9(1):1–130
Kalra MK et al (2005) Computed tomography radiation dose optimization: scanning protocols and clinical applications of automatic exposure control. Curr Probl Diagn Radiol 34(5):171–181
Kang M et al (2010) Dual-energy CT: clinical applications in various pulmonary diseases. Radiographics 30(3):685–698
Katsevich A (2004a) An improved exact filtered backprojection algorithm for spiral computed tomography. Adv Appl Math 32(4):681–697
Katsevich A (2004b) On two versions of a 3-pi algorithm for spiral CT. Phys Med Biol 49(11):2129–2143
Kemmling A et al (2010) Dual energy bone subtraction in computed tomography angiography of extracranial-intracranial bypass: feasibility and limitations. Eur Radiol 21(4):750–756
Konstas AA et al (2009) Theoretic basis and technical implementations of CT perfusion in acute ischemic stroke, part 1: theoretic basis. Am J Neuroradiol 30(4):662–668
Kruger RA et al (1977) Relative properties of tomography, K-edge imaging, and K-edge tomography. Med Phys 4(3):244–249
Kudo K et al (2010) Differences in CT perfusion maps generated by different commercial software: quantitative analysis by using identical source data of acute stroke patients. Radiology 254(1):200–209
Lehmann LA et al (1981) Generalized image combinations in dual KVP digital radiography. Med Phys 8(5):659–667
Lell MM et al (2007) Bone-subtraction CT angiography: evaluation of two different fully automated image-registration procedures for interscan motion compensation. Am J Neuroradiol 28(7):1362–1368
Maass C et al (2009) Image-based dual energy CT using optimized precorrection functions: a practical new approach of material decomposition in image domain. Med Phys 36(8):3818–3829
McCollough CH et al (2006) CT dose reduction and dose management tools: overview of available options. Radiographics 26(2):503–512
McCollough CH et al (2009) Strategies for reducing radiation dose in CT. Radiol Clin North Am 47(1):27–40
Mettler FA et al (2008) Medical radiation exposure in the U.S. in 2006: preliminary results. Health Phys 95(5):502–507
Natterer F (1986) The mathematics of computerized tomography. Wiley, Chichester, New York
Nuzzo S et al (2002) Quantification of the degree of mineralization of bone in three dimension using synchrotron radiation microtomography. Med Phys 19(11):2672–2681
Peyrin F (2009) Investigation of bone with synchrotron radiation imaging: from micro to nano. Osteoporos Int 20(6):1057–1063
Peyrin FC (1985) The generalized back projection theorem for cone beam reconstruction. IEEE Trans Nucl Sci 32(4):1512–1519
Prakash P et al (2010) Radiation dose reduction with chest computed tomography using adaptive statistical iterative reconstruction technique: initial experience. J Comput Assist Tomogr 34(1): 40–45
Radon J (1917) Über die Bestimmung von Funktionen durch ihre Integralwerte längs gewisser Mannigfaltigkeiten (English translation: RADON J.: On determination of functions from their integral values along certain manifolds. IEEE Trans Med imaging 1986 MI, 5(4):170–176) Ber Verh Sächs Akad Wiss Leipzig, Math Phys Kl 69:262–277
Ramachandran GN, Lakshminarayanan AV (1971) Three-dimensional reconstruction from radiographs and electron micrographs: application of convolutions instead of Fourier transforms. Proc Natl Acad Sci USA 68(9):2236–2240
Riederer SJ, Mistretta CA (1977) Selective iodine imaging using K-edge energies in computerized X-ray tomography. Med Phys 4(6): 474–481
Ritman EL et al (1980) Physics and technical considerations in the design of the DSR: a high temporal resolution volume scanner. Am J Roentgenol 134(2):369–374
Roberts HC et al (2001) Multisection dynamic CT perfusion for acute cerebral ischemia: the "toggling-table" technique. Am J Neuroradiol 22(6):1077–1080
Rüegsegger P et al (1996) A microtomographic system for the nondestructive evaluation of bone architecture. Calcif Tissue Int 58(1):24–29
Ruzsics B et al (2009) Comparison of dual-energy computed tomography of the heart with single photon emission computed tomography for assessment of coronary artery stenosis and of the myocardial blood supply. Am J Cardiol 104(3):318–326
Saint-Félix D et al (1994) In vivo evaluation of a new system for 3D computerized angiography. Phys Med Biol 39(3):583–595

Salome M et al (1999) A synchrotron radiation microtomography system for the analysis of trabecular bone samples. Med Phys 26(10): 2194–2204

Salomon EJ et al (2009) Dynamic CT angiography and CT perfusion employing a 320-detector row CT: protocol and current clinical applications. Klin Neuroradiol 19(3):187–196

Shepp LA, Logan BF (1974) The Fourier reconstruction of a head section. IEEE Trans Nucl Sci NS-21:21–34

Söderberg M, Gunnarsson M (2010) Automatic exposure control in computed tomography–an evaluation of systems from different manufacturers. Acta Radiologica (Stockholm, Sweden: 1987) 51(6):625–634

Stanton CL et al (2010) Normal myocardial perfusion on 64-detector resting cardiac CT. J Cardiovasc Comput Tomogr 5(1):52–60

The International Commission on Radiological Protection (2007) Radiation protection in medicine. ICRP Publication 105. Ann ICRP 37(6):1–63

Thieme SF, Hoegl S et al (2010a) Pulmonary ventilation and perfusion imaging with dual-energy CT. Eur Radiol 20(12):2882–2889

Thieme SF, Johnson TR et al (2010b) Dual-energy lung perfusion computed tomography: a novel pulmonary functional imaging method. Semin Ultrasound CT MR 31(4):301–308

Thomas C et al (2010) Automatic lumen segmentation in calcified plaques: dual-energy CT versus standard reconstructions in comparison with digital subtraction angiography. Am J Roentgenol 194(6):1590–1595

Tuy HK (1983) An inversion formula for cone-beam reconstruction. SIAM J Appl Math 43:546–552

UNSCEAR (2008) Sources and effects of ionizing radiation, vol 1, Annex A Medical radiation exposures. United Nations, New York

Valton S et al (2006) Analysis of cone-beam artifacts in off-centered circular CT for four reconstruction methods. Int J Biomed Imaging Article ID 80421, 8 p

Vetter JR et al (1986) Evaluation of a prototype dual-energy computed tomographic apparatus. II. Determination of vertebral bone mineral content. Med Phys 13(3):340–343

Wintermark M et al (2001) Quantitative assessment of regional cerebral blood flows by perfusion CT studies at low injection rates: a critical review of the underlying theoretical models. Eur Radiol 11(7):1220–1230

Wintermark M et al (2005) Comparative overview of brain perfusion imaging techniques. Stroke 36(9):e83–e99

Zerhouni EA et al (1982) Factors influencing quantitative CT measurements of solitary pulmonary nodules. J Comput Assist Tomogr 6(6): 1075–1087

Zhang L et al (2010) Automatic bone removal dual-energy CT angiography for the evaluation of intracranial aneurysms. J Comput Assist Tomogr 34(6):816–824

Zilberman DE et al (2010) In vivo determination of urinary stone composition using dual energy computerized tomography with advanced post-acquisition processing. J Urol 184(6):2354–2359

37 SPECT Imaging: Basics and New Trends

Brian F. Hutton
University College London & UCLH NHS Trust, London, UK

1 Introduction . . . 918

2 The Anger Gamma Camera: Design and Performance . . . 918
2.1 System Components . . . 918
2.2 Detector Characteristics . . . 920
2.3 Collimator Design . . . 921
2.4 Performance Parameters . . . 922
2.4.1 Spatial Resolution . . . 922
2.4.2 Energy Resolution . . . 922
2.4.3 Sensitivity . . . 922
2.4.4 Dead-Time and Count-Rate Capability . . . 923
2.4.5 Uniformity . . . 923

3 Conventional SPECT System Design and Performance . . . 923
3.1 Basic System Design . . . 923
3.2 SPECT Performance . . . 924

4 Factors Affecting SPECT Quantification . . . 925
4.1 Instrument Effects: Resolution and Noise . . . 925
4.2 Physical Effects: Attenuation and Scatter . . . 926
4.3 Changes of Observed Activity Distribution in Time: Motion and Tracer Kinetics . . . 927

5 New Trends in SPECT . . . 927
5.1 Novel Collimators . . . 927
5.2 Organ-Specific Systems . . . 929
5.3 Dual-Modality SPECT . . . 930

6 Conclusions . . . 931

7 Cross-References . . . 932

References . . . 932

C. Grupen, I. Buvat (eds.), *Handbook of Particle Detection and Imaging*,
DOI 10.1007/978-3-642-13271-1_37, © Springer-Verlag Berlin Heidelberg 2012

Abstract: Single Photon Emission Computed Tomography (SPECT) is widely used as a means of imaging the distribution of administered radiotracers that have single-photon emission. The most widely used SPECT systems are based on the Anger gamma camera, usually involving dual detectors that rotate around the patient. Several factors affect the quality of SPECT images (e.g., resolution and noise) and the ability to perform absolute quantification (e.g., attenuation, scatter, motion, and resolution). There is a trend to introduce dual-modality systems and organ-specific systems, both developments that enhance diagnostic capability.

1 Introduction

As mentioned in the introductory chapter on medical imaging (see ❯ Chap. 35, "Radiation-Based Medical Imaging Techniques: An Overview") Single Photon Emission Computed Tomography (SPECT) provides images of the 3D distribution of an administered radiotracer, reflecting functional information depending on the fate of the radioactive component. As the name suggests, SPECT involves tomographic imaging based on single-photon-emitting radionuclides as opposed to Positron Emission Tomography (PET), which involves detection of dual photons that arise from positron annihilation (see ❯ Chap. 38, "PET Imaging: Basics and New Trends"). SPECT is in wide spread use in clinical Nuclear Medicine as it has become a standard feature on the Anger camera systems that are in general used for planar imaging, most commonly based on dual-detector systems. The flexibility of system use has, to a large extent, dominated the development of SPECT designs, with organ-specific designs historically proving unpopular, although more recently finding renewed interest. This chapter introduces the Anger gamma camera as the basic imaging component of SPECT systems; SPECT performance is largely determined by the gamma camera characteristics. System design of conventional SPECT and the factors that affect SPECT performance are discussed. Finally, an overview of new trends in SPECT design is included. The content relates closely to the coverage in other chapters and some overlap is inevitable.

2 The Anger Gamma Camera: Design and Performance

2.1 System Components

The gamma camera initially designed by Hal Anger (1958) continues to be the most widely used instrument in nuclear medicine, and although the performance and robustness have improved since its introduction, the basic design concept remains essentially the same. The basic components of the Anger gamma camera are illustrated in ❯ *Fig. 1*. The conventional camera consists of a single large scintillation detector using thallium-doped sodium iodide (typically 400×500 mm although dimensions can be smaller) and most commonly 9.5 mm thick (chosen to optimize stopping power for the 140 keV photons of technetium-99m, the most commonly used radionuclide). Thicker crystals can be purchased where imaging higher-energy radionuclides is of particular interest. Photons interact with the scintillator producing light that is detected by an array of photo-multiplier tubes (PMTs) optically coupled to the scintillator

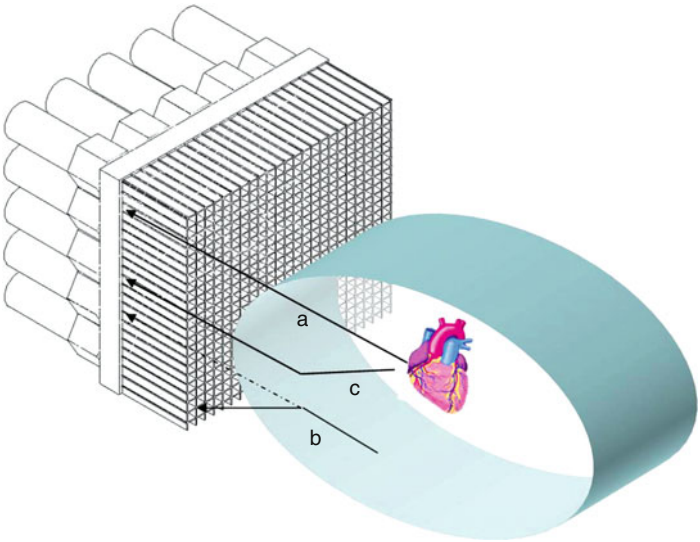


Fig. 1

Schematic of standard Anger gamma camera illustrating the collimator, scintillation detector, and an array of photo-multiplier tubes. The measured output is used to determine both location and energy of the detected photons. Also illustrated are paths for (a) a photon that is detected without any photon interaction in tissue, (b) a photon that is Compton scattered so as to be deflected and not detected (effectively attenuated), and (c) a photon that would not normally be detected but is deflected via Compton interaction so as to reach the detector

(historically using a light guide). The PMTs convert the light detected from a single photon interaction to an electrical signal, which is amplified via a set of dynodes internal to the PMT. The sum of the electrical signals is proportional to the energy deposited in the scintillator; the distribution of light can be decoded to provide an estimate of the location in the scintillator where the photon interacted. As a result, each individual photon interaction is decoded so as to provide an x/y location on the detector and the energy for that photon. Image formation historically involved the use of the x/y signal to deflect a cathode ray tube so as to expose a photographic film for a short time per detected photon; the final image was then acquired via integration of light for the period of acquisition. This process of image formation is now achieved digitally; an image is formed by simply adding a count at an image pixel location corresponding to the x/y location on the camera. On completion of acquisition each image pixel will contain a total number of counts recorded within the pixel area; the final image is formed by transforming the recorded count to a gray shade or color, so that the displayed pixel is perceived as having intensity proportional to the recorded counts. Note that the integrated light signal is proportional to the total energy absorbed in the detector and is indicative of the energy of the photon that is detected. Knowledge of this energy allows some level of discrimination against photons that have undergone Compton scatter in the patient, since these photons will have reduced energy.

It should be noted that image quality is largely determined by two factors: spatial resolution and sensitivity. Spatial resolution determines the ability to distinguish detail and to discriminate

between closely located points of activity; sensitivity determines the number of detected photons (assuming acquisition for a finite tolerable time), which directly determines the observed statistical noise in the resultant image. The statistical noise observed in acquired projections is a direct result of the uncertainty in pixel counts due to the random decay of radiation and so is Poisson distributed. This should be distinguished from electronic noise that is reflected in the uncertainty in both location and energy of the detected photons. (Note that even while accounting for the electronic noise, noise remains Poissonian in the projections). The number of counts can be increased by acquiring for a longer time (but in practice this is limited, otherwise patient motion would be problematic) or by increasing the amount of administered activity (but this is limited by the radiation dose that the subject receives).

An essential component of the system is the collimator, without which the origin of detected photons could not be determined. The collimator limits detection to those photons traveling (ideally) at right angles to the detector; the point of detection therefore provides an estimate of the point of origin of the photon but consequently most emitted photons remain undetected (*Fig. 1*). The collimator is usually fabricated from lead (although tungsten can also be used) and in the case of a conventional parallel hole collimator, the collimator consists of a "honeycomb," whose dimensions determine the collimator properties. In theory, extremely narrow, long holes could be used so as to provide optimal resolution, however this would limit the number of detectable counts; as a result a compromise is made in hole dimensions so that the system sensitivity is acceptable. In practice the sensitivity becomes the limiting factor in performance, since collimator hole size cannot be infinitely small and hence spatial resolution is limited.

There is uncertainty in both the position and energy measured, which results in the detector having a finite intrinsic spatial resolution and energy resolution. Typically an energy window is selected around the photo peak energy of the radionuclide being imaged so as to limit the photons detected to those within the range of energies determined by the energy resolution (e.g., window width twice the energy resolution); this eliminates lower-energy photons that most likely have undergone Compton scatter prior to detection (and whose origin is therefore uncertain). The spatial intrinsic resolution contributes to the overall system resolution in combination with the geometric resolution of the collimator (the latter being usually the dominant factor).

2.2 Detector Characteristics

Sodium iodide (doped with thallium) is in many respects an ideal scintillator for use in gamma cameras, being particularly well optimized for detection of 140 keV photons. A table of properties is provided (*Table 1*). NaI(Tl) has a particularly good light output compared to most

Table 1

Properties of NaI(Tl) scintillation detectors

Density (g/cm^3)	3.67	Good stopping power
Effective atomic number	50	Good absorber
Light decay Time (ns)	230	Reasonably fast
Photon yield (per keV)	38	Higher than most scintillators
Refractive index	1.85	
Peak emission (nm)	415	Well matched to PM tube
Fragile, hygroscopic	–	Easily damaged

scintillators, with reasonably fast decay time and high transparency. Unfortunately, the material is hygroscopic (absorbs water) and has to be hermetically sealed otherwise crystal deterioration would occur. Fortunately large crystals are reasonably robust and easy to grow so that a single large crystal can be utilized. The material is reasonably dense so that crystal thickness does not need to be excessive (which would degrade resolution). The refractive index is higher than that of glass so that optical coupling with PMTs can be problematic and requires specially designed coupling grease.

PMTs have evolved slowly with a number of fairly recent innovations gradually being introduced; however, the conventional PMT is still largely used in commercial systems. The main advantage of PMTs is the high gain (~10^6) and relatively fast response time. The quantum efficiency is relatively low (~25%) and they do suffer from being sensitive to magnetic fields (even the earth's field can affect performance at different orientation without suitable shielding). Typically gamma cameras utilize 30–100 PMTs.

The overall performance of the detector results from variations in the PMT gain (also coupling) and the geometry of the PMTs. As a result, an unprocessed image acquired from a gamma camera has significant defects. Current systems involve mainly digital processing in order to maintain balance of PMT gains so as to provide well-aligned energy spectra from each PMT, linearity correction so as to adjust for spatial distortion in positioning events, and uniformity correction to correct for any further regional variations in sensitivity. A uniformity correction matrix is reacquired periodically to ensure that any drift in performance is corrected; maintaining high-quality uniformity is particularly important for SPECT where uniformity defects can lead to observable artifacts (rings).

2.3 Collimator Design

As pointed out above the collimator is an essential component that effectively limits the overall performance of the imaging system. Collimator dimensions can be chosen to provide some compromise between resolution and sensitivity, and hence a range of collimators is usually available to enable high-resolution or high-sensitivity acquisition. Since collimator septa (the lead separating holes) must be designed to reduce likelihood of penetration these need to be thicker for higher-energy radionuclides, acting as a further constraint to performance. *Table 2* provides some details of typical parallel hole collimators as a guide to possible design. The shape of the hole is most commonly hexagonal in order to provide a symmetric and efficient geometry.

Although parallel hole collimators are most commonly used, there are a wide range of other collimator designs, some of which are specifically designed for use with SPECT. *Figure 5* illustrates some possible designs. These alternative designs aim to improve the balance between resolution and sensitivity, mainly by providing more optimal use of the detector via magnification, although in some recent designs minification is used in combination with new high intrinsic resolution detectors so as to improve sensitivity. There is continuing research into alternative designs, some of which are discussed in the sections below. For example, the pinhole collimator, essentially identical to the pinhole optical camera, was introduced at an early date as a means of providing magnification with potentially improved resolution and sensitivity for small objects; it originally found its clinical role in thyroid imaging. More recently multipinhole collimators have become important for their application in preclinical SPECT systems achieving resolution of <0.4 mm (see Chap. 45, "High-Resolution and Animal Imaging Instrumentation and Techniques").

Table 2

Typical parallel hole collimator specifications

Collimator type	Hole diameter (mm)	Hole length (mm)	Septal thickness (mm)	System resolution at 10 cm (mm)	Relative sensitivity[a]
Low energy high resolution	1.5	35	0.2	7.4	1.0
Low energy general purpose	1.9	35	0.2	9.0	1.69
Medium energy	3.0	58	1.05	9.4	0.90
High energy	4.0	66	1.8	10.7	0.71

[a]Sensitivity for 20% energy window using technetium−99m with this low-energy high-resolution collimator is 72 cps/MBq

2.4 Performance Parameters

2.4.1 Spatial Resolution

Spatial resolution is defined, as in optics, as the minimum distance between two sources (of activity) such that they can be discriminated; in practice the spatial uncertainty in measured counts results in a Gaussian spatial distribution of counts for which resolution is estimated as the Full Width at Half Maximum (FWHM). As mentioned above the spatial resolution has two main components, due to the detector (intrinsic: R_i) and the collimator (geometric: R_g); the system resolution (R_s) is given by $R_s = \sqrt{R_i^2 + R_g^2}$. Note that the intrinsic resolution for conventional gamma cameras tends to be < 4 mm; the geometric resolution for even a high-resolution collimator at 10 cm is ~7 mm degrading linearly with distance. With parallel hole collimation, reducing intrinsic resolution to 1 mm has little effect on the overall system resolution except for points very close to the detector.

2.4.2 Energy Resolution

The energy resolution is expressed as the FWHM of the distribution of energies measured from a mono-energetic source of activity (e.g., technetium-99m). It is usually expressed as a percentage of the photopeak energy, and for NaI(Tl), it is typically ~9.5% at 140 keV. As a result of the energy resolution, the energy window selected around the photopeak will also include scattered photons whose actual energy is lower than the photopeak, but whose measured energy still falls in the energy window.

2.4.3 Sensitivity

Sensitivity is measured as the recorded countrate from a known activity (alternatively, the percent of detected photons from a known activity is used). The sensitivity will depend on photon energy and detector thickness, although the thickness is chosen to at least optimize sensitivity

for technetium-99m. The typical figure for a gamma camera with a high-resolution collimator is ~75 cps/MBq (or <0.01%) reflecting the very poor sensitivity of SPECT, largely due to the collimator (and of course the geometry of detectors, typically limited to two planar detectors).

2.4.4 Dead-Time and Count-Rate Capability

Dead time refers to the finite time immediately following photon detection during which no further events can be detected. There are a number of factors that contribute to dead time; the finite time for scintillation light to decay resulting in an electrical signal of finite length, the ability of electronics to handle a rapid succession of pulses, and the time required to transfer data for digitization/display. The end result is that two photons interacting in the detector within a short time period may not be distinguished; in fact, the combined pulses may pileup with a resulting error in both recorded energy and location. The effect of dead time is to effectively reduce the maximum count-rate capability of the system. In modern systems, fast electronics that limit processing to local regions of the detector and efficient transfer of digital data result in relatively high count rates, sufficient for most practical purposes.

2.4.5 Uniformity

An important parameter is the uniformity of the system; exposure of the detector to a uniform activity should result in a constant count rate per unit area of the detector. This is normally measured either in terms of the maximum global variability in measured count rate (integral uniformity) or the maximum local variability (differential uniformity), the latter being more critical for SPECT. Uniformity is checked on a daily basis as part of routine quality control, and a high-count uniformity acquisition is performed periodically in order to adapt correction to the current performance. As a result differential uniformity can usually be maintained at an acceptably low level (<2%).

3 Conventional SPECT System Design and Performance

3.1 Basic System Design

Although there were many early attempts to design specific SPECT systems, the use of a standard Anger gamma camera fitted to a rotation gantry has proved to be the most popular system, mainly due to its flexible application for combined use as a planar or SPECT system. The earliest commercial systems used a fairly low-cost and easy-to-use balanced gantry system; however, this was later demonstrated to be insufficiently robust to ensure exact controlled rotation. Failure to rotate with well-calibrated camera position and a constant and reproducible center of rotation (COR) can seriously degrade image quality. The standard SPECT design has remained effectively unchanged since its introduction, although this now normally involves dual rectangular detectors rather than the original single circular detector, which introduced truncation in off-center planes. Three- and four-detector systems were introduced but later abolished due to

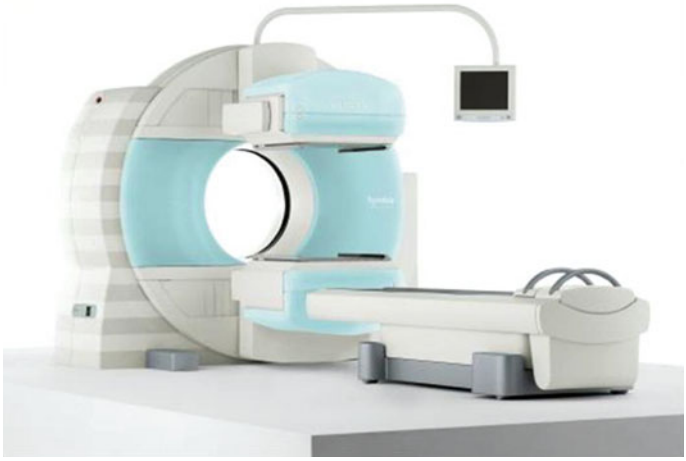


Fig. 2
A current commercial SPECT: note the robust gantry

their lack of flexibility compared to dual-head designs. Gantries are now usually sturdy mechanically sound structures that guarantee stable rotation (see *Fig. 2*). But the simple principle of rotating the planar detector(s) is maintained.

The detector(s) is normally operated in a step-and-shoot mode where the detector is rotated to a specific angle to acquire a static image and then rotated by a small angular increment to continue a new static acquisition; acquisition therefore consists of a set of planar images acquired over 360° or 180°. Each of the images can be considered as a set of axial projections at a given angle; the system therefore operates very similar to the modern multi-detector CT systems, acquiring multiple transaxial planes simultaneously. Due to the relatively poor sensitivity of the Anger camera the acquisition time at each angle is usually 10–60s; the total SPECT acquisition is therefore quite lengthy (typically 10–30 min). Image reconstruction historically involved iterative algorithms but with the development of filtered back projection for use with CT, this became the standard reconstruction approach. More recently however iterative algorithms have again become popular, especially accelerated forms of maximum likelihood reconstruction such as ordered-subsets expectation maximization (OS-EM) (Hudson and Larkin 1994), offering advantages compared to filtered back projection in many studies (see Chap. 39, "Image Reconstruction").

3.2 SPECT Performance

The SPECT system performance is determined largely by the performance of the planar detectors; essentially only the rotational stability and calibration differs from a planar system. As with the planar system the trade-off between resolution and sensitivity is dominated by the collimator and a similar range of collimators is available as used for planar imaging. There are some additional collimators that have been designed specifically for use with SPECT and these will be described later in the chapter. The resolution, as described before, varies with distance from the detector and so a practical issue is the need to rotate the detectors, keeping them as close as

possible to the patient at all angles (typically in a non-circular orbit). Avoiding collision necessitates introduction of collimator collision sensors and/or laser positioning devices. Since the rotation involves acquisition distant as well as close to any specific point in the body, the resultant reconstructed resolution can be considered an average of the resolution over the acquired angles, usually with an asymmetric point spread function, depending on position. The best indication of the expected reconstructed resolution (assuming high enough counts to avoid use of a smoothing filter) is to determine the planar resolution for the maximum used radius of rotation (typically 20–25 cm for the body, less for the brain). The reader will immediately appreciate how much this limits SPECT resolution; clearly planar imaging of near objects can be significantly better but suffers from the inability to discriminate counts originating from more distant structures. Consequently, even for high-resolution collimators the SPECT resolution is typically 12–16 mm for whole-body applications (e.g., cardiac perfusion imaging) and around 9–10 mm for brain (unless alternative collimators are used).

The main factors that are of particular concern in SPECT relate to the stability of rotation and the assumption that the observed distribution of activity remains constant throughout acquisition. Any variation in the activity distribution or its position will give rise to artifacts in the reconstruction; patient or organ motion or redistribution of activity (e.g., bladder filling) are problems that are hard to avoid but must be recognized as potential sources of error. Simple checks must be made that the camera and collimator remain orthogonal to the plane of rotation; otherwise acquired data will be inconsistent. Existence of a localized uniformity defect on the camera can lead to serious artifacts observed as ring artifacts (if 360° acquisition); careful correction of uniformity and regular quality control checks are therefore essential. Errors in the calibration of the system's electronics can result in wrong identification of the mechanical center of rotation with a degradation in resolution that can be hard to identify. Regular checks on the COR calibration are also therefore mandatory. Fortunately, current systems have much improved performance and more stable mechanics than the earlier designs.

4 Factors Affecting SPECT Quantification

There are a number of factors that influence SPECT image quality and the ability to extract quantitative information regarding the measured tracer distribution. These can be summarized as factors due to instrument performance (including resolution and noise), physical factors affecting photon emission (including attenuation and scatter), and patient-related factors (e.g., motion and changes in tracer distribution in time). The factors affecting SPECT quantification will be introduced here but corrections for the various effects will be covered in subsequent Chaps. 40, "Motion Compensation in Emission Tomography" and 41, "Quantitative Image Analysis in Tomography".

4.1 Instrument Effects: Resolution and Noise

As summarized in the gamma camera section, sensitivity and resolution are limited compared to other imaging modalities, the main constraint being the acceptable radiation dose for standard procedures. Most approaches to reducing the resultant statistical noise (whether via instrument design or image processing) result in a degradation of resolution. The result is that the image detail is reduced but also there is reduction in the observed counts (and contrast)

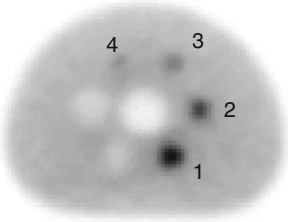


Fig. 3
Phantom with same concentration of radioactivity in four spheres of different volume. Note that, as the volume decreases (1–4), the count density appears to decrease due to the partial volume effect

for small objects. This is normally described as the "partial volume effect" that can be defined as follows: if an object partially occupies the sensitive volume of the imaging instrument the observed maximum count density (estimated activity concentration) will be reduced. In effect, what is observed is an average of the activity in the area of interest with background activity. The effect is clearly visible in *Fig. 3* where four spherical volumes with identical activity concentration appear to have quite different activities.

4.2 Physical Effects: Attenuation and Scatter

It is clear that the idealized model that all photons are emitted and travel to the point of detection without interaction (photon path (a) in *Fig. 1*) is incorrect; in reality photons interact with patient tissues mainly undergoing Compton scatter. The probability of photon interaction in tissue via the photoelectric effect is low since tissues have low atomic number, unlike detector materials where the photoelectric effect is dominant. The result is that some photons are deflected from their original path with loss of energy, with a consequential reduction in the number of photons reaching the detector (photon path (b) in *Fig. 1*); this loss of photons is referred to as attenuation. For a homogeneous material, the attenuation can be estimated by the linear attenuation coefficient μ, which is energy-dependent; for example, at 140 keV the linear attenuation coefficient of soft tissue is approximately 0.15/cm and the number of photons reaching the detector (N) compared to the number emitted in the direction of the detector (N_0) is given by $N = N_0 e^{-\mu d}$ where d is the distance traveled through tissue. Note that for a point of activity 15 cm deep in tissue the attenuation factor is ~9.5; lack of attenuation correction therefore results in a rather large quantitative error. As will be demonstrated in later chapters correction for attenuation is possible but usually demands knowledge of the distribution of attenuation coefficients most often by utilizing an x-ray computerized tomography (CT) scan (or transmission scan based on an external radionuclide).

Some photons are emitted in a direction such that they would not reach the detector but are scattered so as to be within the acceptance angle of the collimator (photon path (c) in *Fig. 1*). These scattered photons will have reduced energy but, due to the finite energy resolution of the detector some of these photons may still be acquired in the selected photopeak energy window. These scattered photons contribute additionally to the expected photon flux, but their origin is not correctly identified by their point of detection. Although the Compton interactions of the detected scattered photons are no different from those that are lost in the

attenuation process, correction of scatter is normally treated independently. The effect of scatter is to degrade contrast since the scattered photons represent a broad distribution of unwanted additional photons.

4.3 Changes of Observed Activity Distribution in Time: Motion and Tracer Kinetics

A basic assumption in tomography is that the measured projections are consistent and therefore form a set of angular views from an unchanging activity distribution. Any variation in the activity distribution with time (e.g., due to tracer redistribution) or variation in the location of the tracer distribution with time (due to movement) results in inconsistency, which leads to potential artifacts in the reconstruction. Changes in tracer distribution during acquisition are normally avoided as much as possible by scanning at a time when equilibrium of the administered radiopharmaceutical has been reached. But in some cases, consistent data cannot be guaranteed (e.g., bladder filling during acquisition). The acquisition time is typically quite long and so patient motion can be difficult to avoid. The issue may be particularly problematic if imaging the thorax where involuntary movements due to the beating heart or breathing cannot be avoided. Both cardiac and respiratory gating are standard procedures that permit data to be acquired for a series of datasets, each acquired for a preselected period during which motion is minimized. The principle of gated acquisition is illustrated in ❯ *Fig. 4*.

5 New Trends in SPECT

5.1 Novel Collimators

As mentioned earlier there are a number of alternative collimators that can be used for SPECT. These are not all new but there is a current trend to consider alternatives as a means of optimizing SPECT performance, usually aimed at specific applications. The wider use of these collimators is closely linked to iterative reconstruction algorithms that until recently were not commercially available; these are easily adapted to different geometry and can include modeling of the system resolution (see ❯ Chap. 39, "Image Reconstruction"). For example, there is some evidence that use of a collimator with higher sensitivity and poorer resolution can provide some advantage in certain applications (Lau et al. 2001; Larsson et al. 2010). A range of collimators have been described for use with SPECT (❯ *Fig. 5*).

These include fan-beam and cone-beam collimators that result in some degree of magnification for limited field of view. Any magnification results in an effective improvement in the intrinsic component of resolution as well as good sensitivity achieved by maximizing use of detector area. Half-cone-beam collimators have also been suggested specifically for brain SPECT (Li et al. 1996), with more recent suggestion to use these in combination with parallel hole collimation on a separate detector. Of recent interest have been designs involving combination of parallel hole and pinhole acquisition via a slit-slat collimator (Metzler et al. 2006). This offers the benefit of pinhole imaging while covering the complete axial field without need for complex acquisition protocols and is considered optimal for brain SPECT. Pinhole collimation was introduced at a very early date and recently has proved very effective for small-animal SPECT (see ❯ Chap. 45, "High-Resolution and Animal Imaging Instrumentation

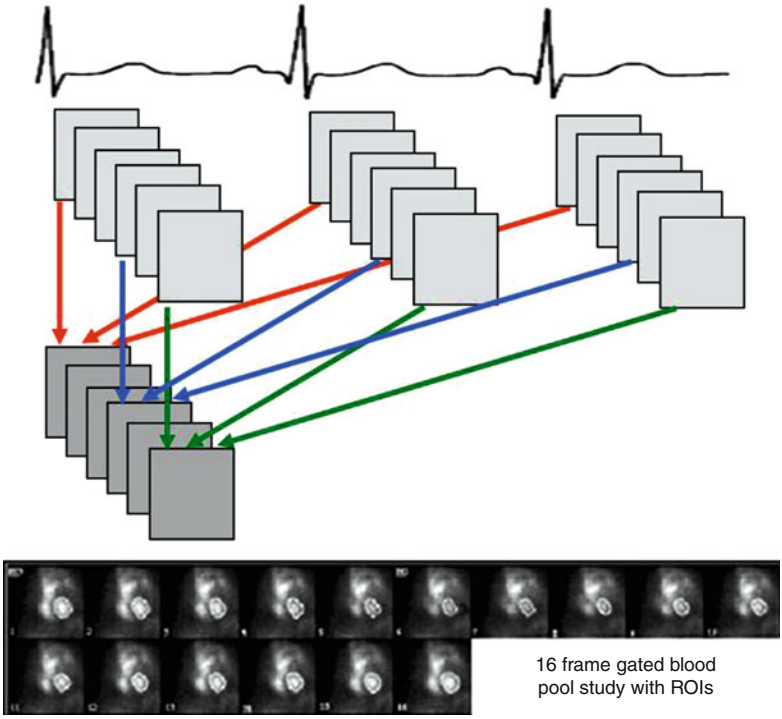


Fig. 4

Principle of using an electrocardiograph (ECG) to acquire gated cardiac data. The ECG signal is used to indicate a fixed time in the cardiac cycle; assuming a cardiac contraction with regular timing relative to the onset of the measured ECG R-wave; the acquired data can be sorted into data frames for different period of the cardiac cycle

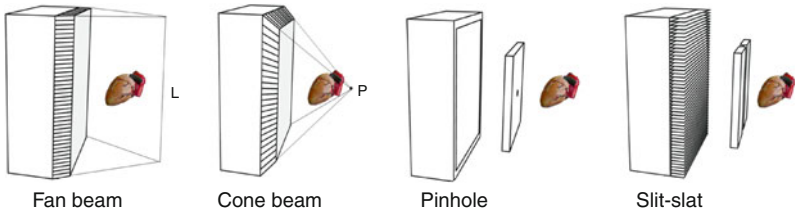


Fig. 5

Range of possible collimator designs available for use with a gamma camera

and Techniques") where the object to be imaged is small compared to the detector area (see Beekman and van der Have 2007). However, a number of authors have demonstrated that multiple pinholes in combination with high-intrinsic-resolution detectors would provide both high resolution and high sensitivity (e.g., Rogulski et al. 1993; Goorden et al. 2009). Multiple-pinhole collimators were used for cardiac work in the early eighties (LeFree et al. 1981) and have been reintroduced as the basis for a new cardiac-specific design (see next section). An issue related to use of multiple pinholes is the possible overlap of projections or multiplexing. Although it

was previously suggested that overlap provided no real advantage (Meikle et al. 2002; Vunckx et al. 2008), there is recent evidence that sensitivity can be improved with careful utilization of multiplexing (Mahmood et al. 2010).

Efforts to replace conventional collimation using solid-state detectors in so-called Compton cameras have until recently proved impractical at the energies of interest for clinical practice (e.g., 140 keV), although recent improvements in energy resolution and noise reduction using silicon detectors (lithium doped) do show some promise of significant improvement in signal-to-noise ratios suggesting a possible future role for this technology (Boston 2010, personal communication).

5.2 Organ-Specific Systems

The earliest SPECT systems were in fact designed specifically for brain, and over many years, there have been many systems designed for specific applications, especially for brain studies. As outlined earlier, however, these have not gained wide acceptance due to their limited use for more general nuclear medicine imaging, requiring a large specific referral base to justify their use. Over time, mainstream application of nuclear medicine tends to shift emphasis and organ-specific systems tend to have limited lifetime compared to more general systems. One area of nuclear medicine that has in recent years expanded to the stage of justifying dedicated systems is nuclear cardiology. The available systems include compact dual-head cameras with smaller than usual detectors, specific collimators that permit either multi-view acquisition (e.g., Xu et al. 2007) or non-truncated acquisition with the central part of the collimator providing magnification (Siemens, IQ-SPECT) and systems with novel designs that are specifically optimized for cardiac acquisition. Central to these new developments has been the introduction of solid-state detectors such as cadmium zinc telluride (CZT), which produce an electrical signal directly from the photon interaction (see ❯ Chaps. 16, "Semiconductor Counters" and ❯ 21, "New Solid State Detectors"). These detectors have the advantages of being compact with much improved energy resolution compared to the conventional sodium iodide (~5%) and so are well suited for simultaneous acquisition of multiple radionuclides. Two of these novel systems are described below.

The D-SPECT system (Spectrum Dynamics) uses a set of nine CZT detectors, each of which rotates on its own axis to acquire data from a preprogrammed region of interest that includes the heart (Gambhir et al. 2009; Erlandsson et al. 2009) (❯ *Fig. 6*). The system utilizes wide-beam tungsten collimators which, combined with the region-centric acquisition, results in significant gain in sensitivity compared to conventional dual-head SPECT. The gain in sensitivity can be used to reduce scan time, reduce activity or improve image quality (e.g., during fast dynamic acquisition). In practice, acquisition time is reduced by a factor of 3–4 compared to conventional systems, a significant jump in performance. GE Healthcare has also introduced a CZT system with similar gains in sensitivity using a non-rotating multi-pinhole design (❯ *Fig. 6*). The appeal in this system is the ability to perform dynamic cardiac acquisitions without any detector motion (Esteves et al. 2009). Various academic groups continue to investigate novel designs for organ-specific imaging, for example, work on dedicated breast imaging is being assessed for clinical use (Cutler et al. 2010).

A further design concept that has recently been introduced, with potential relevance to organ-specific imaging, is adaptive SPECT where components of the imaging system can be adjusted, ideally to optimize acquisition for application in a specific patient (Barrett et al. 2008; Clarkson et al. 2008). The aim is to acquire a scout view in order to determine the optimal

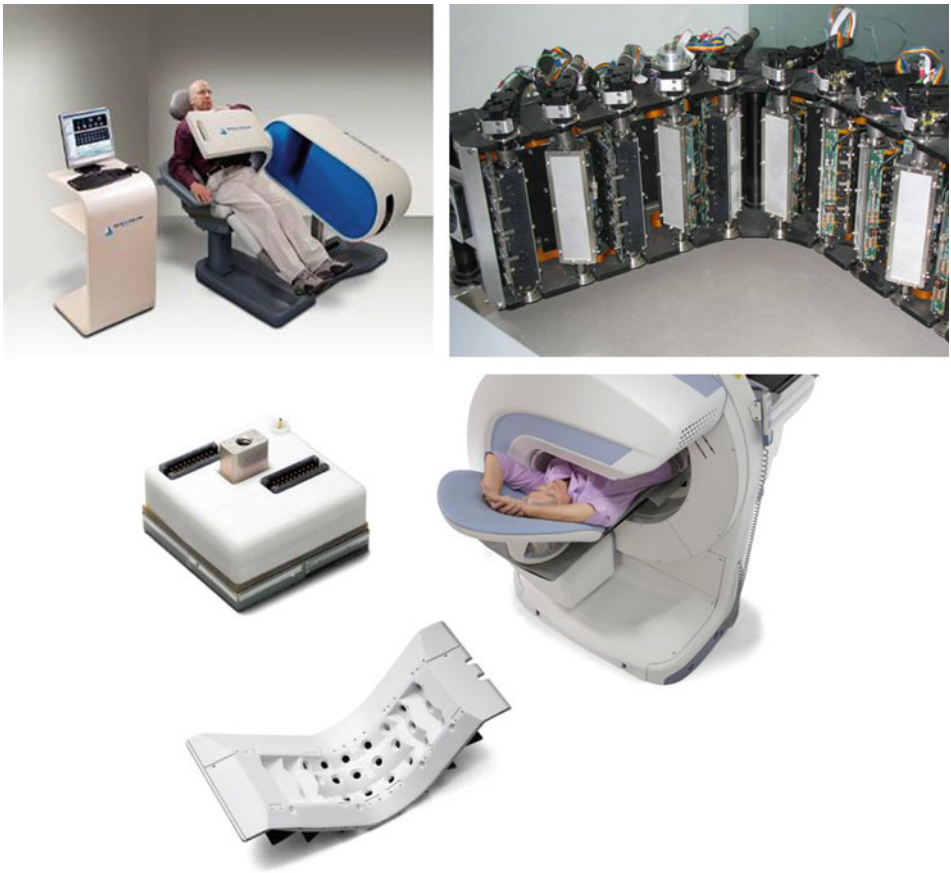


Fig. 6
Two recently introduced dedicated cardiac SPECT systems: D-SPECT (Spectrum Dynamics) (*top*) and Discovery NM530c (GE Healthcare) (*bottom*)

parameters to choose for acquisition using task-based criteria. Components of the acquisition system are then adjusted so as optimize the acquisition. A preclinical system has been designed to meet this aim (Freed et al. 2008). In essence the D-SPECT system outlined above is the first commercial system to offer an adaptive capability; a scout view is acquired so that an individual's heart region can be delineated in order to define a patient-specific volume for region-centric acquisition, thus maximizing the counts acquired from the selected region. There is clearly scope for further development of these concepts.

5.3 Dual-Modality SPECT

The development of a hybrid SPECT/CT system was pioneered by the late Bruce Hasegawa and his group (Hasegawa et al. 1989), initially motivated by the need for measured transmission data

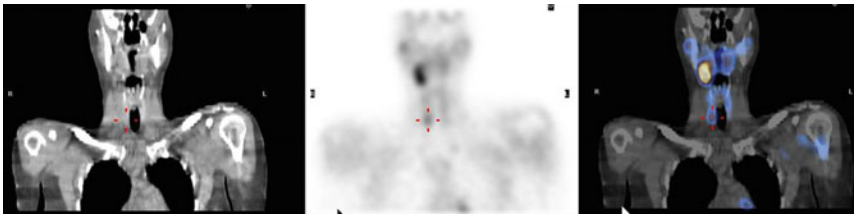


Fig. 7
Clinical SPECT/CT study with CT (*left*), SPECT (*middle*), and fused SPECT and CT data (*right*). The areas of avid uptake are easily localized using the CT anatomy

to facilitate attenuation correction for cardiac SPECT, improving on previous low-cost transmission systems that were developed for this application (Bailey 1998). The development was further motivated by the introduction of coincidence imaging for PET radionuclides on standard SPECT systems. The first commercial system (Hawkeye) was introduced by GE Healthcare in 1999, predating the first commercial PET/CT systems that were developed by CTI/Siemens (Beyer et al. 2000). The impact of PET/CT was very significant with all current commercial PET systems available only in hybrid form; in this case the combined information has proved invaluable. In the case of SPECT/CT clinical demand has been less clear with only limited applications requiring either accurate attenuation correction or direct localization; once again the generality of the system has influenced the development. The CT performance requirements vary depending on intended application; for example, attenuation correction for SPECT mainly needs to delineate areas with attenuation significantly different from soft tissue such as lung, whereas accurate localization usually requires better contrast resolution. The GE Hawkeye is an example of a low-dose (and low-cost) system where the performance at low dose outperforms conventional diagnostic systems operated at similar dose levels (Hamann et al. 2008); however, the slow rotation does render this system prone to motion artifact. Full diagnostic systems are not without problem especially as the fast acquisition can result in positional mismatch with SPECT due to difference in respiration during the two studies (McQuaid and Hutton 2008). The main suppliers now do offer higher-performance CT and clinical demand is increasing. An example of a SPECT/CT study (*Fig. 7*) demonstrates the added value in being able to more clearly localize functional abnormalities detected by the SPECT.

Following the footsteps of PET, dual-modality SPECT/MRI continues to be a work-in-progress with availability of a preclinical system (Hamamura et al. 2010) but no clinical systems available at the time of writing.

6 Conclusions

SPECT continues to be widely used in clinical Nuclear Medicine, providing flexibility to complement continuing planar investigations. The number of studies that involve SPECT as a standard protocol is increasing, as is the use of SPECT/CT dual-modality systems. The limitations in SPECT image quality continue to be determined primarily by the collimator that is an essential component of the system. There is an increasing interest in novel designs for

system geometry and alternative collimators to the conventional parallel hole collimator; the flexibility in design is easily accommodated in iterative reconstruction algorithms, which are now widely used.

7 Cross-References

- Chapter 21, "New Solid State Detectors"
- Chapter 35, "Radiation-Based Medical Imaging Techniques: An Overview"
- Chapter 38, "PET Imaging: Basics and New Trends"
- Chapter 39, "Image Reconstruction"
- Chapter 40, "Motion Compensation in Emission Tomography"
- Chapter 41, "Quantitative Image Analysis in Tomography"
- Chapter 45, "High-Resolution and Animal Imaging Instrumentation and Techniques"

References

Anger HO (1958) Scintillation camera. Rev Sci Instrum 29:27–33

Bailey DL (1998) Transmission scanning in emission tomography. Eur J Nucl Med 25:774–787

Barrett HH, Furenlid LR, Freed M, Hesterman JY, Kupinski MA, Clarkson E, Whitaker MK (2008) Adaptive SPECT. IEEE Trans Med lmag 27:775–788

Beekman FJ, van der Have F (2007) The Pinhole: gateway to ultra-high resolution three-dimensional radionuclide imaging. Eur J Nucl Med Mol Imaging 34:151–161

Beyer T, Townsend DW, Brun T, Kinahan PE, Charron M, Roddy R, Jerin J, Young J, Byars L, Nutt R (2000) A combined PET/CT scanner for clinical oncology. J Nucl Med 41:1369–79

Clarkson E, Kupinski MA, Barrett HH, Furenlid LR (2008) A task based approach to adaptive and multimodality imaging. Proc IEEE 96:500–511

Cutler SJ, Perez KL, Barnhart HX, Tornai MP (2010) Observer detection limits for a dedicated SPECT breast imaging system. Phys Med Biol 55:1903–1916

Erlandsson K, Kacperski K, van Gramberg D, Hutton BF (2009) Evaluation of the performance characteristics of D-SPECT: a novel SPECT system designed for nuclear cardiology. Phys Med Biol 54:2635–2649

Esteves FP, Raggi P, Folks RD, Keidar Z, Askew JW, Rispler S, O'Connor MK, Verdes L, Garcia EV (2009) Novel solid-state-detector dedicated cardiac camera for fast myocardial perfusion imaging: multicenter comparison with standard dual detector cameras. J Nucl Cardiol 16:927–934

Freed M, Kupinski MA, Furenlid LR, Wilson DW, Barrett HH (2008) A prototype instrument for single pinhole small-animal adaptive SPECT imaging. Med Phys 35:1912–1925

Gambhir SS, Berman DS, Ziffer J, Nagler M, Dickman D, Rousso B, Sandler M, Patton J, Hutton B, Dichterman E, Ziv O, Melman H, Zilberstein Y, Ben Haim S, Ben Haim S (2009) A novel high sensitivity rapid acquisition single photon molecular imaging camera. J Nucl Med 50:635–643

Goorden MC, Rentmeester MC, Beekman FJ (2009) Theoretical analysis of full-ring multi-pinhole brain SPECT. Phys Med Biol 54:6593–6610

Hamamura MJ, Ha S, Roeck WW, Muftuler LT, Wagenaar DJ, Meier D, Patt BE, Nalcioglu O (2010) Development of an MR-compatible SPECT system (MRSPECT) for simultaneous data acquisition. Phys Med Biol 55:1563–1575

Hamann M, Aldridge M, Dickson J, Endozo R, Lozhkin K, Hutton B (2008) Evaluation of a low-dose/slow-rotating SPECT-CT system. Phys Med Biol 53:2495–2508

Hasegawa BH, Reilly SM, Gingold EL et al (1989) Design considerations for a simultaneous emission-transmission CT scanner. Radiology 173:414

Hudson HM, Larkin RS (1994) Accelerated image reconstruction using ordered subsets of projection data. IEEE Trans Med Imaging 13:601–9

Larsson A, Jakobson S, Ljungberg M, Riklund K (2010) Dopamine D2 receptor SPECT with 123I-IBZM: evaluation of collimator and post-filtering when using model-based

compensation – a Monte Carlo study. Phys Med Biol 55:1971–1988

Lau YH, Hutton BF, Beekman FJ (2001) Choice of collimator for cardiac SPECT when resolution compensation is included in iterative reconstruction. Eur J Nucl Med 28:39–47

LeFree MT, Vogel RA, Kirch DL, Steele PP (1981) Seven-pinhole tomography – a technical description. J Nucl Med 22:48–54

Li J, Jaszczak RJ, Van Mullekom A, Scarfone C, Greer KL, Coleman RE (1996) Half-cone beam collimation for triple-camera SPECT systems. J Nucl Med 37:498–502

Mahmood ST, Erlandsson K, Cullum I, Hutton BF (2010) The potential for mixed multiplexed and non-multiplexed data to improve the reconstruction quality of a multi-slit-slat collimator SPECT system. Phys Med Biol 55:2247–2268

McQuaid SJ, Hutton BF (2008) Sources of attenuation-correction artefacts in cardiac PET/CT and SPECT/CT. Eur J Nucl Med Mol Imaging 35:1117–1123

Meikle SR, Kench P, Weisenberger AG et al (2002) A prototype coded aperture detector for small animal SPECT. IEEE Trans Nucl Sci 49:2167–2171

Metzler SD, Accorsi R, Novak JR, Ayan AS, Jaszczak RJ (2006) On-axis sensitivity and resolution of a slit-slat collimator. J Nucl Med 47:1884–1890

Rogulski MM, Barberm HB, Barrett HH, Shoemaker RL, Woolfenden JM (1993) Ultra-high-resolution brain SPECT imaging: simulation results. IEEE Trans Nucl Sci 40:1123–1129

Vunckx K, Suetens P, Nuyts J (2008) Effect of overlapping projections on reconstruction image quality in multipinhole SPECT. IEEE Trans Med Imaging 27:972–983

Xu J, Liu C, Wang Y, Frey E, Tsui BMW (2007) Quantitative rotating multisegment slant-hole SPECT mammography with attenuation and collimator-response compensation. IEEE Trans Med Imaging 26:906–916

38 PET Imaging: Basics and New Trends

Magnus Dahlbom
David Geffen School of Medicine at UCLA, University of California, Los Angeles, CA, USA

1 Introduction ***937***

2 Physics ***938***

3 Photon Interactions ***939***
3.1 Photoelectric Interactions 940
3.2 Compton Interactions 940
3.3 Attenuation Coefficients 941
3.4 Relevance to PET 942

4 Detectors ***942***
4.1 The Block Detector 944

5 Coincidence Detection ***946***

6 Event Types ***947***
6.1 Scattered Coincidences 947
6.2 Accidental Coincidences 947
6.3 Multiple Coincidences 948
6.4 Prompt Coincidences 948
6.5 Noise Equivalent Counts 948

7 Resolution Limitations ***949***

8 Data Collection, 2-D and 3-D PET ***951***

9 Data Corrections ***952***
9.1 Normalization 952
9.2 Attenuation Correction 953
9.3 Scatter Correction 957

10 System Calibration and Quantification ***958***
10.1 Partial Volume Effect 959

C. Grupen, I. Buvat (eds.), *Handbook of Particle Detection and Imaging*,
DOI 10.1007/978-3-642-13271-1_38, © Springer-Verlag Berlin Heidelberg 2012

11 Image Reconstruction 961
11.1 Filtered Backprojection 961
11.2 Iterative Reconstruction 962

12 Time-of-Flight PET 963

13 Multi-modality Imaging 965
13.1 PET-CT 965
13.2 PET-MRI 965

14 Dedicated Systems 966
14.1 Animal and Preclinical PET Systems 966
14.2 Organ-Specific PET Systems 966
14.3 Brain Imaging 967
14.4 Breast Imaging 967
14.5 Prostate Imaging 967

15 Summary 968

16 Cross-References 968

References 968

Abstract: Positron Emission Tomography or PET is a noninvasive molecular imaging method used both in research to study biology and disease, and clinically as a routine diagnostic imaging tool. In PET imaging, the subject is injected with a tracer labeled with a positron-emitting isotope and is then placed in a scanner to localize the radioactive tracer in the body. The localization of the tracer utilizes the unique decay characteristics of isotopes decaying by positron emission. In the PET scanner, a large number of scintillation detectors use coincidence detection of the annihilation radiation that is emitted as a result of the positron decay. By collecting a large number of these coincidence events, together with tomographic image reconstruction methods, the 3-D distribution of the radioactive tracer in the body can be reconstructed. Depending on the type of tracer used, the distribution will reflect a particular biological process, such as glucose metabolism when fluoro-deoxyglucose is used. PET has evolved from a relatively inefficient single-slice imaging system with relatively poor spatial resolution to an efficient, high-resolution imaging modality which can acquire a whole-body scan in a few minutes. This chapter will describe the basic physics and instrumentation used in PET. The various corrections that are necessary to apply to the acquired data in order to produce quantitative images are also described. Finally, some of the latest trends in instrumentation development are also discussed.

1 Introduction

Positron Emission Tomography or PET is a well-established research and diagnostic imaging modality, which is used to image the in vivo 3-D distribution of an injected radiotracer. To localize the tracer, PET uses the unique decay characteristics of radionuclides that decay through positron emission. In a typical PET imaging session, a small amount of radiotracer (e.g., a tracer molecule labeled with a positron-emitting radionuclide such as fluoro-deoxy glucose) is injected into the subject. The subject is then placed into a PET scanner, which is designed to detect the high-energetic photons that are emitted as a result of the positron decay. By using an image reconstruction, a 3-D image volume is created which represents the distribution of the radiotracer in the subject. Depending on the characteristics of the radiotracer, the resulting activity distribution reflects the physiological uptake or metabolism of the tracer. Therefore, PET is usually referred to as a functional or molecular imaging method.

PET was for many years used in research to study the function of single organs, primarily the heart and the brain. Although PET is still widely used in research, its main application for the last 10–15 years has been in clinical imaging where it has become an invaluable tool in oncology for diagnosis and staging (Czernin and Phelps 2002; Weber et al. 2008). PET is also used clinically in various neurological diseases such as degenerative diseases (e.g., Alzheimer's and Parkinson's disease (Silverman et al. 2001)) and the evaluation of patients with epilepsy (Lin et al. 2007). With the recent development of long-lived tracers for cardiac imaging, PET may also become used more widespread clinically in the evaluation of cardiac disease (Schindler et al. 2010).

2 Physics

PET imaging relies on the unique emission properties of isotopes decaying through positron decay. Isotopes with an excess of protons can either decay through electron capture or positron decay. One example is ^{18}F which decays to ^{18}O by electron capture 3% of the time and by positron decay the remaining 97%. In the case of positron decay, the decay can be seen as the conversion of a proton (p) into a neutron (n), a positron (β^+), and a neutrino (ν):

$$^{18}\mathrm{F} \rightarrow {}^{18}\mathrm{O} + \beta^+ + \nu_e, \tag{1}$$

$$p \rightarrow n + \beta^+ + \nu_e. \tag{2}$$

The positron is the antiparticle to the electron and has the same mass, but with a different electric charge.

The total amount of energy released in the decay E_{max} is shared between the daughter nucleus, the positron, and the neutrino. Since this energy is shared between the three particles, the positron can be emitted with an energy between zero and the total amount of energy available. E_{max} is determined by differences in atomic masses of the parent and daughter atoms and any possible excited states of the daughter. This energy is therefore dependent on the isotope and the particular decay path. The energy distribution of the positrons from a specific decay path is not uniform but varies with an almost zero probability at zero and E_{max} energies and is peaked at approximately 1/3 E_{max}.

After being emitted in tissue, the energetic positron will rapidly lose its energy through inelastic collisions with atomic electrons. It will eventually have lost most of its energy and will reach thermal energies. At this point, it will be attracted to an electron because of its positive charge and form a hydrogen-like state known as positronium. The positronium is very unstable and the electron–positron will annihilate to produce two photons, each with an energy of 511 keV, which is the rest-mass energy of the electron and the positron (Evans 1955):

$$e^+ + e^- \rightarrow 2h\nu. \tag{3}$$

If the positron–electron pair is at rest at the time of annihilation (i.e., having zero momentum), the two photons will be emitted 180° apart. However, if the momentum is nonzero at the time of annihilation, then there will be a slight deviation from 180° between the two photons.

The simultaneous emission of the two photons with a defined energy (511 keV) in opposite directions is what is utilized in a PET scanner to localize the isotope. The orientation of the emission of the two annihilation photons can be determined using coincidence detection. By placing a pair of detectors on each side of a positron-emitting source, and using coincidence detection (i.e., requiring that both detectors simultaneously detect a 511 keV photon), the annihilation is assumed to have occurred anywhere along the line connecting the two detectors (*Fig. 1*). The line connecting the two detectors is usually referred to as a line of response or LOR. By placing many of these detector pairs around the object to be imaged and collecting a large number of coincidence events along many different LORs, and using tomographic image reconstruction techniques, the 3-D activity distribution within the object can be determined (Phelps et al. 1975).

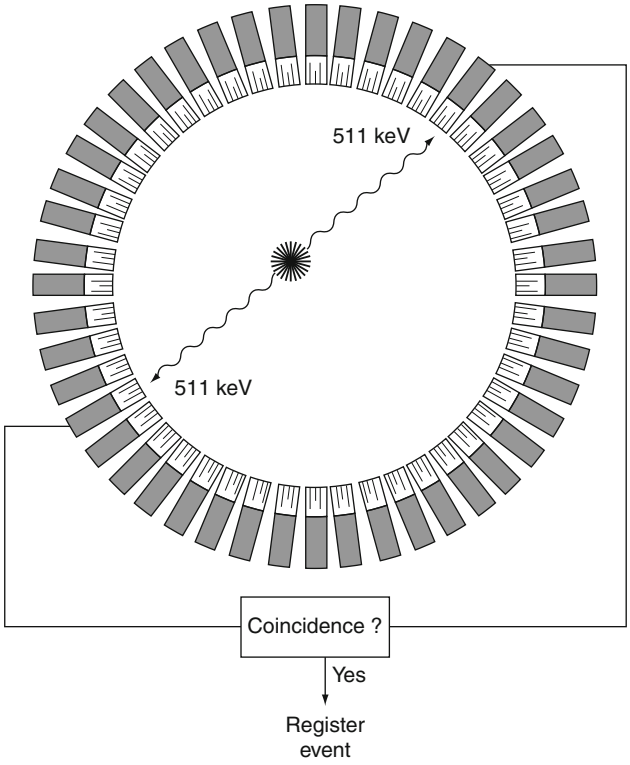


Fig. 1

Principle of PET. In a PET scanner, a large number of detectors are arranged around the object to be imaged. When a pair of detectors simultaneously registers a pair of 511 keV annihilation photons, the locations of the radioactive decay are assumed to have occurred somewhere along the line connecting the two detectors. By collecting a large number of coincidence events between the detector pairs in the system, the distribution of the activity can be reconstructed

3 Photon Interactions

Understanding photon interactions in PET imaging is important for several reasons. In order to detect the 511 keV photons, the detector systems used rely on photon interactions with the detector material to absorb the photon energy and convert this to some other form of energy such as a burst of visible light photons in a scintillation detector. Photon interactions are also a source of image degradation in PET such as the production and detection of scattered photons. The detection of scattered photons will lower image contrast and the loss of primary photons (i.e., attenuation) will produce image artifacts if not corrected for. At 511 keV, the two dominant photon interactions are the photoelectric and Compton interactions.

3.1 Photoelectric Interactions

In the photoelectric interaction, the 511-keV photon will interact with an orbital electron and transfer all of its energy ($h\nu$), minus the binding energy (E_b) as kinetic energy ($E_{p.e.}$) to the electron (Evans 1955):

$$E_{p.e.} \rightarrow h\nu - E_b. \tag{4}$$

This interaction is most likely to occur with electrons in the inner shells of the atom (i.e., K and L shells). The ejection of the electron from the atomic shell will typically result in the emission of the characteristic X-rays as the vacancy is filled by an electron from one of the outer shells. The atom may instead of emitting a characteristic X-ray transfer this energy to a second electron within the same atom. These electrons are referred to as Auger electrons. The energy of the characteristic X-ray and the Auger electrons depends on the electron-shell energy structure of the atom which in turn is dependent on the atomic number.

For soft tissue, photoelectric interaction is only significant at energies below 100 keV. At 511 keV, the photoelectric interaction is the dominant interaction type in materials with a Z (or effective Z) greater than 79. One of the desirable characteristics of a detector material for PET is that the primary interaction type is photoelectric interaction. This will ensure that the photon energy is absorbed in a small detector volume. This ideal detector material for PET should therefore have a high effective Z.

3.2 Compton Interactions

The Compton interaction is an interaction between a photon and free or loosely bound orbital electron. In this interaction, only a portion of the photon energy is transferred as kinetic energy to the electron. The remaining energy appears as a secondary scattered photon, which is emitted in a different direction from the primary photon. Based on the conservation laws of energy and momentum, it can be shown that there is definite relationship between the energies of the primary ($h\nu$) and scattered ($h\nu'$) photons, and the scattering angle (θ) between them (Evans 1955):

$$h\nu' = \frac{h\nu}{1 + \frac{h\nu}{m_0c^2}(1 - \cos\theta)}, \tag{5}$$

where m_0c^2 is the rest-mass energy of the electron (i.e., 511 keV). The energy transferred to the electron is:

$$E_k = h\nu - h\nu' = h\nu \frac{\frac{h\nu}{m_0c^2}(1 - \cos\theta)}{1 + \frac{h\nu}{m_0c^2}(1 - \cos\theta)}. \tag{6}$$

From these equations, it can be seen that the amount of energy transfer to the electron or energy loss of the photon increases with increased scattering angle. It should be noted that for 511-keV photons, a relatively large scattering angle results in a relatively small energy loss. For instance, at a scattering angle of 30°, only 60 keV is lost. The maximum energy transfer (or maximum energy loss) occurs at a scattering angle of 180°. At 511 keV, the maximum energy transfer to the electron is 341 keV and the energy of the scattered photon is 170 keV.

Equations 5 and 6 only describe the relationship between the energy transfer and the scattering angle. The probability distribution of Compton scattering angles depends both on the photon energy and the scattering angle and is described by the Klein–Nishina equation (Evans 1955):

$$\frac{d\sigma}{d\Omega} = Zr_0^2\left(\frac{1}{1+\alpha(1-\cos\theta)}\right)^2\left(\frac{1+\cos^2\theta}{2}\right)\left(\frac{\alpha^2(1-\cos\theta)^2}{(1+\cos^2\theta)(1+\alpha(1-\cos\theta))}\right), \quad (7)$$

where $d\sigma/d\Omega$ is the differential scattering cross section within a solid angle $d\Omega$, and r_0 is the classical electron radius and α is $h\nu/m_0c^2$. At 511 keV, it can be shown that the scattering cross section is peaked in the forward direction (i.e., small scattering angles are favored) and that one third of all events will scatter within 40°, which corresponds to an energy loss of 100 keV or less.

3.3 Attenuation Coefficients

The probability for any interaction to occur is given by the attenuation coefficients for both the photoelectric and Compton interactions and has a dependency on the photon energy. The Compton interaction is directly proportional to the electron density of the interacting material and has an energy dependency described by the Klein–Nishina equation (Eq. 7). The cross section for photoelectric interaction varies strongly with both energy and atomic number. For photon energies near 511 keV, the cross section τ varies approximately as:

$$\tau \propto \frac{Z^3}{E^3}. \quad (8)$$

At energies below 100 keV, the cross section for photoelectric effect shows discontinuities near the electron binding energies for a given Z, where the cross section suddenly decreases at a photon energy slightly below the binding energy.

Photoelectric interactions are relatively insignificant in tissue at 511 keV because of the relatively low effective Z of tissue and the relatively high photon energy. Compton interaction is therefore the predominant interaction type in tissue, and as a result a significant amount of secondary scatter is produced.

The interactions of mono-energetic photons passing through an absorber of thickness x can be described by:

$$I_x = I_0\, e^{-\mu x}, \quad (9)$$

where I_0 and I_x are the number of primary photons before and after passing through the absorber of thickness x and the linear attenuation coefficient is μ. The linear attenuation coefficient μ contains all the attenuation coefficients of the individual processes such as photoelectric and Compton interaction. This equation only describes how many of the primary photons have *not* interacted after passing through the absorber. It does not give any specific information about how many photons have interacted by a specific interaction type or how many secondary scattered photons were generated in the process.

3.4 Relevance to PET

When the isotope is injected into a patient or a phantom, there is a high likelihood that the emitted 511-keV photons will undergo an interaction in the object. At 511 keV, the linear attenuation coefficient is approximately 0.095 cm^{-1}. If an absorber thickness of 15 cm is considered, only 24% of the primary photons will pass through the absorber without any interaction. The remaining 76% will have undergone some sort of interactions, which is primarily Compton interaction since this is the dominant interaction type at 511 keV in water or tissue. The interaction of photons results in an attenuation of the number of primary photons that are emitted and should have been detected. This needs to be corrected for in the reconstruction in order to get an artifact-free image. The large number of interactions will also result in a large number of secondary scattered photons. Although these scattered photons have a lower energy compared to the primary photons, the limited energy resolution of the detectors used in PET systems makes it hard to differentiate these from the primary 511-keV photons. It is therefore a relatively high probability that these scattered photons are detected.

4 Detectors

In order for a PET system to efficiently register the annihilation photons, it is necessary to use a detector material that can stop the 511-keV photons. It is also desirable that these detectors also can stop the photons in a relatively small volume, since this will determine the spatial resolution of the system. This means that the detector material should have a high effective Z since this will ensure that the photons will have a high probability of undergoing a photoelectric interaction. In addition, the detector should be fast enough to be able to separate individual annihilation events from each other at relatively high activity levels. The most efficient way to stop the 511-keV photons is the use of scintillation detectors with a relatively high atomic number.

The energetic photoelectron produced in the photoelectric interaction will produce ionizations and excitations of electrons in the crystal lattice. When some of these electrons de-excite back to the ground state, light is emitted and the amount of light emitted is proportional to the amount of energy that was deposited in the detector. In addition to the photoelectric interaction, there is also the possibility of Compton interactions, in which only a fraction of the photon energy is deposited in the detector (i.e., the energy transferred to the Compton electron). Although in theory these are good events, these interactions are typically rejected since these are indistinguishable from scattered photons originating from Compton interactions within the object that is being imaged.

Following the absorption of a photon in the scintillation detector, there will be an emission of a burst of light photons. This emission is not instantaneous, but will typically follow a function which has a sharp rise time, followed by an exponentially decaying tail with a time constant characteristic for the scintillator material (Knoll 2010). This light is then collected by a photodetector, which will convert the optical signal into an electrical signal that can be analyzed further by either analog and/or digital electronics.

The amount of light and the decay time are in addition to the absorption characteristics very important parameters for a detector to be suitable as a PET detector. The more light that is produced per absorption, the better the energy resolution, which will improve the detector's ability to separate scatter from the primary photons. The decay time will determine how fast the detector can count. In order to ensure a high signal-to-noise ratio (S/N) in the signal, the output

from the detector is integrated over a time equal to 3–4 times the decay constant of the scintillation light. Since the detector cannot process another pulse during this time interval, this will limit how many pulses the detector can process per second. Both the light output and the decay time constant also affect how fast the detector will respond following a photon interaction in the detector. The physical characteristics for different scintillators are listed in ❯ *Table 1*. These are materials that have been or are being used as scintillators in PET. To briefly summarize, the key properties of a suitable detector for PET are a fast scintillation decay time, high effective Z, high density and linear attenuation coefficient, and a high light output. The wavelength of the emitted scintillation light should also be well matched to the spectral response of the photodetector. The detector material should preferably not be hygroscopic since this eliminated the need for detector encapsulation. The material should also be relatively rugged to allow it to be cut into suitable detector dimensions.

The most common type of photodetector used in PET is the photomultiplier tube or PMT (Knoll 2010). The light photons from the scintillation detector are collected by the photocathode of the PMT. In the photocathode, about 20–30% of the incoming light photons will emit an electron that is collected by the first dynode. Since there is a potential drop between the photocathode and the first dynode, the electron will gain kinetic energy, which will be high enough to knock out secondary electron from the first dynode. These will in turn be collected by a subsequent dynode, where further secondary electrons will be produced. This multiplication effect will continue along the chain of dynodes in the PMT. The result of this is both a conversion of the light signal into an electrical signal as well as an amplification of the signal.

In addition to the PMT, there are other types of photodetectors based on different semiconductor technologies. This includes the avalanche photodiodes (APD) (Lecomte et al. 1985) and the silicon photomultipliers (SiPM) (Buzhan et al. 2003). In contrast to conventional photodiodes, both the APDs and SiPM provide an internal amplification of the signal and are very compact devices. Furthermore, both devices are, in contrast to conventional PMTs, insensitive to magnetic fields, which has allowed the construction of PET systems that can operate simultaneously within or near a MRI system. In comparison to PMTs, these devices are still fairly expensive per detector channel which has to date limited a wide spread usage as photodetectors in PET.

Table 1

Examples of scintillation detector materials and their properties at 511 keV

	NaI (Tl)	BGO	GSO	LSO	LYSO	$LaBr_3$
Density [g/ml]	3.67	7.13	6.71	7.35	7.1	5.29
$1/\mu$ [cm]	2.88	1.05	1.43	1.16	1.2	~2
Index of refraction	1.85	2.15	1.85	1.82	1.81	1.9
Hygroscopic	Yes	No	No	No	No	Yes
Rugged	No	Yes	No	Yes	Yes	Yes
Peak emission [nm]	410	480	430	420	420	380
Decay constant [ns]	230	300	60	40	41	25
Light output*	100	15	35	75	75	<100
Energy resolution#	7.8	20	8.9	<9	11	7.5

*Relative to NaI(T1)

#In % at 511 keV

4.1 The Block Detector

The first generations of PET systems used a detector technology where each individual scintillation detector was coupled to an individual PMT. This detector configuration allows an unambiguous identification of the scintillation detector in which a photon interacted since the signal from the PMT originates from a single scintillation crystal. One of the main limitations of this design is that it limits the spatial resolution that can be achieved since the smallest commercial PMT available is about 10 mm in diameter. This limits the size of the scintillation detector to be used to $10 \times 10\ mm^2$. Furthermore, this design is also very expensive since each detector requires its own processing electronics and readout. The photocathode of the small PMTs tends to be very small which results in a poor light collection from the scintillator, which compromises energy and time resolution.

Most modern PET systems use a detector design that is commonly referred to as block detectors (Casey and Nutt 1986). In a block detector, an array of detector elements is created from a large block of scintillation material (◉ *Fig. 2*), which is coupled to four PMTs. The array is generated by saw cuts into the scintillator material, where the depths of the cuts are varied

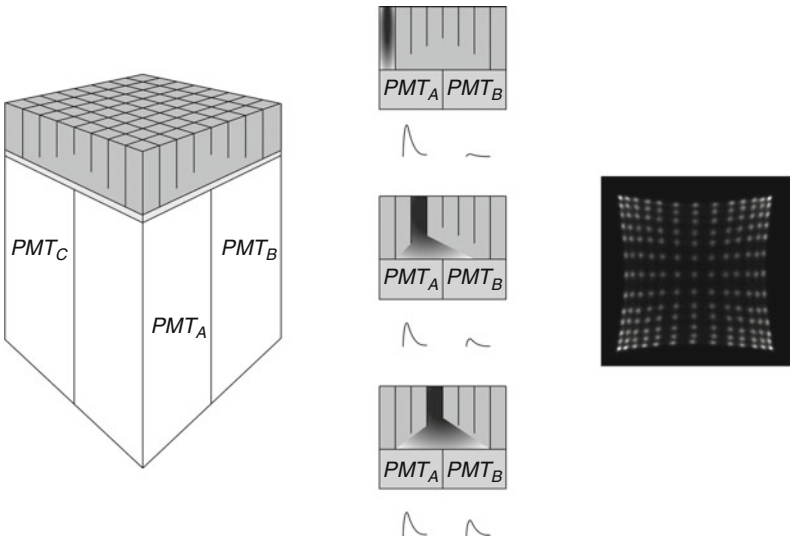


Fig. 2

Principle of the block detector. In a block detector, an array of detector elements is created from a large block of scintillation material (*left*), which is coupled to four PMTs. The array is generated by saw cuts into the scintillator material, where depths of the cuts are varied at different positions in the array. The varying depths of the saw cuts will distribute the light produced in the detector elements in such a way that each element will produce a unique combination of signal intensities in the four PMTs (*middle*). The gray scale image to the *right* shows the distribution of the X_{pos} and Y_{pos} coordinates calculated using ◉ Eq. 10, when a block with 13 by 13 detector elements is uniformly irradiated with 511-keV photons. The brighter spots in this image correspond to individual detector elements in the block detector

at different positions in the array. Alternatively, an array of detectors is glued to a light guide into which the cuts of different depths are sawn. Reflective material is added in the saw cuts to optically isolate the detector elements. The purpose of the varying depths of the saw cuts is to distribute the light from the scintillator elements in such a way that each element will produce a unique combination of signal intensities in the four PMTs. This provides a method for accurate identification of each detector element without the need of a photodetector readout on each detector element. As shown in *Fig. 2*, the cuts near the corner crystal are deep enough to guide most of the light to only one PMT, whereas a more central crystal will produce a signal of different strengths in all four tubes. To identify the detector elements, an Anger-like algorithm is used. Based on the signals from the four PMTs (PMT_A, PMT_B, PMT_C, PMT_D), a coordinate is calculated:

$$X_{\text{pos}} = \frac{PMT_A + PMT_B - PMT_C - PMT_D}{PMT_A + PMT_B + PMT_C + PMT_D},$$
$$Y_{\text{pos}} = \frac{PMT_A - PMT_B + PMT_C - PMT_D}{PMT_A + PMT_B + PMT_C + PMT_D}. \tag{10}$$

The two coordinates X_{pos} and Y_{pos} do not directly correspond to a spatial location of the photon interaction within the detector block due to the nonlinear properties of the light collection. Instead, this coordinate is used together with a look-up table, which maps the coordinate to a specific detector element in the array. The gray scale image in *Fig. 2* shows the distribution of the X_{pos} and Y_{pos} coordinates when a block with 13 by 13 detector elements is uniformly irradiated with 511-keV photons. The brighter spots in this image correspond to individual detector elements in the block detector. As can be seen, the peaks are relatively well separated from each other; however, there are some overlap in the regions between the peaks which results in some mispositioning of events. How well these peaks are separated from each other depends on how much light is collected by the photodetector. This is primarily determined by the light output of the scintillator material used, but other factors such as the size of the detector elements, number of elements in the block, and the reflective materials also play a role. The number of detector elements in a block detector varies between designs and scintillation material, but detectors with as many as 169 elements (13 × 13 array) covering an area of approximately 5 × 5 cm^2 are currently in clinical use (Lois et al. 2010). The block detector technology has allowed the construction of high-resolution systems with resolution to about 3.5 mm for whole-body systems and below 2 mm for dedicated high-resolution brain systems.

There are a number of variations on the original block detector design used in different PET systems. This includes the quadrant sharing detector where larger PMTs are used and each tube is shared among four adjacent detector blocks (Wong 1993). This has the advantage that the number of PMTs and readout channels can be reduced and this allows further cost savings. The drawback is that each PMT and readout channel has to process events from a larger number of detector elements, which limits the count-rate capabilities.

Another detector design is the curved panel detectors where each panel is made up of a large array of individual detector elements that are coupled to an array of PMTs via a light guide (Karp et al. 2003). This design is a hybrid between the block detector and the detector design used in scintillation cameras. In contrast to the readout in a scintillation camera where all PMTs contribute to the positioning of each event, in this detector design only the seven PMTs closest to the detector of interaction are read. This readout scheme improves the count-rate capabilities. The assignment of the event to an actual detector element is done using a look-up table.

5 Coincidence Detection

The registration of the two annihilation photons in PET is based on a technique referred to as coincidence detection. This technique allows the determination of direction of emission of the two annihilation photons, assuming that the photons are emitted 180° apart. *Figure 3* shows a simple coincidence circuit. In the circuit, two photon detectors are placed on opposite sides of the positron-emitting source. Both detectors are connected to power supplies, amplifiers, and pulse-height analyzers (PHA) that are adjusted to only detect 511-keV photons. When a detector is struck by a 511-keV photon, the PHA will generate a logic pulse with a width τ. The outputs from the two PHAs are fed into a coincidence circuit which will sense if there is an overlap of pulses from the two inputs. If there is an overlap of two pulses, a coincidence has been detected and the coincidence circuit will generate a logic pulse that is fed into a counter that will register the event.

Under ideal circumstances, the two PHAs should generate logic pulses with perfect overlap or alignment in time when detecting annihilation photons from the same annihilation occurring at the midpoint between the detectors. However, all scintillation detector systems have an

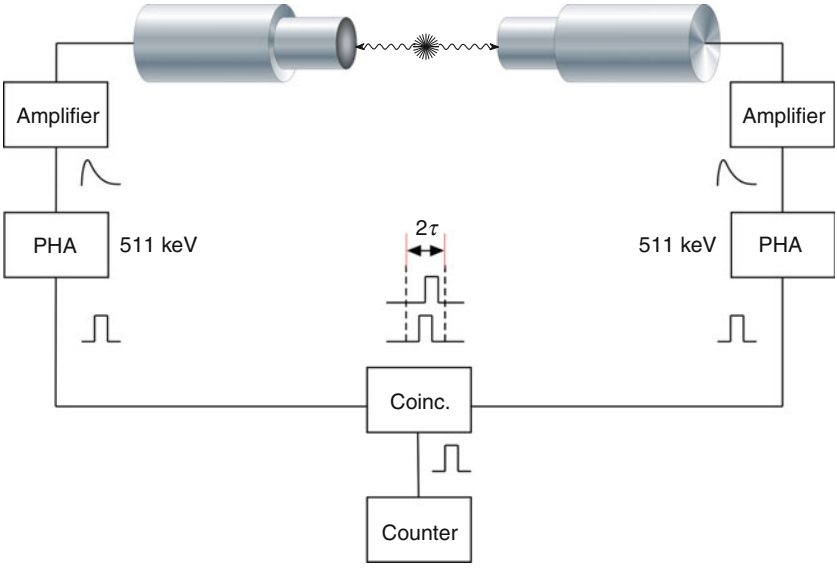


Fig. 3

Schematic of a basic coincidence circuit. The two photon detectors are connected to individual channels of pulse-processing electronics (amplifiers, pulse-height analyzers – PHAs). When a photon interacts in a detector and is identified by the PHA as a 511-keV photon (i.e., energy exceeds a lower energy threshold), a logic pulse with a width of τ is generated. The outputs from the two PHAs are fed into a coincidence circuit that will detect overlapping pulses. If there is an overlap, a coincidence has been detected and the coincidence module will generate a logic pulse that can be fed into a counter to record the event

inherent slowness or finite time resolution. There are a number of reasons for this, but one of the major contributing factors is due to the fact the emission of the scintillation light photons is a stochastic process. This results in a random time delay when the scintillation detector will respond following a photon interaction. In addition, noise in the photodetectors and the pulse processing electronics are other contributing factors to the time delay. In order to prevent loss of events due to the inherent time delay, it is necessary to give the logic pulses a finite width as shown in ❯ *Fig. 3*. Using a width of τ would ensure that all events separated by $\leq 2\tau$ will be registered. The value of τ should be at least as wide as the time resolution of the detector pair. The time resolution depends on the scintillation material and is around 5–6 ns for a BGO system and could be as low as 212 ps for an LSO system with optimized electronics (Moses and Ullisch 2006).

6 Event Types

A PET system should ideally only detect true coincidences, which are coincidences that are generated by two annihilation photons that originate from the same radioactive decay and have not interacted prior to being detected. This ensures correct localization of the event. Due to limitations in the detection system such as the finite energy and time resolution, a number of additional event types are also detected by the coincidence circuitry.

6.1 Scattered Coincidences

A scattered event is a coincidence where one or both of the two photons have interacted and generated secondary scattered photons. If the scattering angle is large enough, the energy loss might be large enough to reject the event by energy discrimination. However, even an energy threshold of 450 keV will allow the detection of photons that have scattered by as much as 30°. Thus, energy discrimination is not very effective in rejecting scatter in PET. Scatter can be reduced by the use of collimators or septa placed between the detector rings and/or increasing the detector separation (Williams et al. 1981). Although both approaches reduce the detection of scattered events, detection efficiency is significantly reduced. A certain amount of scattered events will therefore inevitably be detected and the amount is estimated using modeling and subtracted from the total number of detected events.

6.2 Accidental Coincidences

As discussed earlier, the coincidence detection circuit requires that the trigger pulses from the individual detector outputs have a finite width to ensure overlap of the pulses. As a consequence of the finite width of the trigger pulses, there is a possibility that two photons from unrelated radioactive decays will be recorded as a coincidence. These events are referred to as accidental or random coincidences. To the electronics, the accidental coincidences are indistinguishable from true coincidences and are added as a relatively uniform background that will reduce image contrast if not corrected for. It can be shown that the number of recorded accidental coincidences

N_{acc} is proportional to the product of coincidence timing window 2τ and the count rates of the individual detectors, N_1 and N_2 (Evans 1955):

$$N_{acc} = 2\tau N_1 N_2 \propto 2\tau A^2. \tag{11}$$

Since the individual detector count rates are directly proportional to the activity in the FOV (A), the accidental coincidence count rate is proportional to the square of the activity. Using Eq. 11, the number of accidental coincidences can be estimated for each coincidence pair in the system by monitoring the individual detector singles rates in addition to the coincidence rate. The number of accidental coincidences can also be estimated using a method referred to as the delayed coincidence window technique. In this method, a parallel coincidence circuit is added to the primary with the difference that a time delay beyond the time resolution of the detectors is introduced in one of the detector channels (Williams et al. 1979). This will prevent the coincidence circuit to register any true coincidences, and the only events recorded are accidental coincidences.

Both of these methods provide estimates of the accidental count rate. The singles method provides a statistically more accurate estimate compared to the delayed coincidence method, but may introduce systematic errors due to uncertainties in the coincidence timing window and differences in dead-time behavior between the singles and coincidence circuits.

It should also be noted that although the number of accidental coincidences can be estimated, the correction introduces additional noise to the net true count rate (Hoffman et al. 1981; Strother et al. 1990). It is therefore desirable to always minimize the detection of accidental coincidence, which is primarily controlled by limiting the activity in the FOV during the acquisition.

6.3 Multiple Coincidences

Multiple coincidences refer to events where three or more detectors simultaneously detect a photon. These events may include a true coincidence with one or more single events, but since it is impossible to tell which detector pair registered the true event, the event is discarded. These events do not directly add noise to the data and do not require any corrections. However, since many of the multiple events contain true coincidences, they are a source of data loss or dead time.

6.4 Prompt Coincidences

Prompt coincidences refer to the events that are registered by the coincidence circuit (without any time delay). These events include all true, scatter, and accidental coincidences and are the total number of registered coincidence events from which the accidental and scattered events need to be subtracted from.

6.5 Noise Equivalent Counts

As a consequence of the subtraction of accidental and scattered coincidences from the prompt coincidences, there is an increase in statistical noise in the net true counts. A metric to

characterize this noise increase in the net true count rate is the Noise Equivalent Count rate or NEC rate. The *NEC* is defined as (Strother et al. 1990):

$$NEC = \frac{T^2}{T + S + kfA}, \tag{12}$$

where *T, S*, and *A* are the Trues, Scatter, and Accidental coincidence rates, respectively, and *f* is the fraction of the sinogram or FOV width subtended by the imaged object. The factor *k* depends on which method of estimation of accidental events is used. *k* is unity if the singles method is utilized and 2 if the delayed coincidence method is used. The NEC equation describes the equivalent count rate, in the absence of scatter and accidentals, after the subtraction of these events from the prompt rate. The NEC has also been shown to be directly proportional to the square of the signal-to-noise ratio in the reconstructed images (Dahlbom et al. 2005). However, the NEC does not take into account secondary degrading effects such as detector pile-up, which may result in additional noise in the final images.

7 Resolution Limitations

Like most imaging system, the spatial resolution in PET is limited by the intrinsic resolution of the detector elements. Because of the coincidence detection, the detector geometry, and the physics of the positron decay, there are additional resolution degrading components.

The resolution of the PET system is dependent on the size of the detector elements in the system. Due to the coincidence detection, the resolution varies across the FOV. At the mid-point between a pair of coincidence detectors, it can be shown that the resolution is equal to $\frac{1}{2}\ w$, where *w* is the width of the detector (Hoffman et al. 1982). The shape of the line spread function would be triangular in shape if a pair of ideal detectors were used (i.e., detector with 100% detection efficiency). If the source is moved closer to one of the detectors, the resolution will progressively degrade. The shape of the line spread function will be trapezoidal and will eventual have a square shape with a width equal to the detector width at the surface of one of the detectors.

In addition to the resolution variation in the tangential direction described above, there is also a resolution variation in the radial direction. The reason for this resolution loss is due to the depth in the detectors at which the annihilation photons interact. At off-center positions in the FOV of a PET system, the scintillation detectors are angled relative to each other. When the annihilation photon pair hits the detectors at an angle, there is a probability that one or both photons will penetrate the detector they entered through and deposit their energy in an adjacent detector. The result is a mispositioning of the LOR and a broadening of the line spread function. This resolution loss could be reduced if the depth in the detector where the photon interacted could be determined. A number of different detector designs with depth-of-interaction (DOI) capability have been proposed over the years. These include the use of additional photodetector readouts or the use of multiple layers of different scintillation detector materials and the use of pulse-shape discrimination to determine the DOI.

As described above, following a decay of the mother nucleus, the positron will travel some distance before it annihilates and emits the two 511-keV photons. Therefore, the PET system will localize the position of the annihilation rather than the location of the radioactive source. The random isotropic emission of the positron together with an energy distribution adds an

uncertainty to the localization of the source. Since the energy distribution of the positron is different for different isotopes, the amount of resolution loss depends on the isotope used (Levin and Hoffman 1999). This resolution loss ranges for 0.18 mm full width at half maximum (FWHM) (0.65 mm full width at tenth maximum FWTM) for ^{18}F to 0.56 mm FWHM (4.5 mm FWTM) for ^{82}Rb (Sanchez-Crespo et al. 2004).

Another assumption made in PET is that the two annihilation photons are emitted back to back or 180° apart. However, if the positron–electron pair is not at rest at the time of annihilation, the emission will deviate slightly from 180°. This deviation from the assumed 180° emission translates to a resolution loss that will depend on the detector separation (see *Fig. 4*). The amount of deviation is approximate Gaussian in shape with a FWHM of about 0.5°. This translates into a resolution uncertainty of about 2 mm at a detector separation of 80 cm.

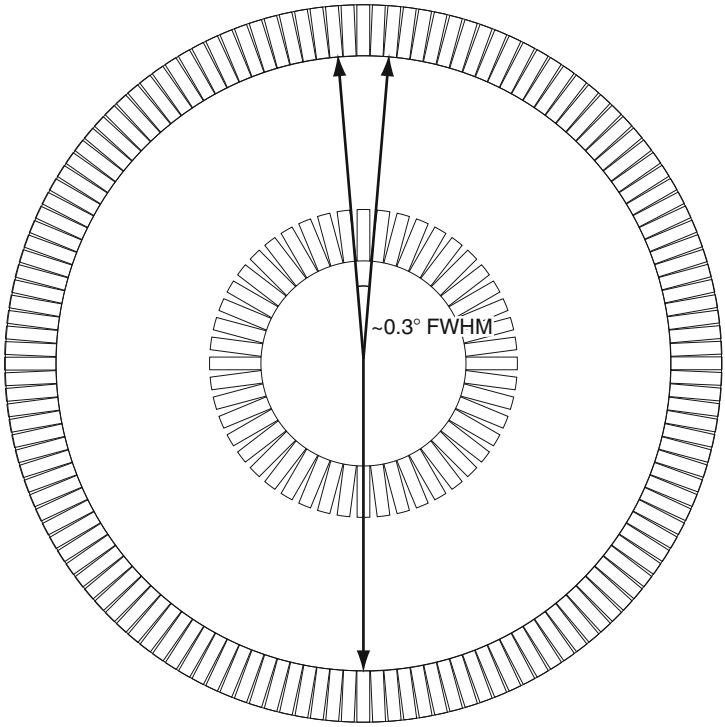


Fig. 4

Illustration of the effect of photon acollinearity on spatial resolution. If the positron–electron pairs are not at rest at the time of annihilation, the two annihilation photons will not be emitted exactly 180° apart. This deviation has approximately a FWHM of 0.5° and translates to a loss in spatial resolution. The amount of resolution loss depends on the detector separation and is about 2–3 mm for a human whole-body system, and only a fraction of a mm for a small-diameter-animal system

8 Data Collection, 2-D and 3-D PET

A complete PET system for routine imaging is made up of a large number of detectors placed around the object to be imaged. To increase the axial coverage and to increase detection efficiency, multiple detector rings are placed next to each other. For a modern PET system, the detector separation or ring diameter is 80–100 cm, with an axial coverage of 15–22 cm.

Traditionally, PET systems collect data in what is referred to as 2-D mode. This means that the recorded coincidences originate from a thin slice of the object. This slice is defined as the space defined by the coincidences collected within the same detector ring or the nearest neighboring detector rings. To better define this slice plane and also to reduce out-of-plane scatter, lead collimators or septa are placed between each detector ring (see *Fig. 5*). By stacking the images from adjacent slices, a 3-D volume can be generated. Limiting the coincidence detection

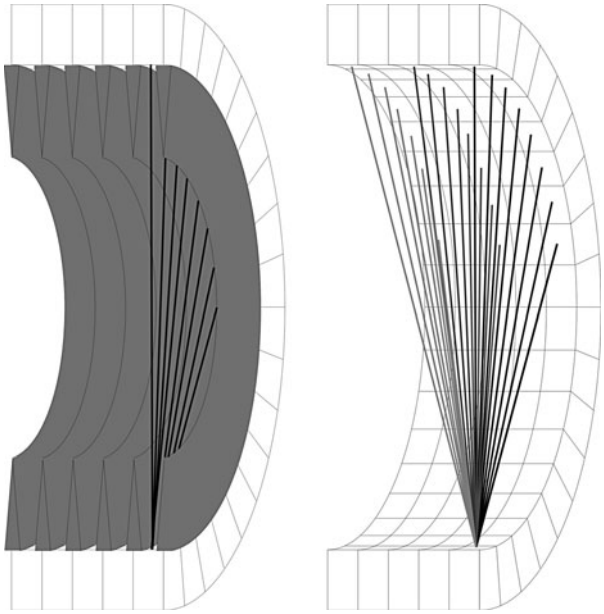


Fig. 5

Illustration of the 2-D and 3-D acquisition modes in PET. PET systems are made up of several adjacent detector rings. Older PET systems only acquired coincidences between detector elements within the same or the closest neighboring detector ring. To better define the imaged slice and to reduce the amount of detector scatter, lead collimators or septa were placed between the detector rings (*left figure*). Since the data collection is limited to a set of thin slices, the acquisition mode is referred to as 2-D. If the lead shields are removed (*right figure*), each detector element can form a coincidence between a larger number of detector elements in the system, which will improve the overall sensitivity. Since coincidences are acquired from a volume, this acquisition mode is referred to as 3-D. Although the sensitivity is dramatically improved in 3-D mode, there is also an increase in detected scatter, which will offset the effective sensitivity

to a thin slice results in a relatively poor detection efficiency of the system. The overall detection efficiency can be significantly improved by removing the inter-plane septa and allowing each detector to record coincidences with detector elements in all detector rings. This results in a collection of events from the entire volume within the FOV instead of just a single slice. This acquisition mode is therefore referred to as 3-D PET.

Although the sensitivity dramatically improves in 3-D mode compared to 2-D, this is offset by a significant increase in detected scatter and accidental coincidences. The removal of the septa results in a significant increase in the number of recorded scattered events from 10–15% in 2-D mode to 40–50% in 3-D mode. Furthermore, this also increases the photon flux each detector element is exposed to, which dramatically increases the number of detected single events. As a consequence, there is an increase in the number of accidental coincidences (see Eq. 11). The increase in photon flux also places a greater demand on the count-rate capabilities on the scintillation material and the electronics. The correction of the prompt count rate for the increase in both scatter and accidental coincidences results in a reduction in the effective sensitivity (i.e., a reduction in NEC).

9 Data Corrections

One of the goals in PET imaging is to produce an image volume that accurately describes the true distribution of the injected activity. Using the projection data acquired by the scanner, this can be accomplished by using tomographic reconstruction techniques. However, there are a number of corrections that need to be applied to the data before reconstruction or have to be included in the reconstruction algorithm in order to produce an image that is quantitatively correct.

9.1 Normalization

A modern PET system may have around 30,000 detector elements. Each of these detector elements will have slightly different detection characteristics due to a number of factors. These include differences in physical dimensions, variations in light output, differences in energy threshold and timing settings, and efficiency variation due to geometry. The purpose of the normalization is therefore to apply calibration factors that equalize the detection efficiency throughout the system.

To generate the normalization correction, the individual detector efficiencies or coincidence efficiencies have to be measured. The normalization also includes corrections for geometrical and plane-to-plane efficiency variations.

The most straightforward way to determine the normalization is the direct measurement using a uniform plane source filled with a long-lived isotope such as ^{68}Ge (Hoffman et al. 1989). This measurement has to be performed for every possible line of response in the system. The inverse of the relative detection efficiency that can be derived from this measurement is then used as the normalization correction. This measurement will take into account all geometrical and detection efficiency variations. Although this method is easy to implement, it has a number of practical limitations. Since the source used for this measurement has to be of relatively low activity in order to avoid pile-up effects in the detector block, the main challenge is to acquire

enough counts per LOR with a minimum of statistical noise that would otherwise propagate into the final image. This is a particular challenge in high-resolution systems where the coincidence detection efficiency is very low. Furthermore, any nonuniformity in the source will propagate into the reconstructed images.

The component-based method is an alternative method to derive the normalization correction which overcomes some of the practical limitations of using a plane source for the normalization measurement (Casey et al. 1995). In its simplest form, this method assumes that the coincidence detection efficiency for a pair of detectors is the product of the individual detection efficiencies, ε, and geometrical factors, g. The normalization factor for a pair of detectors i and j is then given by:

$$n_{ij} = \frac{1}{\varepsilon_i \varepsilon_j g_{ij}}. \tag{13}$$

The geometrical factors include corrections for angle of incidence of the photons on the detector elements, systematic efficiency variations across the detector modules, and relative plane efficiencies. Since these factors do not change, they are usually determined once for a system or system geometry. On the other hand, the individual detector efficiencies have to be determined on a more frequent basis since these are sensitive to any electronic drift in the system. This is usually done with a uniform cylinder source placed at the center of the scanner FOV. For each detector element in the system, the sum of all true coincidences between the detector and all of its opposing detectors is determined. The assumption is that by averaging the detected coincidences over a large number of opposing detectors, this sum is proportional to the detection efficiency of the single detector. In contrast to the plane source method described above, this method provides an estimate of the normalization factors that is almost free of any statistical noise. However, systematic errors may be introduced from the estimation of the geometrical factors.

9.2 Attenuation Correction

The correction that has the most significant impact on the image data is the attenuation correction. Without correction for attenuation, the reconstructed images will give a distorted view of the activity distributions. Furthermore, the images are not quantitative. An example of this is illustrated in ❯ *Fig. 6*. As discussed earlier, at 511 keV, there is a relatively high probability that one or both of the two annihilation photons will interact before escaping the object that is imaged. The result is that primary photons are removed or attenuated from the LORs in the system and the number of measured coincidences is underestimated.

In order to reconstruct an image that is free of artifacts, the amount of attenuation along each LOR in the system has to be determined and a correction factor has to be applied to the acquired number of counts. This correction factor can be determined using a mathematical model or by direct measurement.

The amount of attenuation along a given LOR can be described by the following example illustrated in ❯ *Fig. 7*. Consider a positron-emitting source located at an unknown depth x in a medium with uniform attenuation coefficient μ. If the thickness of the object is D along the LOR, then the probability that photon 1 will escape the object without interacting is (❯ Eq. 9):

$$p_1 = \frac{I(x)}{I_0} = e^{-\mu x}. \tag{14}$$

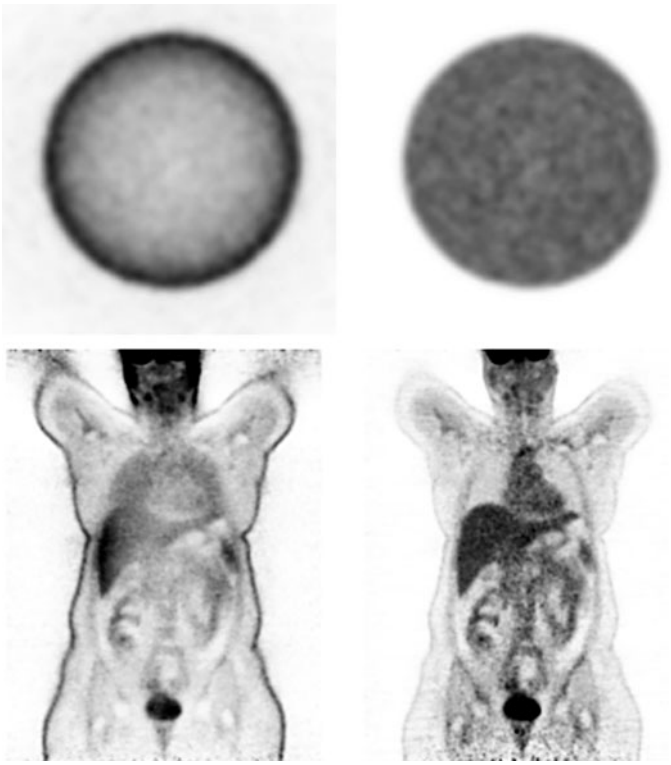


Fig. 6

Illustration of the effect of photon attenuation in PET. The *left images* were reconstructed without attenuation correction. The cross section of the uniform cylinder (*top*) appears to have suppressed activity in the *center*. The coronal cross section of a patient shows an apparent high uptake in the skin and in the lungs. Large organs in the abdomen, such as the liver, appear to have a very heterogeneous uptake. The images to the *right* were reconstructed with attenuation correction. The image of the cylinder is now uniform. The uptake in the internal organs in the patient scan is more uniform

The probability that the second photon will escape the object is similarly:

$$p_2 = \frac{I(D-x)}{I_0} = \mathrm{e}^{-\mu(D-x)}. \tag{15}$$

In order to produce a coincidence event, both photons have to escape and the probability for this is then the product of the individual probabilities:

$$p_{\mathrm{coinc}} = p_1 p_2 = \mathrm{e}^{-\mu x}\mathrm{e}^{-\mu(D-x)} = \mathrm{e}^{-\mu D} = \frac{I(D)}{I_0}. \tag{16}$$

From this equation, it can be seen that the amount of attenuation is independent of the location or depth of the source and is only dependent on the total thickness of the object and the attenuation coefficient of the object along the LOR. This is also true for a distributed source inside the

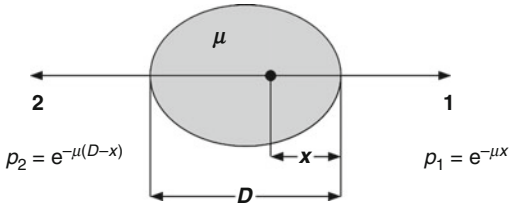


Fig. 7

Equations behind the attenuation correction in PET. The probabilities for the escape of the two annihilation photons emitted at a depth *x* inside a medium of thickness *D* and a uniform attenuation coefficient μ are given by p_1 and p_2. The probability for a coincidence detection is then the product of p_1 and p_2. The correction for attenuation along this LOR is then simply $(p_1 p_2)^{-1}$

object, which can be seen as a large number of point sources. This is a unique property of the coincidence detection method and makes correction for attenuation relatively straightforward in PET. The attenuation correction factor f_{ac} that needs to be applied to the acquired number of emission counts along the LOR is simply the reciprocal of the expression in Eq. 16:

$$f_{ac} = e^{\mu D}. \tag{17}$$

The attenuation correction factor has to be applied to every measured LOR in the system. Therefore, the thickness of the object has to be estimated along each LOR. Furthermore, the object has to have a uniform attenuation coefficient. This method is usually referred to as calculated attenuation correction.

To generate a calculated attenuation correction, the emission data is initially reconstructed without attenuation correction. The resulting images are then used to estimate the outline of the object. The outline can either be approximated with a simple geometrical shape or the actual outline is determined (Bergström et al. 1982). This method was frequently used in brain imaging in early PET systems, and the method produces images that are largely free of attenuation artifacts. The method is, however, prone to introduce artifacts if the outline of the object is not accurate. Furthermore, the method also tends to introduce quantitative errors, especially in brain imaging due to the higher attenuation in the skull, which is not taken into account. Methods have been developed to reduce these effects (Siegel and Dahlbom 1992); however, this method is rarely used in modern PET systems where measured attenuation corrections are almost exclusively used.

The derivation of Eq. 17 assumes a uniform attenuation coefficient throughout the object to be imaged. In most cases, the attenuation is heterogeneous and the simple calculated attenuation correction cannot be used. For a heterogeneous attenuating object, it can be shown that the amount of attenuation or the probability for a coincidence along a given LOR is given by:

$$p_{coinc} = e^{\int_0^D \mu(l)\,dl}. \tag{18}$$

It is therefore necessary to know the distribution of attenuation coefficients within the object in order to generate an attenuation correction. This can be done using an external positron-emitting source or an X-ray CT, as described in the next section.

As shown by Eq. 17, the attenuation correction factor f_{ac} is independent of the location and the distribution of the positron-emitting source along the LOR. Therefore, an external point or line source placed along the LOR will experience the same amount of attenuation as an internal distributed source. This is what is utilized in the measured attenuation correction method. In this method, one or more positron-emitting sources are used to directly measure the attenuation along each LOR in the system (Ranger et al. 1989). Initially, a reference scan is acquired where coincidences are acquired without anything within the FOV. This scan, which is referred to as a blank scan, represents the measurement of I_0 in Eq. 16. Following this scan, the object to be imaged is placed in the FOV and the acquisition is repeated and this scan, which is referred to as the transmission scan, represents the measurement of $I(D)$ in Eq. 16. The attenuation correction factor $I_0/I(D)$ is then simply the ratio of the number of counts acquired along each LOR of the blank and the transmission scans, respectively.

This method directly measures the attenuation along each LOR, and there are no assumptions made regarding the shape of the object or the attenuation coefficients. The main challenge is to acquire enough counts in the transmission scan along each LOR in order to avoid excessively noisy estimates of the attenuation correction factors. This is especially a problem in imaging of the chest and abdomen where the attenuation is the greatest, and for large patients it is not unusual to have attenuation factors of 40–50 in this region (Dahlbom and Hoffman 1987). The statistical quality of the transmission scan is therefore usually very poor, and this noise will propagate into the emission image if left unprocessed. A common method to reduce the noise is to apply a smoothing filter to the transmission and blank data. However, excessive smoothing may introduce artifacts in the reconstructed emission images, especially at interfaces with very different attenuation coefficients (e.g., lung and soft tissue in the chest) (Huang et al. 1979). Hybrid techniques have therefore been developed which combine the measured and calculated attenuation correction techniques (Huang et al. 1981; Xu et al. 1991; Meikle et al. 1993). In these methods, the transmission scan data is used to reconstruct a CT-like image of the attenuation coefficients. Although the result is very noisy, the information content is enough for an image segmentation algorithm to classify the pixels in the image into attenuation coefficients of distinct tissue types (e.g., soft, lung, and bone tissue). The result of this classification of the pixel values is an image of the attenuation coefficients virtually free of noise. This image can then be used to calculate the attenuation correction for each LOR in the system.

Although the transmission scan with the combination of image segmentation results in a direct measurement of the amount attenuation, this correction is still a very lengthy procedure which may take up one third to half of the total scanning time. As a result, motion may occur between the emission and transmission scans, which will introduce artifacts in the final image. As an alternative to the use of transmission sources, a CT scanner can be added to the PET system to acquire the attenuation map (Beyer et al. 1994). The use of the CT image for attenuation correction results in an almost noiseless attenuation correction. However, since the CT uses an X-ray tube, operating at a 120–140 kVp with an average energy of about 70 keV, the measured attenuation map has to be converted into a map at 511 keV. The most common method for achieving this is to use the bilinear scaling method (Hasegawa et al. 1993; Kinahan et al. 1998). In this method, the CT image is first segmented into two parts, typically values above (bone tissue) and below (soft tissue) 0 HU or water level. The segmented images are then scaled with two different scale factors that yield approximately correct attenuation values for 511-keV

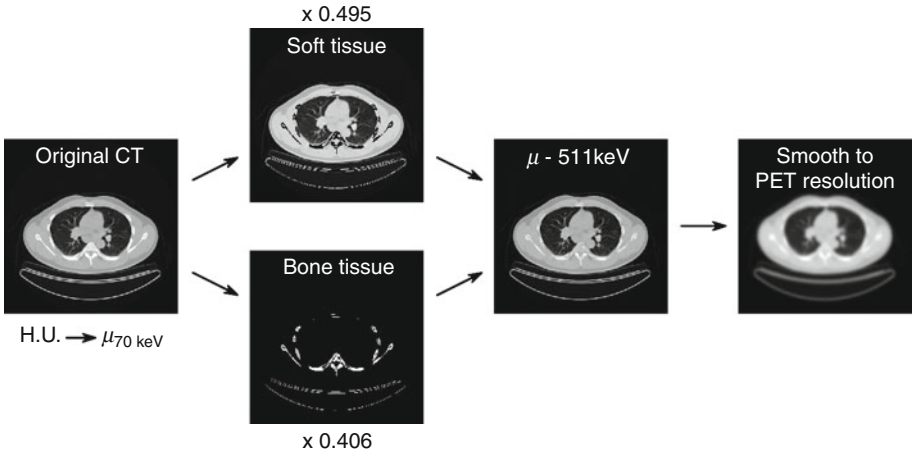


Fig. 8

Illustration of the use of CT images for attenuation correction. CT images cannot directly be used for attenuation correction in PET since the mass attenuation coefficient is significantly different in soft and bone tissues at X-ray CT energies compared to 511 keV. The CT images are therefore segmented into "soft tissue" and "bone tissue" using thresholding. Different scale factors are then applied to the different tissue types to convert the attenuation coefficients at CT energies to attenuation coefficients at 511 keV. The scaled images are then added and spatially smoothed to the resolution of the PET system. The final image is then used to calculate the attenuation correction factor along each LOR in the system

photons. The reason for the two scaling factors is that at HU values below 0, Compton interaction is the predominant interaction at both 70 and 511 keV. However, at higher HU values, there will be a mix of Compton and photoelectric interactions. Since the relative probability for Compton and photoelectric interactions is significantly different at 70 and 511 keV, a different scale factor has to be applied. This process is illustrated in *Fig. 8*.

9.3 Scatter Correction

Scatter and attenuation are closely related since scatter is sometimes a product of the interaction process. Correction for scatter is one of the most difficult corrections to apply in PET since these events are indistinguishable from true events. Scattered photons have a lower energy than the primary 511-keV photons; however, the energy loss is relatively nominal and approximately one third of the single scatter photons will only lose 100 keV or less. Since the energy window is set relatively wide, 400–600 keV for a system with high-energy-resolution detectors, a significant number of scattered events will be detected. There are two reasons for setting a fairly wide energy window. First, as all scintillation detectors have a finite energy resolution, it is necessary to set an energy window to ensure that most of the events that fall within the photopeak will be recorded. Second, a significant fraction of all primary photons will only deposit a portion of the energy within the detector volume. Although these are good events, they are detected in the

same energy range as scattered photons. Using a more narrow energy window will reject some of the scattered photons but will also result in an efficiency loss due to the rejection of primary photons. It will therefore always be a tradeoff between detection efficiency and the acceptable level of detected scattered events.

The amount of detected scatter depends on a number of parameters including energy thresholds, activity distribution, dimensions of the scattering medium (i.e., patient size), distribution of attenuation coefficients, and scanner geometry (i.e., 2-D with septa or 3-D). One efficient way to reduce scatter is the use of lead collimators, placed between each detector ring in the system as described above. However, this also reduces the overall detection efficiency of the system. The latest generation of PET systems all operate exclusively in 3-D mode, since it has been well established that with fast scintillation detectors 3-D outperforms 2-D, in terms of sensitivity despite the increase in scatter. Because of the high scatter fraction in most imaging situations in 3-D PET, it is necessary to apply an accurate scatter correction. There have been a number of proposed methods for scatter corrections. In the simplest approach, the amount of scatter is estimated by fitting a smoothly varying function (e.g., Gaussian or polynomial) to the detected counts outside object in measured projection data. This method assumes that the scatter varies smoothly across the FOV and is relatively independent of the source distribution and the scattering medium. Since this is usually not the case, this method only works reasonably well in situations where the activity distribution is uniform and is symmetrical (e.g., uniform phantoms). A widely used analytical method in 2-D PET imaging is the Bergström method (Bergström et al. 1980). This method is based on measurements of the scatter distributions of line sources placed at different positions inside a scattering media, usually a uniform cylindrical phantom 15–18 cm in diameter. These scatter distributions are then used to deconvolve the scatter from the measured data. The advantage of this method is that it takes into account the activity distribution in the estimation of the scatter. Since the scatter distribution is highly dependent on the size, shape, and distribution of attenuation coefficients, this method is still an approximation.

The most accurate methods to estimate scatter are the simulation-based methods. In these methods, the scatter is estimated based on a simulation of the scatter using approximate source and attenuation distributions. As an initial approximation of the actual source distribution, the emission data, reconstructed without scatter correction, is used. The scatter is then estimated using a single-scatter model (Watson et al. 1997), or an approximate analytical model (Ollinger 1996), or using a Monte Carlo simulation (Levin et al. 1995). This process may be repeated after the initial scatter estimate, where a scatter-corrected distribution is used to produce a new scatter estimate. The advantage of this method over other scatter correction methods is that both the actual distribution of activity and attenuation coefficients of the imaged object are used in the estimation of the scatter. The drawback is that these methods are computationally more expensive; however, with the power of modern computers, this is no longer an issue. Because of its high accuracy, the simulation-based scatter correction is routinely used in human imaging systems. Yet, it does not account for activity outside the FOV of the scanner that is not recorded at any scanner bed position.

10 System Calibration and Quantification

The reconstructed images from an emission data set (corrected for accidentals, scatter, and attenuation) represent the relative activity distribution and the pixel values are typically in some

arbitrary units. Although this may be adequate for visual interpretation of the data, it is many times desirable to extract quantitative information from the images such as the absolute activity concentration within a region of interest. Since the emission data can relatively accurately be corrected for attenuation and scatter, absolute quantification is possible in PET, at least in objects where the partial volume effect is negligible.

To achieve this, the system needs to be calibrated against a source of known activity concentration. This source is typically a cylinder source or phantom of known volume filled with a known amount of activity of ^{18}F. The phantom is imaged in the system, and images are reconstructed with all corrections applied. An ROI or VOI is then drawn on the image volume, which will produce some arbitrary average image value. A calibration factor for the system can then be calculated from:

$$calibration\ factor = \frac{Activity/Volume}{Average\ ROI\ counts}. \tag{19}$$

The activity is the measured activity of the dose injected into the phantom, which is typically determined using a dose calibrator. This activity has to be decay-corrected to the time of the image acquisition. This calibration factor is then used in all subsequently acquired images as a multiplicative factor to produce images where the pixel values represent activity concentration. The greatest uncertainty in this calibration procedure is the determination of the accuracy of the activity calibration used in the phantom measurement, which is typically no better than 5–10% for most dose calibrators.

10.1 Partial Volume Effect

The calibration procedure described above allows the PET scanner to produce images that are quantitative. The main challenge in PET is the quantification of structures that have a physical extent of the same magnitude as the spatial resolution of the reconstructed image. The effect the limited spatial resolution has on a small structure is to spread out the activity over a larger area. Small structures therefore appear to have a lower activity concentration with poorly defined edges compared to larger structures with the same activity concentration. This is referred to as the Partial Volume Effect and is illustrated in ❯ *Fig. 9* for a series of spheres with the same activity concentration for a system with a resolution of 8 mm FWHM. The degree of suppression in the spheres is illustrated by the line profiles. The amount of suppression is a function of both the dimensions of the object and the reconstructed image resolution and is characterized by the recovery coefficient *RC* (Hoffman et al. 1979):

$$RC = \frac{Measured\ peak\ activity\ concentration}{True\ activity\ concentration}. \tag{20}$$

The calculated RC for different sphere diameters, normalized to the image resolution, is shown in ❯ *Fig. 10*. This graph shows that if the object is about three times the image resolution in diameter, a small region placed at the center of the object should accurately represent the true activity concentration. The ROI used under realistic imaging conditions should be large enough to minimize statistical noise (by voxel averaging), but small enough to reduce averaging or partial volume effects. The RC can be determined if the dimensions of the object are known, which would make it possible to correct for the partial volume effect. The challenge is that it is difficult to estimate the actual object dimensions from the PET image due to the blurring of the partial volume effect. The dimension of the structure can instead be estimated if registered

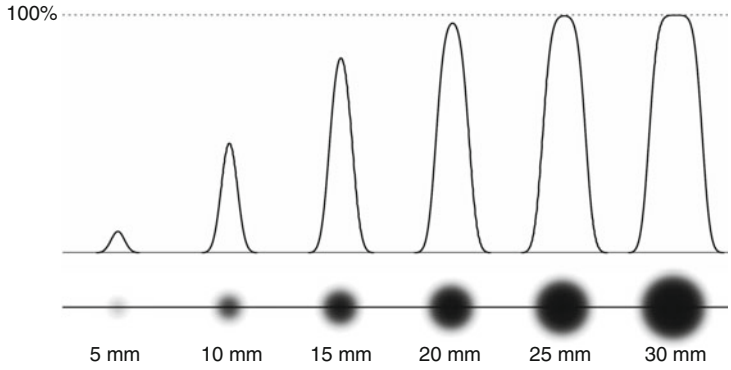


Fig. 9

Illustration of the partial volume effect on a series of spheres of different diameters but with equal activity concentration imaged on a simulated system with a spatial resolution of 8 mm FWHM (no noise added). Due to the limited spatial resolution of the system, the smaller spheres will appear to have a lower intensity compared to the larger spheres

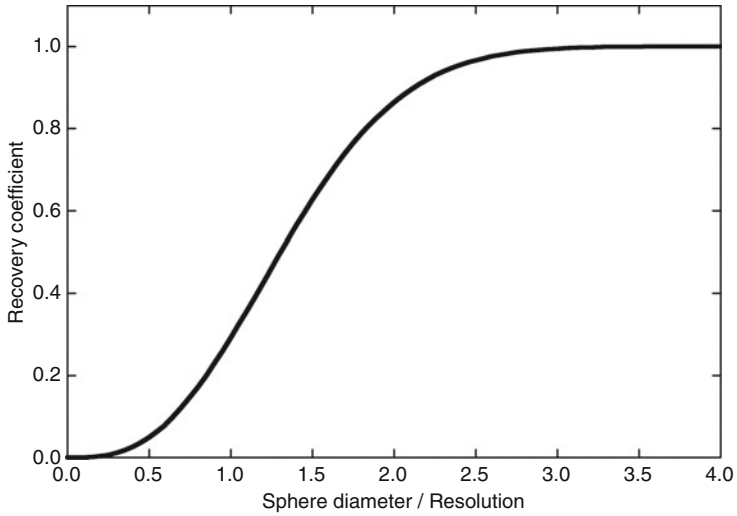


Fig. 10

Calculated recovery coefficients for spheres imaged on a noise-free system with an isotropic spatial resolution. As can be seen from the curve, the activity in the spheres is only fully recovered if the sphere diameter is at least three times the image resolution (in FWHM)

anatomical image information such as CT or MR is available, and if the edges of the anatomical structures genuinely coincide with the edges of the metabolically active structure, which is not always the case.

The issue of the partial volume correction becomes further complicated by the fact that the apparent uptake in a region not only contains suppressed activity from the structure itself, but

may also contain activity added from uptake of surrounding structures. To accurately quantify the activity in a region, it is thus necessary to correct for this spillover activity.

11 Image Reconstruction

The data collected by a PET scanner is a set of projections acquired between 0° and 180° around the object. If the 3-D acquisition mode is used, then there are also a number of azimuthally acquired angles. This data set is then used to reconstruct the 3-D activity distribution within the object. There are several methods for reconstructing the data; however, these can be divided into two distinct groups: analytical and iterative reconstructions.

11.1 Filtered Backprojection

The most commonly used analytical image reconstruction method is filtered backprojection or FBP, which is also routinely used in X-ray CT and SPECT imaging (Brooks and Di Chiro 1976). In this reconstruction method, the corrected projection data of each slice are backprojected onto the image matrix at each angle. The backprojection operation will result in an image that resembles the true activity distribution but with lower contrast and resolution. As a result of the backprojection, the relative quantification is also lost. It can be shown that the resulting image from a backprojection is the true activity distribution convolved with a $1/r$ filter:

$$a(x,y)_{\mathrm{bp}} = a(x,y) \otimes \frac{1}{r} = a(x,y) \otimes \frac{1}{\sqrt{x^2+y^2}}. \tag{21}$$

In the frequency domain, ❯ Eq. 21 can be written as:

$$A(u,v)_{\mathrm{bp}} = A(u,v)\frac{1}{\rho} = A(u,v)\frac{1}{\sqrt{u^2+v^2}}, \tag{22}$$

where $A(u,v)$ is the 2-D Fourier transform of the $a(x,y)$ and u and v are the spatial frequencies. This shows that the backprojection operation will dampen high frequencies in the image, which explains the low resolution in the image.

This blurring can be reversed if a filter is applied to the data, which linearly amplifies the spatial frequencies. This filtering step will remove the blurring of the image that would otherwise occur if the projections were simply backprojected. It will also preserve the relative quantification between structures in the image.

One of the problems with the filtering is that it will also amplify high-frequency noise (e.g., statistical noise). Since most PET studies contain relatively high levels of statistical noise, it is always necessary to apply a smoothing filter, which dampens the high-frequency noise. Since the high spatial frequencies carry the information to visualize small objects and edges, this filtering always results in a loss in resolution and ability to visualize and accurately quantify small objects. The amount of smoothing that is necessary to apply to the data depends on the level of statistical noise, which in turn depends on how many counts were acquired. Furthermore, the level of smoothing also depends on what level of noise the observer finds acceptable in order to identify objects of a certain size or contrast in the image (i.e., contrast resolution).

Images reconstructed with FBP from data acquired under clinical conditions will typically always have a much lower resolution than the intrinsic spatial resolution of the scanner.

11.2 Iterative Reconstruction

The second class of image reconstruction methods used in PET imaging is the iterative reconstruction algorithms. These algorithms have to a great extent replaced the traditional FBP in PET imaging. The main reason for this is the improved image quality that can be achieved using iterative image reconstruction, especially in terms of improved S/N.

The basic principle for iterative reconstruction is shown in *Fig. 11*. The goal of the algorithm is to reconstruct an image that through a series of iterations best represents the acquired data by the scanner. The iterative process starts with an initial estimate of the image, which is many times a uniform distribution. The next step is to sum up the image intensities along the same LORs as the scanner would acquire the data. This process is usually referred to as forward projection (i.e., the opposite of the backprojection operation). The forward-projected data is then compared to the measured projection data. Any difference between the measured

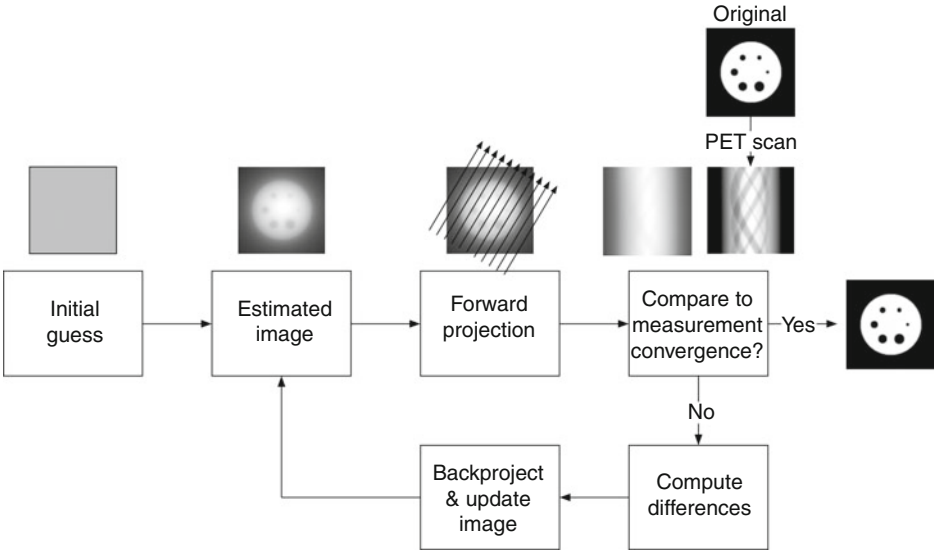


Fig. 11

Diagram of the iterative image reconstruction process. Following an initial guess of the activity distribution, the data are forward projected along the LORs of the system. The forward-projected data is compared to the measured data. Following the comparison, adjustment factors are calculated and backprojected into image space and the image is then updated. This updated image is then the starting point for the next iteration. This process is repeated until the difference between the measured and forward-projected data set has reached a specific level (i.e., convergence has occurred)

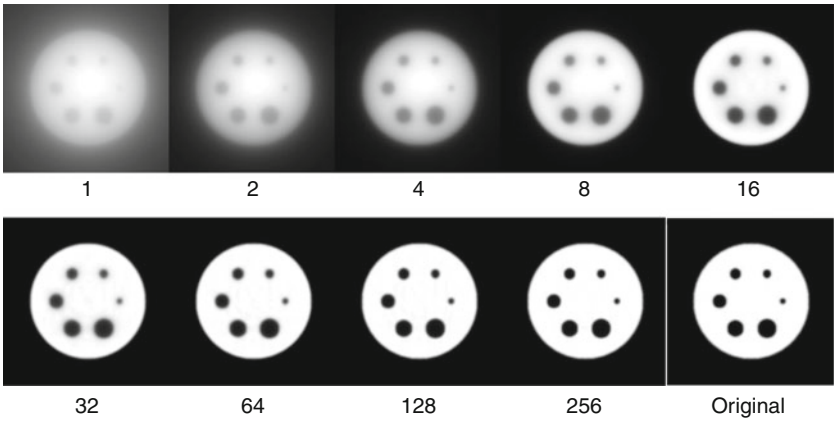


Fig. 12

Illustration of the progression of a ML-EM reconstruction of a noise-free simulated phantom. As the number of iterations increases, there is only small improvements in image quality indicating that convergence has been reached

and forward-projected data is then used to make adjustments to the initial estimate along the measured LORs (e.g., a backprojection operation). This process is then repeated until the difference between the forward-projected and measured data sets has reached a specified level. The reconstruction of a phantom for different numbers of iterations is shown in *Fig. 12*.

One of the advantages of the iterative reconstruction algorithms is that they attempt to arrive at an image or activity distribution that most likely would have created the measured projection data, taking into account both the effects of counting statistics and the physics of the measurements (e.g., scatter, normalization, detector resolution, etc.) (Shepp and Vardi 1982; Lange and Carson 1984). The drawback of the iterative reconstruction algorithms is that they are very computationally expensive. In contrast to FBP where one backprojection is required, iterative reconstruction algorithms may require a large number of iterations before convergence is reached, where each iteration requires one forward and one backprojection. A number of methods have therefore been developed to accelerate the reconstruction. One of the most commonly used acceleration techniques is the ordered subsets (Hudson and Larkin 1994). In this method, the projection data is first partitioned into a number of subsets. During the reconstruction, the forward projection is only performed at the different subset of the projection data. The result is a speedup factor close to the number of subsets the data is divided into.

12 Time-of-Flight PET

The coincidence detection method described earlier only registers the detector pair that was struck by the two annihilation photons. Therefore, the event is only localized to a thin channel or a LOR connecting the two detectors. Since the two photons are traveling a distance at the speed of light, it will take each photon a measurable amount of time to reach each detector.

If the annihilation occurred at the midpoint between the detector pairs, then the time it would take each photon to the each detector would be the same. However, if the annihilation occurred closer to one of the two detectors, then one photon would reach the near detector before the other one. This time difference Δt can then be used to localize where along the line of response the annihilation occurred. Using classical mechanics, it can be shown:

$$t_1 - t_2 = \Delta t = \frac{2x}{c} \Rightarrow x = \frac{c\Delta t}{2}. \tag{23}$$

Thus, with a very fast detection system (i.e., infinite time resolution), each annihilation event could potentially be localized in the FOV, without the need for any image reconstruction. As discussed above, current scintillation detectors materials unfortunately do have a finite time resolution, which limits how precise the events can be localized in time. Some of the fastest materials today have a time resolution of a few hundred ps. Using the equation above, this time resolution translates into a uncertainty in distance. For instance, a time resolution of 300 ps translates into a positioning uncertainty of approximately 4.5 cm.

If the time-of-flight (TOF) information is used in the image reconstruction, it can be shown that this will result in a gain in signal-to-noise ratio (SNR) in the reconstructed image (Budinger 1983):

$$SNR_{\text{TOF}} = \sqrt{\frac{D}{\Delta x}} SNR_{\text{non-TOF}}, \tag{24}$$

where SNR_{TOF} and $SNR_{\text{non-TOF}}$ are the SNR with and without the TOF information used in the reconstruction, respectively. D is the object diameter and Δx is the TOF resolution. This expression shows that as the time resolution improves, the gain in signal to noise increases. Furthermore, this gain is also the greatest for objects of large extent. The reason for this improvement in S/N is that the statistical noise in the measured data is not spread over a large area during the reconstruction. In non-TOF reconstruction algorithms, the noise in the measured projection data is backprojected along each LOR onto the image matrix and is therefore adding noise over a larger area. In a TOF reconstruction, the measured data will only be backprojected over a region with an extent corresponding to the time resolution of the system and centered at a position determined by the measured time differences Δt. Since the noise is contained within a smaller area, it is expected, according to the expression above, that the imaging situations that will gain most from TOF are the large cross-sectional areas of the body such as the abdomen. It is also expected that larger patients would benefit more from the use of TOF PET than thin patients.

Although TOF has received a lot of attention in recent years, this technology was first proposed in the early 1980s by the LETI group in France (Allemand et al. 1980). TOF instrumentation and image reconstruction were studied extensively during this time. A number of TOF scanners were built based primarily on CsF and BaF_2 detectors. Although these systems used very fast detector materials capable of time resolution of a few hundred ns, the detection efficiency of these materials is very poor due to the relatively low density and low effective Z. The gain by the TOF information was therefore never capable of offsetting the loss in detection efficiency. In addition, the electronics used in these systems was not stable enough to allow reliable operation. This technology was therefore abandoned for many years in favor of the use of denser detector materials such as BGO and GSO. It was not until the discovery and widespread availability of new fast and dense detector materials such as LSO, LYSO, and $LaBr_3$ that the TOF technology was revisited and has been implemented in several commercial PET systems

(Conti et al. 2005; Surti et al. 2007). Preliminary studies on systems using TOF have shown, as expected, an improvement in S/N and lesion contrast (Kadrmas et al. 2009; Lois et al. 2010).

13 Multi-modality Imaging

13.1 PET-CT

The combination of PET and CT into a single imaging system has made a significant impact on several aspects of PET imaging. The addition of the CT has allowed a dramatic reduction in the overall scan time, for some protocols by a factor of two. With the latest generation of multi-detector CT systems, a whole-body scan only takes a few seconds as opposed to 15–20 min using transmission sources. The potential noise contamination from the transmission scan into the emission data has been eliminated when using the CT images for attenuation correction. The greatest advantage of a combined PET/CT is that by acquiring the image data on the same system under same imaging conditions, fusion of the functional PET images and anatomical CT images can be made with very high accuracy. Having the ability of the image fusion has turned out to be an invaluable tool for the interpreting physicians. One of the many challenges in interpreting PET images is that they contain somewhat limited anatomical information, and only the largest organs are clearly visualized. Having the CT volume registered to the PET volumes allows functional abnormalities observed on the PET to be accurately localized on the anatomical CT image. Furthermore, nonspecific tracer uptake in normal tissue can be distinguished from disease. The merging of the two imaging modalities into one imaging procedure has resulted in an increased diagnostic accuracy compared to PET alone. Because of this, all clinical PET systems that are manufactured today are PET/CT systems.

13.2 PET-MRI

Following the success of combining PET and CT, several research groups investigated the possibility of integrating PET and MRI. MRI has several attractive features including higher tissue contrast compared to CT, potential for providing complementary functional imaging to PET, spectroscopic information, and there is no exposure to ionizing radiation. There are a number of challenges in building a combined PET/MRI system. The PMTs used in the PET system are highly sensitive to magnetic fields and cannot be used in near vicinity of the magnet. The first functional MRI-compatible PET system therefore used very long optical fibers to lead the scintillation light to a PMT outside the fringe field of the magnet (Shao et al. 1997). Although this system successfully showed that it is possible to simultaneous acquire PET and MRI images, this system was limited to a single slice.

Solid-state photodetectors do not have the same sensitivity to magnetic fields as the PMTs and can therefore operate in an MRI system. However, these detectors have to be designed without any ferromagnetic material that would otherwise produce distortions in the magnetic field. Pichler et al. have successfully constructed a multislice PET insert (Pichler et al. 2006), which is currently in use for small-animal imaging (Judenhofer et al. 2008).

Siemens Medical systems have designed a PET brain insert that is inserted into the MRI bore to allow simultaneous PET and MRI imaging (Schmand et al. 2007). The detectors used in this

system are similar to what is used in the small-animal system described above. This system has a relatively long axial FOV of 19.25 cm, but because of the small detector diameter of 35.5 cm, it is limited to brain imaging. A clinical study comparing PET/CT and PET/MRI imaging showed that the image quality was similar for both systems (Boss et al. 2010). However, PET/MRI offers several advantages such as higher tissue contrast, allowing advanced MRI imaging techniques such as perfusion and diffusion imaging and MR spectroscopy, without adding radiation dose to the patient.

14 Dedicated Systems

14.1 Animal and Preclinical PET Systems

PET imaging has also become an important tool in the development of new radiopharmaceuticals, basic biological research, and in the studies of human disease. In 1992, the first PET system dedicated to animal imaging was designed (Cutler et al. 1992). This system was based on similar detector technologies used in high-resolution systems at that time but with a smaller system diameter (64 cm) to reduce the effect of the resolution loss due to the photon acollinearity. However, the system diameter was large enough to accommodate imaging of relatively large animals including primates and dogs.

Many models of human disease use mice as the animal of choice. Due to the small organ sizes, a spatial resolution less than 2 mm in all spatial dimensions is required. In the late 1990s, the first microPET® system was built by Cherry et al. (Cherry et al. 1997). This was the first dedicated small-animal PET systems and had a spatial resolution of around 2 mm in all spatial directions. In contrast to autoradiography, this system allows repeated noninvasive in vivo imaging of the same animal at multiple time points. This technology has opened up new opportunities for researchers to study biology.

Since the first microPET system was introduced, numerous other PET systems dedicated to small-animal imaging have been designed. Many of these have also been commercialized. Some of these systems are also multimodality systems allowing sequential or simultaneous imaging with MRI, CT, and SPECT.

14.2 Organ-Specific PET Systems

The success of PET and PET/CT in clinical imaging has also led to the development of organ-specific PET systems. This includes systems for breast imaging, neuroimaging, prostate imaging, and also imaging of the extremities. Common for all of these systems is that they are designed to overcome some of the limitations of whole-body PET systems, which can be seen as all-purpose systems. The primary limitation of the whole-body system is a limited spatial resolution, which is primarily dominated by the photon acollinearity effect at a detector separation of 80–100 cm. Furthermore, the detection efficiency is also relatively poor in a whole-body system. For these two reasons, conventional PET systems are typically not able to detect lesions with moderate tracer uptake that are smaller than 10 mm. Like the small-animal systems, the organ-specific systems have a significantly higher spatial resolution and sensitivity and should therefore be able to detect smaller lesions or structures.

14.3 Brain Imaging

Some of the first organ-specific PET systems were designed for imaging of the brain. An example of this is the HRRT system which consists of eight high-resolution panel detectors arranged in an octagon (Wienhard et al. 2002). The detector separation of this system is 46.9 cm and has 31.2 cm wide FOV and a 25.2 cm axial coverage. The detector in this system has DOI capability to a resolution of 7.5 mm, (two detector layers of 7.5 mm each). This reduces the DOI resolution loss seen at radial offsets. The spatial resolution of this system is approximately 2.5–3.0 mm across the entire FOV compared to about 5 mm for a high-resolution whole-body PET system.

14.4 Breast Imaging

Breast imaging with positron-emitting tracers is sometimes referred to as PEM (positron emission mammography). Several system designs have been proposed for breast imaging, but these can in general be divided into two major groups: partial and fully tomographic systems. In the partial tomographic systems, the patient's breast is slightly compressed between a pair of detectors (Weinberg et al. 1996). The imaging geometry is intentionally similar to what is used in conventional X-ray mammogram to allow easier image correlation. The detectors are either a pair of stationary panel detectors as in the commercial system from Naviscan or a pair of translating detector banks. Because of this particular geometry, only coincidence with limited angular sampling is acquired. As a result of the limited angular sampling, the resolution in the reconstructed images will not be isotropic. With an EM reconstruction algorithm, a resolution of 2 mm can be achieved within the image plane; resolution between image planes is on the order of 5–7 mm due to the limited tomography. Overall the volumetric resolution is significantly higher compared to whole-body PET systems, and has been shown to have a significantly better sensitivity for sub-centimeter-sized lesions (Berg et al. 2006).

A number of fully tomographic imaging systems have also been constructed (Doshi et al. 2000; Wang et al. 2006; Raylman et al. 2008; Tai et al. 2008). There are a number of different designs, but typically consist of a pair of panel detectors that can rotate around the breast to collect a complete tomographic data set. Since a full data set is acquired, the spatial resolution in the reconstructed images will be relatively isotropic. However, the motion of the detector panels adds complexity and cost to the system and the overall scanning time is extended.

One of these fully tomographic systems has also been combined with an X-ray CT system that allows fusion of the functional image with structural images (Bowen et al. 2009). The setting of the X-ray CT acquisition was selected to not give a higher radiation dose than a conventional two-view mammogram. Although the preliminary data from this device look very promising, larger clinical trials are needed to determine its clinical usefulness.

14.5 Prostate Imaging

Dedicated devices have also been developed for imaging the prostate (Huber et al. 2001; Turkington et al. 2004). The prostate is also a very challenging organ to image with conventional whole-body PET systems due to low tracer uptake and high uptake of tracer in the bladder.

Dedicated systems with higher efficiency have therefore been proposed and prototype systems have been developed. These are conceptually similar to some of the breast imaging devices in that they have a smaller detector separation to increase efficiency. These systems are still in a relatively early stage of development and the clinical utility has to be tested in clinical trials.

15 Summary

PET was for many years used primarily as an in vivo imaging tool limited to research studies of the brain and the heart. With the introduction of whole-body PET imaging and the combination of PET and CT into one imaging system, PET has become a diagnostic imaging tool that is routinely used in oncology, neurology, and cardiology. PET has also become widely accepted by researchers in preclinical research.

16 Cross-References

- Chapter 1, "Interactions of Particles and Radiation With Matter"
- Chapter 13, "Photon Detectors"
- Chapter 15, "Scintillation Counters"
- Chapter 16, "Semiconductor Counters"
- Chapter 17, "Gamma-Ray Detectors"
- Chapter 21, "New Solid State Detectors"
- Chapter 35, "Radiation-Based Medical Imaging Techniques: An Overview"
- Chapter 36, "CT Imaging: Basics and New Trends"
- Chapter 39, "Image Reconstruction"
- Chapter 41, "Quantitative Image Analysis in Tomography"
- Chapter 42, "Compartmental Modeling in Emission Tomography"
- Chapter 43, "Evaluation and Image Quality in Radiation-Based Medical Imaging"
- Chapter 44, "Simulation of Medical Imaging Systems: Emission and Transmission Tomography"
- Chapter 45, "High-Resolution and Animal Imaging Instrumentation and Techniques"

References

Allemand R, Gresset C et al (1980) Potential advantages of a cesium fluoride scintillator for a time-of-flight positron camera. J Nucl Med 21(2): 153–155

Bergström M, Bohm C et al (1980) Corrections for attenuation, scattered radiation, and random coincidences in a ring detector positron emission transaxial tomography. IEEE Trans Nucl Sci 27(1):549–554

Bergström M, Litton J et al (1982) Determination of object contour from projections for attenuation correction in cranial positron emission tomography. J Comput Assist Tomogr 6(2): 365–372

Berg WA, Weinberg IN et al (2006) High-resolution fluorodeoxyglucose positron emission tomography with compression ("positron emission mammography") is highly accurate in depicting primary breast cancer. Breast J 12(4):309–323

Beyer T, Kinahan PE et al (1994) The use of X-ray CT for attenuation correction of PET data. In: Nuclear Science Symposium and Medical Imaging Conference, vol 4. Norfolk, VA, pp 1573–1577

Boss A, Bisdas S et al (2010) Hybrid PET/MRI of intracranial masses: initial experiences and comparison to PET/CT. J Nucl Med 51(8):1198–1205

Bowen SL, Wu Y et al (2009) Initial characterization of a dedicated breast PET/CT scanner during human imaging. J Nucl Med 50(9):1401–1408

Brooks RA, Di Chiro G (1976) Principles of computer assisted tomography (CAT) in radiographic and radioisotopic imaging. Phys Med Biol 21(5): 689–732

Budinger TF (1983) Time-of-flight positron emission tomography: status relative to conventional PET. J Nucl Med 24(1):73–78

Buzhan P, Dolgoshein B et al (2003) Silicon photomultiplier and its possible applications. Nucl Instrum Methods Phys Res Sect A-Accelerat Spectrom Detect Associat Equip 504(1–3):48–52

Casey ME, Nutt R (1986) A multicrystal two dimensional BGO detector system for positron emission tomography. IEEE Trans Nucl Sci 33: 460–463

Casey ME, Gadakar H et al (1995) A component based method for normalization in volume PET. In: 3 rd international meeting on fully three-dimensional image reconstruction in radiology and nuclear medicine, Aix-les-Bains, France, pp 67–71

Cherry SR, Shao Y et al (1997) MicroPET: a high resolution PET scanner for imaging small animals. IEEE Trans Nucl Sci 44(3):1161–1166

Conti M, Bendriem B et al (2005) First experimental results of time-of-flight reconstruction on an LSO PET scanner. Phys Med Biol 50(19): 4507–4526

Cutler PD, Cherry SR et al (1992) Unique design features and performance of a new PET system for animal research. J Nucl Med 33(4):595–604

Czernin J, Phelps ME (2002) Positron emission tomography scanning: current and future applications. Annu Rev Med 53:89–112

Dahlbom M, Hoffman EJ (1987) Problems in signal-to-noise ratio for attenuation correction in high resolution PET. IEEE Trans Nucl Sci 34(1): 288–293

Dahlbom M, Schiepers C et al (2005) Comparison of noise equivalent count rates and image noise. IEEE Trans Nucl Sci 52(5):1386–1390

Doshi NK, Shao Y et al (2000) Design and evaluation of an LSO PET detector for breast cancer imaging. Med Phys 27(7):1535–1543

Evans RD (1955) The atomic nucleus. Krieger, Malabar

Hasegawa BH, Lang TF et al (1993) Object-specific attenuation correction of SPECT with correlated dual-energy X-Ray CT. IEEE Trans Nucl Sci 40(4):1242–1252

Hoffman EJ, Huang S-C et al (1979) Quantitation in positron emission computed tomography: I. Effect of object size. J Comput Assist Tomogr 3(3):299–308

Hoffman EJ, Huang S-C et al (1981) Quantitation in positron emission computed tomography: IV. Effect of accidental coincidences. J Comput Assist Tomogr 5(3):391–400

Hoffman EJ, Huang S-C et al (1982) Quantitation in positron emission computed tomography: VI. Effect of nonuniform resolution. J Comput Assist Tomogr 6(5):987–999

Hoffman EJ, Guerrero TM et al (1989) PET system calibration and corrections for quantitative and spatially accurate images. IEEE Trans Nucl Sci 36(1):1108–1112

Huang SC, Hoffman EJ et al (1979) Quantitation in positron emission computed tomography: 2. Effects of inaccurate attenuation correction. J Comput Assist Tomogr 3(6):804–814

Huang SC, Carson RE et al (1981) A boundary method for attenuation correction in positron computed tomography. J Nucl Med 22(7):627–637

Huber JS, Derenzo SE et al (2001) Conceptual design of a compact positron tomograph for prostate imaging. IEEE Trans Nucl Sci 48(4): 1506–1511

Hudson HM, Larkin RS (1994) Accelerated image reconstruction using ordered subsets of projection data. IEEE Trans Med Imaging 13(4): 601–609

Judenhofer MS, Wehrl HF et al (2008) Simultaneous PET-MRI: a new approach for functional and morphological imaging. Nat Med 14(4): 459–465

Kadrmas DJ, Casey ME et al (2009) Impact of time-of-flight on PET tumor detection. J Nucl Med 50(8):1315–1323

Karp JS, Surti S et al (2003) Performance of a brain PET camera based on anger-logic gadolinium oxyorthosilicate detectors. J Nucl Med 44(8):1340–1349

Kinahan PE, Townsend DW et al (1998) Attenuation correction for a combined 3D PET/CT scanner. Med Phys 25(10):2046–2053

Knoll GF (2010). Radiation Detection and Measurement. Wiley, New York

Lange K, Carson R (1984) EM reconstruction algorithms for emission and transmission tomography. J Comput Assist Tomogr 8(2):306–316

Lecomte R, Schmitt D et al (1985) Performance characteristics of BGO-silicon avalanche photodiode detectors for PET. IEEE Trans Nucl Sci 32(1):482–486

Levin CS, Hoffman EJ (1999) Calculation of positron range and its effect on the fundamental limit of positron emission tomography system spatial resolution. Phys Med Biol 44(3):781–799

Levin CS, Dahlbom M et al (1995) A Monte Carlo correction for the effect of Compton scattering in 3-D PET brain imaging. IEEE Trans Nucl Sci 42(4):1181–1185

Lin TW, de Aburto MA et al (2007) Predicting seizure-free status for temporal lobe epilepsy patients undergoing surgery: prognostic value of quantifying maximal metabolic asymmetry extending over a specified proportion of the temporal lobe. J Nucl Med 48(5):776–782

Lois C, Jakoby BW et al (2010) An assessment of the impact of incorporating time-of-flight information into clinical PET/CT imaging. J Nucl Med 51(2):237–245

Meikle SR, Dahlbom M et al (1993) Attenuation correction using count-limited transmission data in positron emission tomography. J Nucl Med 34(1):143–150

Moses WW, Ullisch M (2006) Factors influencing timing resolution in a commercial LSO PET camera. IEEE Trans Nucl Sci 53(1):78–85

Ollinger JM (1996) Model-Based Scatter Correction For Fully 3d PET. Phys Med Biol 41(1):153–176

Phelps ME, Hoffman EJ et al (1975) Application of annihilation coincidence detection to transaxial reconstruction tomography. J Nucl Med 16(3):210–224

Pichler BJ, Judenhofer MS et al (2006) Performance test of an LSO-APD detector in a 7-T MRI scanner for simultaneous PET/MRI. J Nucl Med 47(4):639–647

Ranger NT, Thompson CJ et al (1989) The application of a masked orbiting transmission source for attenuation correction in PET. J Nucl Med 30(6):1056–1068

Raylman RR, Majewski S et al (2008) The positron emission mammography/tomography breast imaging and biopsy system (PEM/PET): design, construction and phantom-based measurements. Phys Med Biol 53(3):637–653

Sanchez-Crespo A, Andreo P et al (2004) Positron flight in human tissues and its influence on PET image spatial resolution. Eur J Nucl Med Mol Imaging 31(1):44–51

Schindler TH, Schelbert HR et al (2010) Cardiac PET imaging for the detection and monitoring of coronary artery disease and microvascular health. JACC Cardiovasc Imaging 3(6): 623–640

Schmand M, Burbar Z et al (2007) Brain PET: first human tomograph for simultaneous (functional) PET and MR imaging. J Nucl Med meeting abstracts 48(MeetingAbstracts_2):45P

Shao Y, Cherry SR et al (1997) Development of a PET detector system compatible with MRI/NMR systems. IEEE Trans Nucl Sci 44(3): 1167–1171

Shepp LA, Vardi Y (1982) Maximum likelihood reconstruction for emission tomography. IEEE Tran Med Imaging 1(2):113–122

Siegel S, Dahlbom M (1992) Implementation and evaluation of a calculated attenuation correction for PET. IEEE Trans Nucl Sci 39:1117–1121

Silverman DH, Small GW et al (2001) Positron emission tomography in evaluation of dementia: regional brain metabolism and long-term outcome. JAMA 286(17):2120–2127

Strother SC, Casey ME et al (1990) Measuring PET scanner sensitivity: relating countrates to image signal-to-noise ratios using noise equivalent counts. IEEE Trans Nucl Sci 37(2):783–788

Surti S, Kuhn A et al (2007) Performance of Philips Gemini TF PET/CT scanner with special consideration for its time-of-flight imaging capabilities. J Nucl Med 48(3):471–480

Tai YC, Wu H et al (2008) Virtual-pinhole PET. J Nucl Med 49(3):471–479

Turkington TG, Hawk TC et al (2004) PET prostate imaging with small planar detectors. In: IEEE nuclear science symposium and medical imaging conference, vol 5. Rome, Italy, pp 2806–2809

Wang GC, Huber JS et al (2006) Characterization of the LBNL PEM camera. IEEE Trans Nucl Sci 53(3):1129–1135

Watson CC, Newport D et al (1997) Evaluation of simulation-based scatter correction for 3-D PET cardiac imaging. IEEE Trans Nucl Sci 44(1): 90–97

Weber WA, Grosu AL et al (2008) Technology insight: advances in molecular imaging and an appraisal of PET/CT scanning. Nat Clin Pract Oncol 5(3):160–170

Weinberg I, Majewski S et al (1996) Preliminary results for positron emission mammography: real-time functional breast imaging in a conventional mammography gantry. Eur J Nucl Med 23(7):804–806

Wienhard K, Schmand M et al (2002) The ECAT HRRT: performance and first clinical application

of the new high resolution research tomograph. IEEE Trans Nucl Sci 49(104–110)

Williams CW, Crabtree MC et al (1979) Design and performance characteristics of a positron emission computed axial tomograph – ECAT-II. Trans Nucl Sci 26(1):619–627

Williams CW, Crabtree MC et al (1981) Design of the Neuro-ECAT: a high-resolution, high efficiency positron tomograph for imaging the adult head or infant torso. IEEE Trans Nucl Sci 28(2): 1736–1740

Wong WH (1993) A positron camera detector design with cross-coupled scintillators and quadrant sharing photomultipliers. IEEE Trans Nucl Sci 40(4):962–966

Xu EZ, Mullani N et al (1991) A segmented attenuation correction for PET. J Nucl Med 32(1): 161–165

39 Image Reconstruction

Claude Comtat
CEA, Orsay, France

1 Introduction 975

2 Analytical Reconstruction Algorithms 977
2.1 Scanning and Reconstruction Geometry 978
2.2 X-ray and Radon Transforms 979

3 2D Analytical Reconstruction 980
3.1 The 2D X-Ray Transform and Its Dual 981
3.2 Parallel-Beam Filtered Backprojection 983
3.2.1 Discretization 983
3.2.2 Ill-Posedness 984
3.3 Fan-Beam Filtered Backprojection 984
3.3.1 Flat Detector 986
3.3.2 Curved Detector 987
3.3.3 Short Scan 987
3.3.4 Helical Fan Beam 987

4 3D Analytical Reconstruction 988
4.1 Parallel-Beam Geometry for 3D PET Systems 988
4.1.1 The Reprojection Algorithm 989
4.1.2 Rebinning Techniques 990
4.2 Cone-Beam Reconstruction 990
4.2.1 The Feldkamp, Davis, and Kress Algorithm 991
4.2.2 Exact Reconstruction Algorithms 991

5 Iterative Reconstruction Algorithms 992
5.1 Scanning Model 993
5.2 Objective Function and Minimization Algorithm 995
5.2.1 The EM-ML Algorithm in Emission Tomography 996
5.2.2 ML Algorithms in Transmission Tomography 999
5.3 Regularization 1000
5.3.1 Early Termination 1000

C. Grupen, I. Buvat (eds.), *Handbook of Particle Detection and Imaging*,
DOI 10.1007/978-3-642-13271-1_39, © Springer-Verlag Berlin Heidelberg 2012

5.3.2 Post-Reconstruction Smoothing 1000
5.3.3 Penalized Objective-Function or MAP Reconstruction 1001

6 Concluding Remarks 1003

7 Cross-References 1004

References 1004

Abstract: This chapter describes the mathematical algorithms that are commonly used to reconstruct a three-dimensional image of the body being scanned in computed tomography. It covers emission and transmission tomography with electromagnetic ionizing radiation for three radiological modalities: positron emission tomography (PET), single photon emission computed tomography (SPECT), and X-ray computed tomography (X-ray CT). Analytical reconstruction algorithms are presented first in two dimensions for parallel and diverging rays. Then, the extension to three-dimensional analytical algorithms is described. Three-dimensional scanning is characterized by truncated measured data, requiring specific adaptation of the reconstruction algorithms. Finally, iterative algorithms are presented. A detailed presentation is given of the two most common reconstruction algorithms, the analytical filtered-backprojection algorithm (FBP) and the iterative expectation maximization – maximum-likelihood algorithm (EM-ML).

1 Introduction

In medical imaging with X-ray or gamma radiation, two different types of image are produced: planar and cross-sectional images of the body (*Fig. 1*). In planar imaging, it is not possible to localize in the depth of the patient the structures seen in the image. To produce cross-sectional images of the patient, it is necessary to satisfy geometrical constraints during the scanning of the patient and to use tomographic algorithms to reconstruct the images. This is referred to as computed tomography. This chapter is about such tomographic algorithms.

Tomographic image reconstruction is an inverse problem: given a scanning model, reconstruct from the measurements the spatial distribution of the quantity of interest. There are two computed tomography applications using X-ray or gamma radiation: emission and transmission tomography. In emission computed tomography, the radiation source is inside the patient: a radiopharmaceutical labeled with a gamma- (SPECT) or positron- (PET) emitting radionuclide is injected to the patient. The quantity to be reconstructed is the activity concentration of the radiopharmaceutical inside the body. In transmission computed tomography, the radiation source is outside the patient and rotates around the patient. The radiation source can be an X-ray tube (X-ray CT) or a long-lived radionuclide. The quantity to be reconstructed is the photon linear attenuation coefficient of the body. Radionuclide-based transmission tomography was used on emission tomography systems in nuclear medicine to correct the emission scan for attenuation. Today, an X-ray CT is usually attached to a PET or SPECT system and used for attenuation correction.

The factors limiting the diagnostic utility of tomographic images are quite different between emission imaging and X-ray CT. In emission tomography or transmission tomography with a radionuclide source, the measured data are blurred and very noisy. As a consequence, the effective spatial resolution of these systems is of the order of half a centimeter to one centimeter. Because of the high level of stochastic noise, there is no benefit for developing systems for clinical applications with a much finer spatial resolution. In X-ray CT, the flux of photons is much higher, and stochastic noise is less of an issue. The effective spatial resolution is below

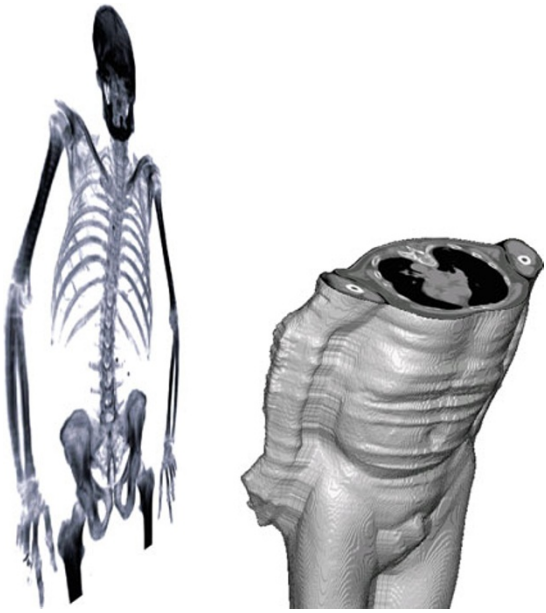


Fig. 1
Planar X-ray image (*left*) and cross-sectional image in X-ray CT (*right*)

one millimeter. The reading of the X-ray CT images is also different: they are usually viewed at a compressed gray scale, looking for small contrast. The expectations from a tomographic image-reconstruction algorithm are different between a PET or SPECT and an X-ray CT application. It is important to evaluate the "quality" of an image-reconstruction algorithm according to the task that is performed on the reconstructed image. Some algorithms may be more appropriate for tumor detection tasks but not optimal for a quantitative estimation task like tracer kinetic modeling.

The choice of the reconstruction algorithm plays an important role in the overall performance of any computed tomography system (Qi and Leahy 2006). The key element in a tomographic image-reconstruction algorithm is the scanning model that is inverted. In a first approximation, X-rays and gammas are assumed to propagate along straight lines and the acquired data provide us with the line integrals of the function to be reconstructed. This line integral model can be analytically inverted and an image reconstructed under certain scanning geometry conditions. The scanning model can be more complex to account for physical effects such as the stochastic nature of gamma emission or for the spatial resolution properties of the detectors. If the scanning model becomes too complex for an analytical inversion, iterative techniques are used to reconstruct an image. Iterative image-reconstruction algorithms are currently a very active research field as they allow for accurate scanning modeling. Both analytical and iterative reconstruction techniques are described in this chapter. Many image-reconstruction algorithms are proposed in the specialized publications. Only conventional algorithms will be presented.

2 Analytical Reconstruction Algorithms

Analytical reconstruction techniques in emission and transmission computed tomography are based on the line integral model. The quantity to be reconstructed is modeled as a continuous function $f(\mathbf{x})$ on the support $\mathfrak{R} \subset \mathbb{R}^3$ (the reconstructed field of view). In the following, $f(\mathbf{x})$ is assumed not to depend on time. If $f(\mathbf{x})$ varies with time, the scanning period is subdivided into temporal frames. Each frame is reconstructed independently, and $f(\mathbf{x})$ is supposed static over the duration of the frame.

In transmission tomography, $f(\mathbf{x})$ represents the photon linear attenuation coefficient of the tissue at location $\mathbf{x}$. Let L be a beam of X-rays along a straight line between the X-ray tube and one detection element, I_0 the initial intensity of the beam, and I its intensity after having crossed the body (◉ *Fig. 2, left*). The line integral model and the Beer–Lambert law state that

$$-\ln\left(\frac{I}{I_0}\right)_{\text{X-ray CT}} = \int_L f(\mathbf{x})\,\mathrm{d}\mathbf{x} = g(L). \tag{1}$$

In emission tomography, $f(\mathbf{x})$ represents the activity concentration of the radiopharmaceutical inside the body at point $\mathbf{x}$. Let L be a straight line originating from one detection element of a gamma camera (◉ *Fig. 2, right*). The detector head is collimated so as to accept only gammas traveling along the line L. In SPECT, for a fixed detector head position, the number I of detected gammas per unit of time for line L is modeled as

$$I_{\text{SPECT}} = \int_L f(\mathbf{x}) \mathrm{e}^{-\int_{L(\mathbf{x})} \mu(\mathbf{y})\,\mathrm{dy}}\,\mathrm{d}\mathbf{x} = g(L), \tag{2}$$

where $\mu(\mathbf{y})$ represents the gamma attenuation coefficient and $L(\mathbf{x})$ the section of L between $\mathbf{x}$ and the detection element. This model is referred to as the attenuated line integral model. In the following, for the sake of simplicity, we will ignore the effect of attenuation in SPECT and have

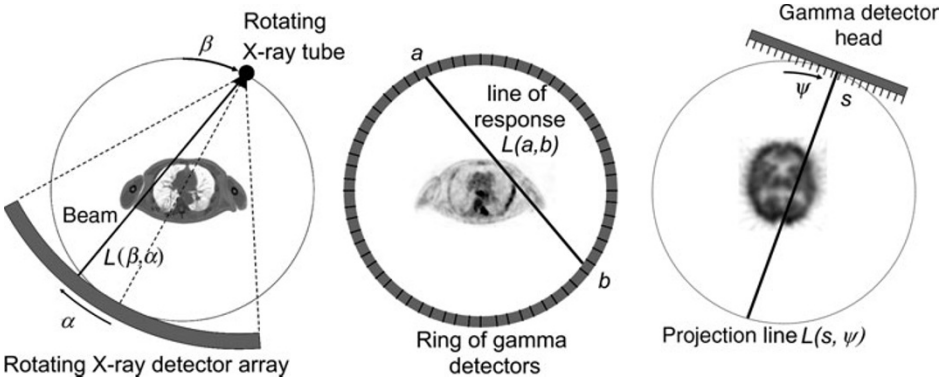


Fig. 2

Transaxial views of a X-ray CT (*left*), a PET (*middle*), and a SPECT (*right*) tomograph. The detectors are repeated linearly along the axial direction, orthogonal to the figure. In X-ray CT, the line *L* is parametrized by the angular position β of the source and the fan angle α of the detector. In PET, the line *L* is defined by the locations *a* and *b* of the pair of detectors in coincidence. In SPECT, the line *L* is parametrized by the angular position ψ of the detector head and the position *s* of the detection element

$g(L) = \int_L f(\mathbf{x})\,\mathrm{d}\mathbf{x}$. It should be noted that exact analytical image-reconstruction algorithms for the attenuated line integral model have been developed (Novikov 2002). They will not be presented in this chapter.

Let L be the line of response (LOR) between two detectors in coincidence in a PET scanner (Fig. 2, *middle*). The number I of coincidences per unit of time for line L is modeled in PET as

$$I_{\mathrm{PET}} = \mathrm{e}^{-\int_L \mu(\mathbf{x})\,\mathrm{d}\mathbf{x}} \int_L f(\mathbf{x})\,\mathrm{d}\mathbf{x}. \tag{3}$$

Unlike SPECT, the effect of attenuation in PET can be corrected for prior to reconstruction. The model can be expressed as

$$I_{\mathrm{PET}} \cdot \mathrm{e}^{+\int_L \mu(\mathbf{x})\,\mathrm{d}\mathbf{x}} = \int_L f(\mathbf{x})\,\mathrm{d}\mathbf{x} = g(L). \tag{4}$$

As a result of these models, the scanning process in emission and transmission computed tomography provides us with the line integrals $\int_L f(\mathbf{x})\,\mathrm{d}\mathbf{x}$ of the function f. In order to satisfy the line integral model, measured data are corrected prior to reconstruction for background events (scattered photons and random coincidences in PET) and nonuniform detection efficiency. In the following, the measured data $g(L)$ will be assumed to be corrected for such effects and expressed such as $g(L) = \int_L f(\mathbf{x})\,\mathrm{d}\mathbf{x}$.

2.1 Scanning and Reconstruction Geometry

In a real tomograph, the detection elements are not continuous, and the detectors have a limited spatial extent. As a result, the number of measured lines is finite. The design of the tomograph determines the organization of the measured lines. There are three basic geometries in use with parallel or divergent lines:

- Parallel geometry: ensemble of equally spaced parallel lines, for a number of equally distributed directions. Used in PET and SPECT with parallel hole collimator.
- Fan-beam geometry: ensemble of divergent lines in a 2D plane originating from a single point (fan vertex) for a number of equally distributed point positions along a circle or a line. Used in X-ray CT when the data are recorded by a linear detector and in SPECT with 2D convergent hole collimator.
- Cone-beam geometry: ensemble of divergent lines in a 3D cone originating from a single point (cone vertex) for a number of equally distributed point positions along a cylinder or a planar detector. Used in X-ray CT when the data are recorded by a 2D detector array (multi-row detector) and in SPECT with 3D convergent hole collimator.

In the following, unless specified, a parallel-beam geometry is assumed by default.

A fixed Cartesian coordinate system $(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$ and a rotating coordinate system $(\mathbf{e}_\theta, \mathbf{e}_{\theta_1^\perp}, \mathbf{e}_{\theta_2^\perp})$ are defined in Fig. 3. The unit vector $\mathbf{e}_3$ is oriented along the axial axis of the tomograph, the direction of the patient bed. It corresponds to the rotation axis for SPECT and X-ray CT, and to the direction perpendicular to the detector rings for PET. The plane defined by $(\mathbf{e}_1, \mathbf{e}_2)$ is called the transaxial plane. The unit vector $\mathbf{e}_\theta$ defines the orientation of the measured lines. It is defined by the co-polar angle ϕ and the azimuthal angle ψ. The projection plane defined by $\left(\mathbf{e}_{\theta_1^\perp}, \mathbf{e}_{\theta_2^\perp}\right)$ is noted $\theta^\perp$. For clarity, a point in $\mathbb{R}^3$ will be noted $\mathbf{x}$, and a point in $\theta^\perp$

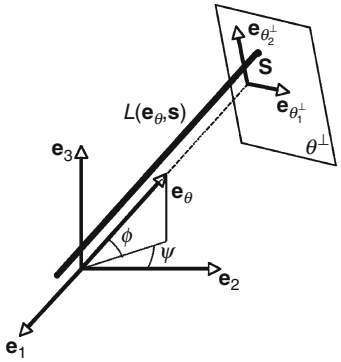


Fig. 3
The orthonormal basis $\mathbf{e}_1$, $\mathbf{e}_2$ and $\mathbf{e}_3$ defines a Cartesian coordinate system in $\mathbb{R}^3$, fixed relative to the patient bed. The orthonormal basis $\mathbf{e}_\theta$, $\mathbf{e}_{\theta_1^\perp}$, and $\mathbf{e}_{\theta_2^\perp}$ defines a coordinate system for a parallel-beam geometry, with $\mathbf{e}_{\theta_1^\perp}$ orthogonal to $\mathbf{e}_3$

will be noted $\mathbf{s}$. A line L is parametrized by its orientation $\mathbf{e}_\theta$ and its intersection $\mathbf{s}$ with the projection plane $\theta^\perp$.

If the co-polar angle ϕ of all measured lines is equal to zero, then each transaxial plane of the image can be reconstructed independently from the others. The reconstruction reduced to a two-dimensional (2D) problem. The three-dimensional volume is obtained by stacking the transaxial slices. If lines with different co-polar orientations are measured, then all line integrals have to be processed simultaneously. The reconstruction is a three-dimensional (3D) problem: the 3D image cannot be decomposed into 2D planes that are reconstructed independently. 2D tomography refers to single-row X-ray CT scanners (fan-beam geometry), SPECT scanners operated with parallel-hole collimator (parallel geometry) or convergent fan-beam collimator (fan-beam geometry), and PET scanners operated with annular collimator (septa) between the detector rings (parallel geometry). 3D tomography refers to multi-row X-ray CT (cone-beam geometry), SPECT scanners operated with convergent cone-beam collimator or pinhole collimator (cone-beam geometry), and PET scanners operated without annular collimator (parallel geometry). 3D PET imaging allows for increased sensitivity and cone-beam X-ray CT allows for faster acquisition time.

2.2 X-ray and Radon Transforms

The X-ray transform and the Radon transform play a central role in computed tomography. Both transforms can be generalized for any n-dimensional space, with n superior to one, and they coincide for n equals to 2.

The n-dimensional X-ray transform $\mathbf{P}$ maps a function f on $\mathbb{R}^n$ into the sets of its line integrals (Natterer 2001). For a line oriented along $\mathbf{e}_\theta \in \mathbb{S}^{n-1}$ (unit sphere in $\mathbb{R}^n$) and defined by the point $\mathbf{s}$ in the subspace $\theta^\perp$ orthogonal to $\mathbf{e}_\theta$ ($\mathbf{s} \cdot \mathbf{e}_\theta = 0$), we have

$$\mathbf{P}f(\mathbf{e}_\theta, \mathbf{s}) = \int_{-\infty}^{+\infty} f(\mathbf{s} + t\mathbf{e}_\theta)\, \mathrm{d}t, \tag{5}$$

where the symbol $\cdot$ denotes the scalar product between two vectors. $\mathbf{P}f(\mathbf{e}_\theta, \mathbf{s})$ is a function on the manifold $\mathbb{T}^{(2n-2)} = \{(\mathbf{e}_\theta, \mathbf{s}) | \mathbf{e}_\theta \in \mathbb{S}^{n-1}, \mathbf{s} \in \theta^\perp\}$ of lines in $\mathbb{R}^n$. The scanning process corresponds to the X-ray transform of $f(\mathbf{x})$. This transform also corresponds to the linear forward projection operator. The reconstruction consists in the inversion of the X-ray transform.

The n-dimensional Radon transform $\mathbf{R}$ maps a function f on $\mathbb{R}^n$ into the sets of its integrals over the hyperplanes (Natterer 2001). For a hyperplane Π normal to the unit vector $\mathbf{e}_\Pi$ and defined by the signed distance l from the origin along $\mathbf{e}_\Pi$, we have

$$\mathbf{R}f(\mathbf{e}_\Pi, l) = \int_{\mathbf{x}\cdot\mathbf{e}_\Pi = l} f(\mathbf{x})\, \mathrm{d}\mathbf{x}. \tag{6}$$

In the 2D case ($n = 2$), the Radon and the X-ray transforms coincide, with a change in the notation of the arguments: the vector $\mathbf{e}_\Pi$ for the Radon transform is normal to the vector $\mathbf{e}_\theta$ of the X-ray transform and corresponds to $\mathbf{e}_{\theta^\perp}$. The analytical 2D reconstruction is based on the inversion of the 2D Radon transform. In the 3D case, the Radon transform integrates over all points in a 2D plane. This does not correspond to the scanning process in 3D computed tomography. Nonetheless, the 3D Radon transform plays a role in cone-beam reconstruction.

The Fourier slice theorem is a key element for defining an analytical reconstruction in computed tomography. For the n-dimensional X-ray transform, the $(n-1)$-dimensional Fourier transform of $\mathbf{P}f(\mathbf{e}_\theta, \mathbf{s})$ with respect to $\mathbf{s} \in \theta^\perp$ is given by

$$\mathcal{F}_{n-1}\{\mathbf{P}f\}(\mathbf{e}_\theta, \boldsymbol{\nu}_\perp) = \int_{\theta^\perp} \mathrm{e}^{-2\pi i \mathbf{s}\cdot\boldsymbol{\nu}_\perp} \mathbf{P}f(\mathbf{e}_\theta, \mathbf{s})\, \mathrm{d}\mathbf{s}, \tag{7}$$

with $\boldsymbol{\nu}_\perp \in \theta^\perp$ ($\boldsymbol{\nu}_\perp \cdot \mathbf{e}_\theta = 0$). The n-dimensional Fourier transform of $f(\mathbf{x})$ is given by

$$\mathcal{F}_n\{f\} = \int \mathrm{e}^{-2\pi i \mathbf{x}\cdot\boldsymbol{\nu}} f(\mathbf{x})\, \mathrm{d}\mathbf{x}. \tag{8}$$

From the definition of the X-ray transform and the Fourier transform, one can deduct the Fourier slice theorem

$$\mathcal{F}_{n-1}\{\mathbf{P}f\}(\mathbf{e}_\theta, \boldsymbol{\nu}_\perp) = \mathcal{F}_n\{f\}(\boldsymbol{\nu} = \boldsymbol{\nu}_\perp). \tag{9}$$

The $(n-1)$-dimensional Fourier transform of the n-dimensional X-ray transform for a given $\mathbf{e}_\theta$ samples the n-dimensional Fourier transform of $f(\mathbf{x})$ on the frequency plane through the origin and normal to $\mathbf{e}_\theta$. This theorem shows the possibility to reconstruct the image f from its line integrals if they are measured such as to sample entirely the n-dimensional Fourier space of f.

3 2D Analytical Reconstruction

A detailed presentation of analytical reconstruction in 2D computed tomography can be found in the seminal book of F. Natterer (2001).

When all line integrals are perpendicular to the tomograph axis $\mathbf{e}_3$, each line contributes only to one transaxial slice. Therefore, each transaxial slice can be reconstructed independently with a 2D algorithm.

Let the orthonormal vectors $(\mathbf{e}_1, \mathbf{e}_2)$ and $(\mathbf{e}_\theta, \mathbf{e}_{\theta^\perp})$ define the 2D coordinate systems in $\mathbb{R}^2$. $\theta^\perp$ is the line perpendicular to $\mathbf{e}_\theta$ passing through the origin O. Points in $\mathbb{R}^1$ along $\theta^\perp$ will be noted $s = \mathbf{s} \cdot \mathbf{e}_{\theta^\perp}$. A line L is defined on $\mathbb{T}^2$ and is parametrized by its azimuthal angle ψ and its

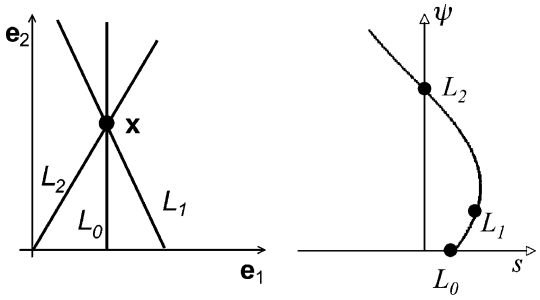


Fig. 4
All lines *L* crossing a point x in $\mathbb{R}^2$ (*left*) define a sinus curve in $\mathbb{T}^2$ (*right*)

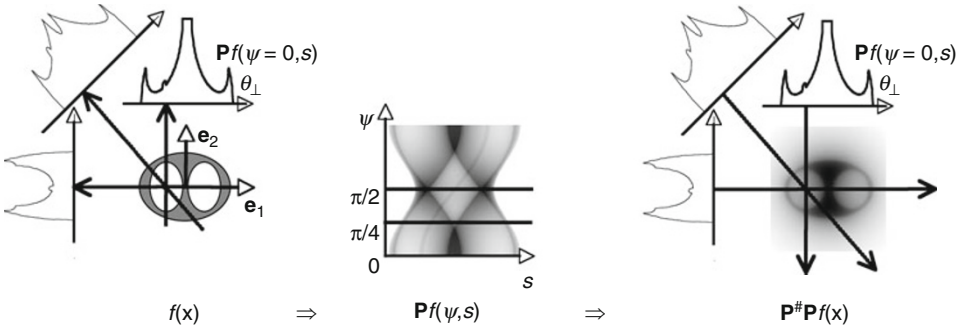


Fig. 5
The projection of the image $f(\mathbf{x})$, followed by its backprojection, does not result in the original image

signed radial distance s: $L(\psi, s)$. All points $\mathbf{x}$ that belong to $L(\psi, s)$ satisfy the relation $\mathbf{x}\cdot\mathbf{e}_{\theta^\perp} = s$. In 2D computed tomography, $\mathbb{T}^2$ is called a sinogram. All lines L that pass through the point $\mathbf{x}$ in $\mathbb{R}^2$ describe a sinus curve in $\mathbb{T}^2$, as illustrated in ❯ *Fig. 4*.

3.1 The 2D X-Ray Transform and Its Dual

The X-ray and the Radon transforms coincide in 2D. They map a function $f(\mathbf{x})$ on $\mathbb{R}^2$ into $\mathbb{T}^2$. They are defined by the 1D projection operator

$$\mathbf{P}f(\psi, s) = \int_{\mathbf{x}\cdot\mathbf{e}_{\theta^\perp} = s} f(\mathbf{x})\, \mathrm{d}\mathbf{x} = \int_{-\infty}^{+\infty} f(s\cos\psi - t\sin\psi, s\sin\psi + t\cos\psi)\, \mathrm{d}t, \qquad (10)$$

with s a signed variable. $\mathbf{P}f(\psi, s)$ is an even function: $\mathbf{P}f(\mathbf{e}_\theta, s) = \mathbf{P}f(-\mathbf{e}_\theta, -s)$. ❯ *Figure 5* illustrates the relation between an image f and its X-ray transform. From the line integral model, it follows that the 2D scanning process of f provides us with its 2D X-ray transform: $g(\psi, s) = \mathbf{P}f(\psi, s)$, where $g(\psi, s)$ are the measured data. The image-reconstruction algorithm is based on the inversion of the X-ray transform.

The Fourier slice theorem states in 2D that the 1D Fourier transform of $\mathbf{P}f(\psi, s)$ with respect to s is equivalent to the 2D Fourier transform of f along $\mathbf{e}_{\theta^\perp}$. As a consequence, the

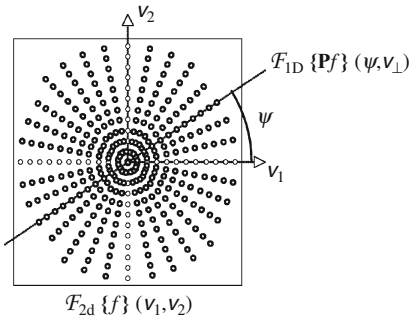


Fig. 6
Polar sampling of the 2D Fourier transform of the image *f* provided by the 1D Fourier transform of the X-ray transform of *f* over $[0, \pi[$

2D Fourier space of f is entirely known if the line integrals are measured over $[0, \pi[$ as illustrated in Fig. 6. In other words, f can be reconstructed if the line integrals of f are measured over $[0, \pi[$. If the azimuthal range of the measured lines is limited to $[\psi_0, \psi_1[$ with $\psi_0 > 0$ and $\psi_1 < \pi$, the 2D Fourier space of the image will not be entirely covered: the measured data are incomplete and the reconstruction of these data will lead to an incoherent image. For this reason, the detector head in SPECT and the X-ray tube in CT rotate around the patient. The 2D Fourier slice theorem suggests one direct method to reconstruct f: build the 2D Fourier transform of f for each $\mathbf{e}_{\theta^\perp}$ by taking the 1D Fourier transform of $\mathbf{P}f(\psi, s)$ with respect to s, and then take the inverse 2D Fourier transform. This algorithm is called the *direct Fourier reconstruction*. The difficulty with this algorithm comes from the polar discretization of the measured data: in order to use a fast Fourier transform (FFT) to compute the 2D inverse Fourier transform of f, a Cartesian 2D grid has to be interpolated from the polar grid.

The dual operator of the 2D X-ray transform is the backprojection operator, noted $\mathbf{P}^\#$. It maps a function $g(\psi, s)$ on $\mathbb{T}^2$ into $\mathbb{R}^2$. The value of $g(\psi, s)$ is backprojected over the image space along the corresponding line (ψ, s):

$$\mathbf{P}^\# g(\mathbf{x}) = \int_{\mathbb{S}^1} g(\psi, \mathbf{x}\cdot\mathbf{e}_{\theta^\perp})\,\mathrm{d}\psi = \int_0^{2\pi} g(x_1\cos\psi + x_2\sin\psi, \psi)\,\mathrm{d}\psi. \tag{11}$$

The projection of f followed by its backprojection is not equal to f ($\mathbf{P}^\#\mathbf{P}f(\mathbf{x}) \neq f(\mathbf{x})$), as illustrated in Fig. 5. For f on $\mathbb{R}^2$, we have

$$\mathbf{P}^\#\mathbf{P}f(\mathbf{x}) = \frac{1}{|\mathbf{x}|} ** f(\mathbf{x}), \tag{12}$$

where $**$ is the 2D convolution product and $|\cdot|$ the euclidean norm (see Natterer 2001, Theorem II.1.5) In the Fourier domain, noting that the 2D Fourier transform of $\frac{1}{|\mathbf{x}|}$ is $\frac{1}{|\boldsymbol{\nu}|}$, Eq. 12 becomes

$$\mathcal{F}_{2D}\{f(\mathbf{x})\}(\boldsymbol{\nu}) = \mathcal{F}_{2D}\{\mathbf{P}^\#\mathbf{P}f\}(\boldsymbol{\nu})|\boldsymbol{\nu}|. \tag{13}$$

This relation suggests a four-step 2D algorithm to reconstruct f from the measured projections $g = \mathbf{P}f$: (1) backprojection of g, (2) 2D Fourier transform of $\mathbf{P}^\# g$, (3) multiplication by the filter $|\boldsymbol{\nu}|$, and (4) inverse 2D Fourier transform. In practice, another algorithm is usually preferred,

the *Filtered-BackProjection* (FPB) algorithm. In this algorithm, the ordering of the filtering step and the backprojection step is inverted.

3.2 Parallel-Beam Filtered Backprojection

In a real scanner, the line integrals are not measured continuously. Given the radial sampling Δs of the scanner, the maximum frequency that can be recovered without aliasing is $1/(2\Delta s)$ as specified by Shannon's sampling theory. As a consequence, the image $W_c ** f$ is reconstructed in practice, where W_c is a low-pass filter with cutoff frequency $\nu_c \leq 1/2\Delta s$,

$$\mathcal{F}_2\{W_c\}(\boldsymbol{\nu}) = \Phi\left(\frac{|\boldsymbol{\nu}|}{\nu_c}\right), \tag{14}$$

with $0 \leq \Phi(|\boldsymbol{\nu}|/\nu_c) \leq 1$ and $\Phi(|\boldsymbol{\nu}|/\nu_c) = 0$ for $|\boldsymbol{\nu}|/\nu_c \geq 1$. The filtered-backprojection algorithm is based on the following relation (Theorem II.1.3 in Natterer (2001)):

$$f_{\mathrm{FBP}} = W_c ** f = \mathbf{P}^{\#}(w_c * \mathbf{P}f), \tag{15}$$

where $w_c \in \mathbb{T}^2$, $\mathbf{P}^{\#} w_c = W_c$, and $*$ is the 1D convolution product with respect to s. From theorem II.1.4 in Natterer (2001), we infer for w_c

$$\mathcal{F}_1\{w_c\}(\psi, \nu_\perp) = \frac{1}{2}|\nu_\perp|\Phi\left(\frac{|\nu_\perp|}{\nu_c}\right). \tag{16}$$

The expression of the filter w_c in the 1D Fourier domain is given by the product of the ramp filter $|\nu_\perp|$ with a low-pass filter with cutoff frequency ν_c. It does not depend on the azimuthal angle ψ. The evaluation of $f_{\mathrm{FBP}} = W_c ** f$ with the filtered-backprojection algorithm is given by the one-dimensional convolution or filtering of the projections $g(\psi, s) = Pf$ with w_c for each direction ψ, followed by the backprojection of the filtered projections

$$f_{\mathrm{FBP}}(\mathbf{x}) = \frac{1}{2}\int_0^{2\pi} g^*(\psi, s(\mathbf{x}, \psi))\, \mathrm{d}\psi, \tag{17}$$

where

$$s(\mathbf{x}, \psi) = \mathbf{x} \cdot \mathbf{e}_{\theta_\perp}(\psi). \tag{18}$$

The filtered projections are given by

$$g^*(\psi, s) = g(\psi, s) * w_c(s). \tag{19}$$

3.2.1 Discretization

For a given 2D-scanning parallel geometry, with a circular field of view of radius r, the measured data $g = \mathbf{P}f$ are sampled at $\{(\psi_j, s_k) | j = 1, \ldots, p; k = -q, \ldots, +q\}$ with $\psi_j = \pi(j-1)/p$, $s_k = \Delta s\, k$, and $\Delta s = r/q$. The filtered backprojection with linear interpolation and discrete quadrature is given explicitly by the following algorithm:

for $j = 1$ to p **do**
  for $k = -q$ to q **do**
    $g^*_{j,k} = \Delta s \sum_{l=-q}^{q} w_c(s_k - s_l) g_{j,l}$
  end for
  for each reconstruction point **x** **do**
    $k = \text{truncate}\{\mathbf{e}_{\theta^\perp} \cdot \mathbf{x}/\Delta s\}$
    $u = \mathbf{e}_{\theta^\perp} \cdot \mathbf{x}/\Delta s - k$
    $f_{\text{FBP}}(\mathbf{x}) = f_{\text{FBP}}(\mathbf{x}) + \frac{\pi}{p}\left((1-u)\, g^*_{j,k} + u g^*_{j,k+1}\right)$
  end for
end for

If the low-pass filter $\Phi\left(\frac{|\nu_\perp|}{\nu_c}\right)$ is a rectangular function with cutoff ν_c, the filter $w_c(s)$ is expressed as

$$w_c(s) = \begin{cases} \frac{\nu_c^2}{4\pi^2}\left(\frac{\cos s\nu_c - 1}{(s\nu_c)^2} + \frac{\sin s\nu_c}{s\nu_c}\right) & s \neq 0 \\ \frac{\nu_c^2}{8\pi^2} & s = 0 \end{cases} . \tag{20}$$

Alternatively, the calculation of the discrete convolution can be performed using FFTs.

3.2.2 Ill-Posedness

The inversion of the X-ray transform by the FBP algorithm is an *ill-posed problem*. The solution $f_{\text{FBP}} = \mathbf{P}^{\#}(w_c * g)$ does not depend continuously on the data g. In other words, an arbitrary small perturbation of g might cause an arbitrary large error in f_{FBP}. In computed tomography, perturbations in the signal typically occur because of the stochastic noise in the measurements. This is particularly true in emission tomography, where the number of detected events per projection element is very low. This feature can be intuitively understood by noting the effect of the ramp filter $|\nu_\perp|$ on the noise. Usually, the image to be reconstructed is essentially band-limited: its Fourier transform or power spectral density is near zero above a certain finite frequency. In contrast, noise has a much higher power spectral density at high frequencies. The multiplication of the Fourier transform of g with the ramp filter exacerbates the impact of noise at high frequency. An illustration of the effect of noisy projections on the reconstructed image is given in ❯ *Fig. 7*. To limit the effect of noise, regularization can be achieved by adjusting the cutoff frequency ν_c and the shape of the low-pass filter $\Phi(|\nu_\perp|/\nu_c)$. Rather than a rectangular function, a smoother low-pass filter is chosen, like the Hann apodization window $\Phi(|\nu_\perp|/\nu_c) = 1/2\,(1 + \cos(|\nu_\perp|/\nu_c))$. An illustration of the effect of the low-pass filter $\Phi(|\nu_\perp|/\nu_c)$ on the quality of the image is given in ❯ *Fig. 8*. The price to pay for this reduction in noise is a degraded spatial resolution. The choice of the apodization window depends on the application and is a compromise between the limitation of the noise amplification introduced by the reconstruction algorithm and a degradation of the spatial resolution.

3.3 Fan-Beam Filtered Backprojection

When the scanning geometry measures diverging rays instead of parallel rays, two approaches are possible: either the data are rebinned into parallel-beam projections, followed by a parallel-beam filtered-backprojection reconstruction, or a dedicated fan-beam filtered-backprojection

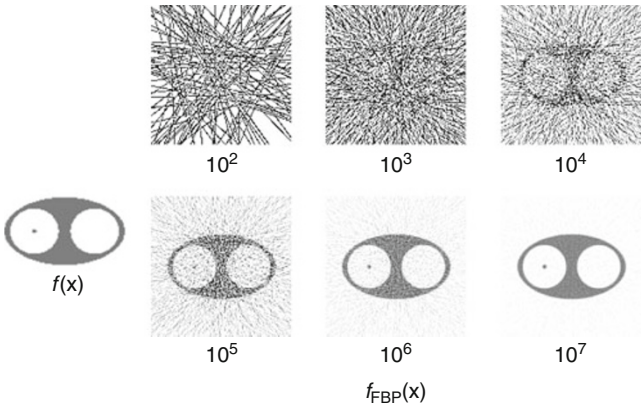


Fig. 7

The effect of the number of detected photons on the quality of the reconstructed image. A synthetic 2D image *f* was projected and Poisson noise was added to the projections to simulate various levels of detected counts (from 10^2 to 10^7). The noisy projections were reconstructed with the FBP algorithm and a rectangular low-pass filter $\Phi(\nu_\perp)$ with Nyquist cutoff frequency $1/(2\Delta s)$. The *streak* artifact typical of FBP images is particularly visible in the low-count images. The highest-count images are representative of X-ray CT, and the noise of the lower-count images is typical of emission tomography

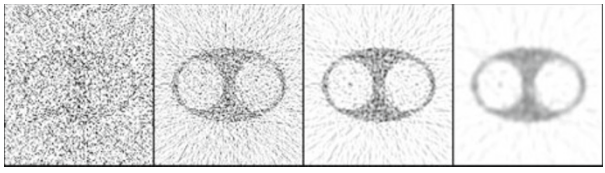


Fig. 8

The effect of the regularization on the "quality" of the reconstructed image. The same simulated data set was reconstructed with a rectangular low-pass filter with Nyquist cutoff frequency (*left*, no regularization), and with the Hann apodization window and a cutoff frequency varying between the Nyquist frequency ($1/(2\Delta s)$) and one fifth of the Nyquist frequency ($1/(10\Delta s)$). Some amount of regularization is obviously needed to get an "acceptable" image

reconstruction algorithm is used. In the first case, bilinear interpolations are required for the rebinning. In the second case, the native scanning geometry is used for the reconstruction.

In 2D fan-beam scanning geometry, the beam vertex (the source for transmission tomography or the focal point of the collimator for SPECT) moves along a circle of radius R, bigger than the reconstructed field-of-view radius. Its position is given by

$$\mathbf{a}(\beta) = R\mathbf{e}_a(\beta), \tag{21}$$

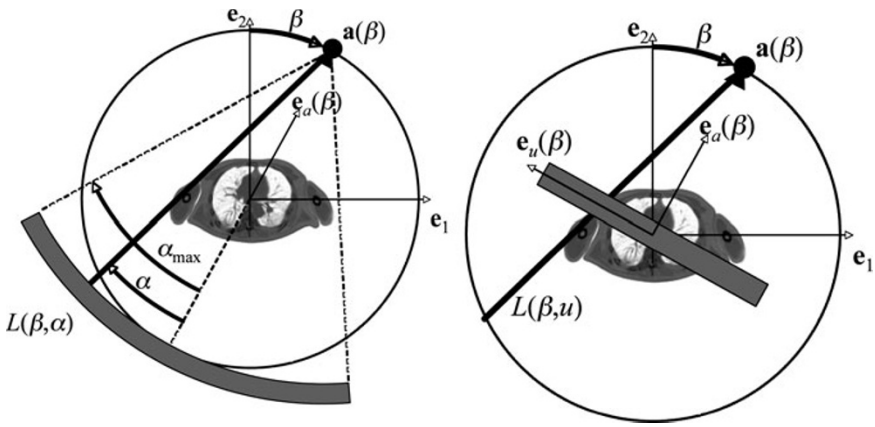


Fig. 9

Scanning geometry for fan-beam projections. *Left*: curved detector with equiangular detection elements. *Right*: flat detector with equidistant detection elements, virtually located at the center of rotation of the beam vertex

with $\beta \in [0, 2\pi)$ the angle describing the rotation. The detector rotates with the beam vertex and may be flat or curved.

3.3.1 Flat Detector

The detector is perpendicular to $\mathbf{e}_a(\beta)$ and the rays $L(\beta, u)$ are sampled equidistantly along its axis $\mathbf{e}_u(\beta)$, see *Fig. 9, right*. The origin of the detector ($u = 0$) is at the orthogonal projection of the beam vertex. For simplicity, the detector is assumed centered on the beam vertex center of rotation. For a real system, where the detector is placed further away from the beam vertex, the detection coordinate is obtained by a single scaling of u. The filtered-backprojection algorithm for fan-beam data $g(\beta, u)$ with a flat detector and a full scan ($\beta \in [0, 2\pi)$) is given by

$$f_{\mathrm{FBP}}(\mathbf{x}) = \frac{1}{2}\int_0^{2\pi} \frac{R^2}{(R - \mathbf{x}\cdot\mathbf{e}_a(\beta))^2} g^*(\beta, u(\mathbf{x},\beta))\,\mathrm{d}\beta, \tag{22}$$

where the filtered projections are given by

$$g^*(\beta, u) = \left(g(\beta, u)\frac{R}{\sqrt{R^2 + u^2}}\right) * w_c(u). \tag{23}$$

The filter $w_c(u)$ is the same as for parallel-beam data. The detector coordinate is given by

$$u(\mathbf{x},\beta) = R\frac{\mathbf{x}\cdot\mathbf{e}_u(\beta)}{R - \mathbf{x}\cdot\mathbf{e}_a(\beta)}. \tag{24}$$

3.3.2 Curved Detector

For X-ray CT, the detector is usually placed on a circular arc with its origin on the beam vertex (see ❯ *Fig. 9, left*). All detector elements are then at the same distance from the source and the rays $L(\beta, \alpha)$ are sampled equiangularly according to the fan angle α. The filtered-backprojection algorithm for fan-beam data $g(\beta, \alpha)$ with such a curved detector and a full scan ($\beta \in [0, 2\pi)$) is given by

$$f_{\mathrm{FBP}}(\mathbf{x}) = \frac{1}{2} \int_0^{2\pi} \frac{R^2}{||\mathbf{x} - R\mathbf{e}_a(\beta)||^2} g^*(\beta, \alpha(\mathbf{x}, \beta))\, \mathrm{d}\beta, \tag{25}$$

where the filtered projections are given by

$$g^*(\beta, \alpha) = (g(\beta, \alpha) \cos\alpha) * \left(R \left(\frac{\alpha}{\sin\alpha} \right)^2 w_c(R\alpha) \right). \tag{26}$$

The filter $w_c(R\alpha)$ is the same as for parallel-beam data. The detector coordinate is given by

$$\alpha(\mathbf{x}, \beta) = \arctan \frac{\mathbf{x} \cdot \mathbf{e}_{a_\perp}(\beta)}{R - \mathbf{x} \cdot \mathbf{e}_a(\beta)}. \tag{27}$$

3.3.3 Short Scan

In the previous sections, the projections were assumed to be measured over a full interval $\psi \in [0, 2\pi)$ or $\beta \in [0, 2\pi)$. For parallel-beam data, a projection interval $\psi \in [0, \pi)$ is sufficient as $g(\psi, s) = g(\psi + \pi, -s)$. The reconstruction can be limited to $f_{\mathrm{FBP}}(\mathbf{x}) = \int_0^\pi g^*(\psi, s(\mathbf{x}, \psi))\, \mathrm{d}\psi$. The same consideration holds for fan-beam data. For a curved detector, a projection interval $\beta \in [0, \pi + 2\alpha_{\max})$ is sufficient, where $\alpha_{\max}$ is the maximum fan aperture. This is referred as a *short-scan* acquisition, as opposed to a *full-scan* acquisition over $\beta \in [0, 2\pi)$. In transmission tomography, a short-scan acquisition avoids unnecessary radiations and reduces the absorbed dose to the patient.

A short-scan acquisition can be reconstructed with a long-scan algorithm, using $g(\beta, \alpha) = g(\beta + \pi + 2\alpha, -\alpha)$ to estimate the missing data for $\beta \in [\pi + 2\alpha_{\max}, 2\pi)$. This is referred as the complementary rebinning method. Alternatively, a short-scan reconstruction can be used over the interval $\beta \in [0, \pi + 2\alpha_{\max})$. However, some data are redundant in the rectangular interval $\beta \in [0, \pi + 2\alpha_{\max})$ and $\alpha \in [-\alpha_{\max}, +\alpha_{\max}]$. Prior to the short-scan reconstruction, short-scan data are windowed with a smooth window $A(\beta, \alpha)$,

$$g^w(\beta, \alpha) = A(\beta, \alpha) g(\beta, \alpha) \quad \text{with} \quad A(\beta, \alpha) + A(\beta + \pi + 2\alpha, -\alpha) = 1.0. \tag{28}$$

Usually, the Parker windowing (Parker 1982) is applied.

3.3.4 Helical Fan Beam

In practice, for volume scanning of the body in X-ray CT, the bed supporting the patient is continuously translated along the scanner axis $\mathbf{e}_3$, while the source–detector system is rotating, yielding a helical source trajectory around the patient. The pitch of the spiral scan is defined

by the distance translated by the bed during one 360 degree rotation of the source. The 3-dimensional position of the source is given by

$$\mathbf{a}(\beta) = R\mathbf{e}_a(\beta) + \frac{\beta\,\text{pitch}}{2\pi}\mathbf{e}_3. \tag{29}$$

For a single-row detector, a fan-beam filtered-backprojection algorithm can be used to reconstruct each transaxial slice of the scanned volume, provided an interpolation scheme is used to generate consistent fan-beam data of the reconstructed slices.

With the 360 degree linear interpolation method for a given slice $z_{2\pi}$, the fan-beam data $g_{z_{2\pi}}(\beta,\alpha)$ are interpolated from the adjacent measured data $g(\beta(z_1),\alpha)$ and $g(\beta(z_2),\alpha)$ such as $\beta(z_1) = \beta(z_2) = \beta$ and $z_1 < z_{2\pi} < z_2 = z_1 + \text{pitch}$. A full-scan fan-beam filtered-backprojection algorithm is then performed on the $z_{2\pi}$ slice.

With the 180-degree linear interpolation method, the data $g_{z_{2\pi}}(\beta,\alpha)$ are interpolated from the measured data $g(\beta(z_1),\alpha)$ and $g(\beta(z_2),-\alpha)$ such as $\beta(z_1) = \beta$, $\beta(z_2) = \beta + \pi + 2\alpha$, and $z_1 < z_{2\pi} < z_2 = z_1 + \text{pitch}(\pi + 2\alpha)/(2\pi)$. A short-scan filtered-backprojection algorithm is then performed on the $z_{2\pi}$ slice.

4 3D Analytical Reconstruction

If the line integrals cross several transaxial planes, a 3D algorithm has to be used to reconstruct simultaneously all slices. Unlike in two dimensions, it is not feasible to design a scanning geometry with a 4π coverage that provides the 3D X-ray transform over all directions $\mathbf{e}_\theta \in \mathbb{S}^2$. The 3D scanners have a limited aperture Ω defined by the set of measured directions $\mathbf{e}_\theta$. As a consequence, a direct generalization from two-dimensional to three-dimensional reconstruction is not possible. In addition, the axial coverage of the detectors is *de facto* limited, leading to unavoidable truncated projections.

4.1 Parallel-Beam Geometry for 3D PET Systems

In order to reconstruct a three-dimensional image, the Fourier slice theorem (Eq. 9) requires that for any frequency $\boldsymbol{\nu} \in \mathbb{R}^3$, at least one projection $\mathbf{e}_\theta$ which satisfies $\mathbf{e}_\theta \cdot \boldsymbol{\nu} = 0$ is measured by the scanner. Geometrically, this condition, known as the Orlov's condition, requires that the aperture Ω should intersect all equatorial circles on the unit sphere $\mathbb{S}^2$ (Orlov 1975). In particular, it shows that the subset Ω_0 of measured directions characterized by a co-polar angle ϕ equal to zero ($\Omega_0 = \{\mathbf{e}_\theta | \psi \in [0,\pi[, \phi = 0\}$) is sufficient to reconstruct an image. On the contrary, if the aperture $\Omega_{\phi_{\max}}$ includes oblique directions ($\Omega_{\phi_{\max}} = \{\mathbf{e}_\theta | \psi \in [0,\pi[, \phi \in [-\phi_{\max}, +\phi_{\max}], \phi_{\max} > 0\}$), then for any frequency $\boldsymbol{\nu}$, several directions $\mathbf{e}_\theta$ will be orthogonal to $\boldsymbol{\nu}$. This feature illustrates data redundancy specific to 3D imaging. It is used in 3D PET to improve signal-to-noise ratio by increasing the number of detected events.

The scanning system is assumed to satisfy the Orlov's condition and be shift invariant, that is, for each direction $\mathbf{e}_\theta \in \Omega$, all lines satisfying $\mathbf{s} \cdot \mathbf{e}_\theta = 0$ and crossing the support $\mathfrak{R}$ of f are measured (complete projections). The scanning model is defined by

$$g(\mathbf{e}_\theta, \mathbf{s}) = \mathbf{P}f(\mathbf{e}_\theta, \mathbf{s}), \tag{30}$$

where $g \in \mathbb{T}^4$ are the measured data, $\mathbf{P}$ the 3D X-ray transform, and f the image to reconstruct. The backprojection operator is defined by

$$\mathbf{P}^{\#}_{\Omega} g(\mathbf{x}) = \int_{\Omega} g\left(\mathbf{e}_\theta, \mathbf{x} - (\mathbf{x}\cdot\mathbf{e}_\theta)\mathbf{e}_\theta\right) \mathrm{d}\mathbf{e}_\theta. \tag{31}$$

The 3D filtered-backprojection algorithm is defined as

$$f_{\mathrm{FBP}}(\mathbf{x}) = \mathbf{P}^{\#}_{\Omega}\left(h \ast\ast g\right)(\mathbf{x}), \tag{32}$$

where $\ast\ast$ is the 2D convolution product on $\theta^{\perp}$ and $h(\mathbf{e}_\theta, \mathbf{s})$ a kernel defined on $\mathbb{T}^4$. The Fourier slice theorem shows that to get a valid reconstruction, $H(\mathbf{e}_\theta, \boldsymbol{\nu}_{\mathrm{s}})$, the 2D Fourier transform of the filter $h(\mathbf{e}_\theta, \mathbf{s})$, must satisfy

$$\int_{\Omega} \delta(\mathbf{e}_\theta \cdot \boldsymbol{\nu}) H(\mathbf{e}_\theta, \boldsymbol{\nu})\, \mathrm{d}\mathbf{e}_\theta = 1 \tag{33}$$

for any $\boldsymbol{\nu}$ in $\mathbb{R}^3$ (Defrise et al. 1989), where $\delta(x)$ is the Dirac distribution. Unlike 2D FBP, the filter is not unique; an infinite number of filters can be derived. Although they all yield identical images for consistent data, they differ for noisy data in the way they propagate noise in the reconstructed image. For 3D PET, the filter derived by Colsher minimizes the variance in the reconstructed image (Colsher 1980)

$$H_{\mathrm{Colsher}}(\boldsymbol{\nu}) = \frac{|\boldsymbol{\nu}|}{\int_{\Omega} \delta\left(\mathbf{e}_\theta \cdot \frac{\boldsymbol{\nu}}{|\boldsymbol{\nu}|}\right) \mathrm{d}\mathbf{e}_\theta}. \tag{34}$$

The Colsher filter does not depend on $\mathbf{e}_\theta$, and its frequency dependence is similar to that of the ramp filter in 2D FBP. The same regularization method using an apodization window can be used to limit the noise in the reconstructed image. Colsher's filter is usually used for 3D PET.

4.1.1 The Reprojection Algorithm

The 3D FBP algorithm defined by ❯ Eq. 32 requires complete data. For a real PET scanner, this requirement is only valid for non-oblique projections measured in Ω_0. For oblique projections ($\phi \neq 0$), the system is not shift invariant: only a subspace of the scanner field of view is covered. Measured data are missing at both axial ends of the scanner field of view. To overcome this difficulty, the missing data are estimated prior to the reconstruction. This defines the 3D reprojection algorithm (3DRP) of Kinahan and Rogers (1989). It is organized as follow for continuous data

- Perform a 2D FBP reconstruction with non-oblique data.
- Forward-project the 2D FBP reconstructed image to estimate the missing oblique data.
- Merge the measured and the estimated missing data.
- Perform a 3D FBP reconstruction with the merged data. For each measured direction $\mathbf{e}_\theta$:
  - Compute the 2D Fourier transform of the projection.
  - Multiply by the apodized Colsher filter.
  - Compute the inverse 2D Fourier transform of the filtered projection.
  - Backproject the filtered projection in the 3D image volume.

The discretization of the algorithm follows the same principle as for 2D FPB.

4.1.2 Rebinning Techniques

Fully 3D reconstructions take significantly more time to compute than a series of 2D reconstructions of the same transaxial slices. To keep the advantage of higher sensitivity of 3D scanning geometry and still use fast 2D reconstruction algorithms, measured data can be rebinned into a stack of 1D projections $g(s, \psi, z)$ of each transaxial slice z. Each slice is then reconstructed with a 2D reconstruction algorithm.

The rebinning techniques usually use a parametrization of the 2D projections in *oblique* sinograms by analogy to the sinogram in the 2D case

$$g_s(s, \psi, z, \phi) = g(\mathbf{e}_\theta(\psi, \phi), \mathbf{s}) \cos\phi, \tag{35}$$

where $s = \mathbf{s} \cdot \mathbf{e}_{\theta_1^\perp}$ is the radial coordinate and $z = \mathbf{s} \cdot \mathbf{e}_{\theta_2^\perp} \cos\phi$ the axial coordinate of the point of the LOR closest to the $\mathbf{e}_3$ axis.

The simplest rebinning technique is the single-slice rebinning (SSRB) algorithm (Daube-Witherspoon and Muehllehner 1987). It neglects the polar obliquity of the LOR and assumes that $g_s(s, \psi, z, 0) \cong g_s(s, \psi, z, \phi)$. It is defined by

$$g_{\mathrm{SSRB}}(s, \psi, z) = \frac{1}{2\phi_{\max(s,z)}} \int_{-\phi_{\max(s,z)}}^{\phi_{\max(s,z)}} g_s(s, \psi, z, \phi)\, \mathrm{d}\phi, \tag{36}$$

where $\phi_{\max(s,z)}$ is the maximum co-polar aperture. The numerical implementation of this approximate rebinning algorithm is straightforward, but it results in severe axial blurring when the co-polar aperture is important.

In practice, the approximate Fourier rebinning (FORE) algorithm (Defrise et al. 1997) is preferred as is it more accurate than the SSRB algorithm. It is based on the frequency–distance relation (Edholm et al. 1986) and proceeds in the Fourier domain. It is based on the following approximate relation

$$\mathcal{F}_{2D}\{g_s\}(\nu_\perp, k, z, 0) \cong \mathcal{F}_{2D}\{g_s\}\left(\nu_\perp, k, z' = z + \frac{k \tan\phi}{2\pi\nu_\perp}, \phi\right), \tag{37}$$

where k is the azimuthal Fourier index. In practice, the algorithm is adequately accurate when the co-polar aperture $\phi_{\max}$ is smaller than 20°. For a higher aperture, exact rebinning algorithms can be used (Liu et al. 1999). The FORE algorithm is used on many 3D PET systems.

4.2 Cone-Beam Reconstruction

In X-ray CT, the number of detector rows usually largely exceeds one. As a consequence, the co-polar angle of the projection ray can be important, requiring 3D reconstruction algorithms. Development of exact reconstruction algorithms for cone-beam data is a very active research field as the number of detector rows increases continuously for new systems and the demand for significant reduction of the patient exposure to X-ray is getting more stringent.

Unlike fan-beam data, cone-beam data cannot be expressed as non-truncated 2D parallel projections, followed by a parallel-beam 3D FBP reconstruction. The scanning geometry of cone-beam tomography requires dedicated reconstruction algorithms. For a small cone aperture (few detector rows), cone-beam data can be rebinned into a stack of approximate fan-beam projections corresponding to each reconstructed slice. Two-dimensional reconstruction methods combined with axial filtering can provide sufficient image quality (Hu 1999).

4.2.1 The Feldkamp, Davis, and Kress Algorithm

The most popular algorithm over the past decades is the algorithm proposed by Feldkamp, Davis, and Kress (FDK) in 1984 (Feldkamp et al. 1984) to reconstruct cone-beam data for a circular orbit of the X-ray source around the patient and a planar 2D detector (multi-row detector). Despite it is not an exact reconstruction algorithm and considerable progress in cone-beam reconstruction has been achieved, it is still widely used in its original or modified form for helical cone-beam CT with cylindrical detectors.

The FDK algorithm is an empirical extension of the 2D fan-beam filtered-backprojection algorithm presented in ❯ Sect. 3.3. For the measured cone-beam data $g(\beta, u, v)$, where u refers to the transaxial coordinate (like in 2D) and v to the axial coordinate (row) of the detector, the reconstructed image is given by

$$f_{\mathrm{FDK}}(\mathbf{x}) = \frac{1}{2}\int_0^{2\pi} \frac{R^2}{(R - \mathbf{x}\cdot\mathbf{e}_a(\beta))^2} g^*(\beta, u(\mathbf{x},\beta), v(\mathbf{x},\beta))\,\mathrm{d}\beta, \tag{38}$$

with a backprojection performed along the cone-beam rays defined by the detector coordinates

$$u(\mathbf{x},\beta) = R\frac{\mathbf{x}\cdot\mathbf{e}_u(\beta)}{R - \mathbf{x}\cdot\mathbf{e}_a(\beta)}, \quad v(\mathbf{x},\beta) = R\frac{\mathbf{x}\cdot\mathbf{e}_3}{R - \mathbf{x}\cdot\mathbf{e}_a(\beta)}. \tag{39}$$

The filtering of the projection is 1D, similar to the fan-beam algorithm

$$g^*(\beta, u, v) = \left(g(\beta, u, v)\frac{R}{\sqrt{R^2 + u^2 + v^2}}\right) * w_c(u). \tag{40}$$

The weighting factor can be interpreted as the cosine of the angle between the (β, u, v) ray and the central ray $(\beta, 0, 0)$. The algorithm is exact in the plane of the circular orbit of the X-ray source.

The major advantage of the FDK algorithm is its simplicity. For moderate cone angles, the deviations of the reconstructed image to the true image are small and acceptable. The algorithm also handles data truncated in the axial direction (limited axial coverage of the detector) as the filtering is only performed in the non-truncated fan direction $\mathbf{e}_u(\beta)$. The main distortion is a blurring of the reconstructed image in the axial direction.

4.2.2 Exact Reconstruction Algorithms

The Tuy condition (Tuy 1983) provides a sufficient condition on the data for an exact reconstruction based on the 3D Radon transform of the object $\mathbf{R}f(\mathbf{e}_\Pi, l)$ (❯ Eq. 6). The 3D Radon transform of all planes intersecting the support $\mathfrak{R}$ of the function $f(\mathbf{x})$ has to be known in order to perform an exact reconstruction of $f(\mathbf{x})$. For parallel projections, a plane integral is obtained from a one-dimensional integration of projection data along a straight line in the plane. For cone-beam data, the condition implies that all planes which cross the support $\mathfrak{R}$ should also intersect the cone vertex trajectory at least once. This means, in particular, that a circular (non-spiral) trajectory does not satisfy Tuy's condition as any transaxial plane above or below the circular trajectory plane is not intersected by the cone vertex trajectory.

For a cone-beam geometry, a plane integral is not equal to a 1-dimensional integral of divergent projection data. P. Grangeat derived a relation to calculate the radial derivative of

the 3D Radon transform from divergent projections and then processed an exact reconstruction (Grangeat 1991). If the Tuy's condition is fulfilled, the first derivative of the 3D Radon transform $\frac{\partial \mathbf{R}f}{\partial l}(\mathbf{e}_\Pi, l)$ can be computed for each plane $\Pi(\mathbf{e}_\Pi, l)$ which crosses the support $\mathfrak{R}$ using Grangeat's formula. Then, the function $f(\mathbf{x})$ can be reconstructed exactly by inversion of the 3D Radon transform

$$f_{\mathrm{CB}}(\mathbf{x}) = -\frac{1}{4\pi^2}\int_0^\pi \int_0^\pi \sin\phi \frac{\partial \mathbf{R}f}{\partial l}(\mathbf{e}_\Pi(\phi,\psi), l = \mathbf{x}\cdot\mathbf{e}_\Pi)\,\mathrm{d}\phi\,\mathrm{d}\psi, \tag{41}$$

where ϕ is the polar angle and ψ to azimuthal angle of $\mathbf{e}_\Pi$.

This method requires that $g(\beta, u, v)$ is not truncated. However, $g(\beta, u, v)$ is usually truncated along the axial direction v, and Grangeat's formula cannot be used directly to compute the value $\frac{\partial \mathbf{R}f}{\partial l}(\mathbf{e}_\Pi, l)$ for all planes containing the cone-beam vertex $\mathbf{a}(\beta)$. The method has been adapted to compute $\frac{\partial \mathbf{R}f}{\partial l}(\mathbf{e}_\Pi, l)$ by data combination using several vertex locations intersecting the plane $\Pi(\mathbf{e}_\Pi, l)$, see, for example, Kudo et al. (1998).

In practice, only a part of the body is scanned. The "object" (the body) extends axially far beyond the limits of $\mathfrak{R}$. The *long-object* problem consists in the reconstruction of a region of interest in a long object from truncated helical cone-beam data using a helical path which only extends slightly beyond the axial limits of the region of interest. A. Katsevich derived an exact FBP long-object algorithm for helical cone-beam CT. The processing of each cone-beam projection consists of two steps: shift-invariant and $\mathbf{x}$-independent filtering along a family of tilted lines on the detector, followed by the backprojection of the filtered projection (Katsevich 2002). Among other exact algorithms, it allows for fast reconstruction and involves only weak restrictions on the scanner geometry.

5 Iterative Reconstruction Algorithms

Analytical reconstruction algorithms are based on the line integral model: the measured data g provide us with the projection of the function f to be reconstructed: $g = \mathbf{P}f$. A direct inversion of this model is achieved with the filtered-backprojection algorithm: $f_{\mathrm{FBP}} = \mathbf{P}^{\#}(w_c * g)$. The line integral model is an idealization of the scanning process. In practice, the scanning model is more complex due to several effects:

- Data are noisy: the measurements are random variables following a probability distribution.
- Measurements are discrete.
- Detectors are subject to measurement errors like mis-positioning of the photon.
- Not all photons travel along straight lines through the body: some are scattered and others are absorbed.
- The scanning geometry does not necessary provide complete data, data can be truncated.

The use of a more complex and realistic scanning model for the reconstruction of the image offers the potential benefit of a better image quality, for example, by improving the noise versus spatial resolution trade-off in the reconstructed image. However, a complex scanning model does not necessary lead to a feasible direct inversion. It is possible to model analytically some of these effects such as to obtain a direct inversion; but these are usually not optimized for noisy data that are typically encountered in emission tomography. Iterative algorithms are more

flexible to solve such problems as they do not require a direct inversion of the model. An objective function $\Phi(f)$ expressing a measure relating the measured data to the estimated image is solved iteratively. The solution is given by

$$\widehat{f}_{\text{iter}} = \arg\min_{f \in \Psi} \Phi(f), \tag{42}$$

where Ψ defines the set of feasible solutions. Typically, the value of f should be nonnegative. The objective function can be seen as some distance metric between the measured data and the scanning model applied to the estimated image.

Iterative reconstruction algorithms are commonly used in emission tomography. Statistical noise is an important limitation of emission tomography and iterative reconstructions with a noise model can improve image quality over analytical algorithms. In whole-body PET imaging, iterative algorithms are now routinely used. In X-ray computed tomography, the noise is usually not such an issue (at least, when compared to emission tomography), and data sets are much larger. However, low-dose X-ray scans and truncated data geometries drive interest for iterative algorithms.

5.1 Scanning Model

Iterative reconstruction algorithms are characterized by

- A finite parametrization of the image
- A model for the discrete measured data
- A noise model, that is a probability distribution function for the measured data
- An objective function relating the measured data to the estimated image
- An iterative algorithm to solve the objective function

The first three items define the scanning model. If the model includes a noise model, the reconstruction algorithm is also called *statistical*.

For a scanner with n detection elements, the measured data are represented by a vector $\mathbf{y} = \{y_i | i = 1, \ldots, n\} \in \mathbb{R}^n$. Each element y_i represents the number of detected photons for detection element i. For iterative algorithms, the native scanning geometry of the data can be conserved. There is no need to rebin the data into a regular grid of equidistant parallel lines. If some data are missing, there is no need to estimate them prior to the reconstruction. In their principles, there is no difference between 2D and 3D iterative reconstruction algorithms. They will not be covered separately.

The image $f(\mathbf{x})$ is decomposed into m spatial basis functions $f_j(\mathbf{x})$ with $f(\mathbf{x}) = \sum_{j=1}^{m} \lambda_j f_j(\mathbf{x})$ in emission tomography and $f(\mathbf{x}) = \sum_{j=1}^{m} \mu_j f_j(\mathbf{x})$ in transmission tomography. The reconstruction in emission tomography consists in the determination of the vector $\boldsymbol{\lambda} = \{\lambda_j | j = 1, \ldots, m\} \in \mathbb{R}^m$ from the data $\mathbf{y}$, respectively the vector $\boldsymbol{\mu}$ in transmission tomography.

The usual choice for the spatial basis functions is the voxel

$$f_j(\mathbf{x}) = v(\mathbf{x} - \mathbf{x}_j) = \begin{cases} 1 & \text{if} \quad |(\mathbf{x} - \mathbf{x}_j) \cdot \mathbf{e}_k| \leq \Delta x_k \text{ for } k = 1, \ldots, 3 \\ 0 & \text{else} \end{cases}, \tag{43}$$

where the voxel centers $\mathbf{x}_j$ are on a regular 3D lattice with spacing $(\Delta x_1, \Delta x_2, \Delta x_3)$. As an alternative to cubic voxels, overlapping functions with spherical symmetries are also used, like the bell-shaped *blobs* proposed by R.M. Lewitt (1992) for iterative reconstructions

$$f_j(\mathbf{x}) = b(\|\mathbf{x} - \mathbf{x}_j\|), \tag{44}$$

where $\|\cdot\|$ is the euclidean norm and b a generalized Kaiser–Bessel function. The blobs have the following important features: their values and derivatives are continuous, they are completely localized in space, are band limited, and their projection has a convenient analytical form. Wavelets have also been used as basis elements (Frese et al. 2002).

In emission tomography, let $a_i(\mathbf{x})$ be the sensitivity function for the detection element i such that the number y_i of detected events is a random variable with an expectation value given by the linear model

$$\mathrm{E}\{y_i\} = T \int_{\mathbb{R}^3} a_i(\mathbf{x}) f(\mathbf{x})\, \mathrm{d}\mathbf{x} + b_i, \tag{45}$$

where T is the acquisition duration and b_i represents the average number of background events. For simplicity, the acquisition duration T will be omitted from now in the equations.

Let $\mathbf{A} = \{A_{ij}\}$ be the $n \times m$ system matrix defined by

$$A_{ij} = \int_{\mathbb{R}^3} a_i(\mathbf{x}) f_j(\mathbf{x})\, \mathrm{d}\mathbf{x}. \tag{46}$$

In vector notation, the acquisition model is

$$\mathrm{E}\{\mathbf{y}\} = \mathbf{A}\boldsymbol{\lambda} + \mathbf{b}. \tag{47}$$

The system matrix elements can be computed numerically (Selivanov et al. 2000), experimentally measured with a point-like radioactive source (Panin et al. 2006), or estimated by Monte Carlo simulation (Rafecas et al. 2004). For the line-integral model used by analytical reconstruction algorithms, the sensitivity function $a_i(\mathbf{x})$ is zero except when $\mathbf{x}$ lies on the line centered on detection element i. The corresponding matrix element A_{ij} can be computed from the intersection length between the line i and the voxel j. Iterative algorithms allow for the modeling in $\mathbf{A}$ of the 3D extension of the line. It includes geometrical effects defined by the size of the detectors or the solid angle of the collimator hole, and photon mislocalization in the detectors. Photon attenuation in the body and detector efficiency can be included in the system matrix. The modeling of photon scattering in the tissues in the sensitivity function $a_i(\mathbf{x})$ introduces long-range effects from the line i, making an efficient numerical implementation of the matrix difficult. For PET, random coincidences are nonlinear effects that cannot be modeled in a linear model. For these reasons, the average spatial distributions of scattered events and random coincidences in the measured data are usually assumed to be known prior to image reconstruction and included in the background term $\mathbf{b}$.

In order to ease the numerical implementation and manipulation of the system matrix, it is common to factorize the different contributions to the model in separate sparse matrices. In PET, the following decomposition has been proposed for the $n \times m$ system matrix (Qi et al. 1998)

$$\mathbf{A}\boldsymbol{\lambda} = \mathbf{A}_{\text{d-eff}}\mathbf{A}_{\text{d-blur}}\mathbf{A}_{\text{attn}}\mathbf{A}_{\text{geom}}\mathbf{A}_{\text{p-range}}\boldsymbol{\lambda}, \tag{48}$$

where

- $\mathbf{A}_{\text{p-range}}$ is an $m \times m$ blurring matrix to model positron range.
- $\mathbf{A}_{\text{geom}}$ is an $n \times m$ matrix for the geometrical forward projection.

- $\mathbf{A}_{\text{attn}}$ is an $n \times n$ diagonal matrix with the inverse of the attenuation correction factors.
- $\mathbf{A}_{\text{d-blur}}$ is an $n \times n$ blurring matrix to model photon mislocalization in the detectors.
- $\mathbf{A}_{\text{d-eff}}$ is an $n \times n$ diagonal matrix with the inverse of the normalization factors.

In transmission tomography, by analog to $\mathbf{A}$, let $\mathbf{M} = \{M_{ij}\}$ be the $n \times m$ system matrix such that the expectation for the data $\mathbf{y}$ is given by

$$\mathrm{E}\{y_i\} = I_i \mathrm{e}^{-\sum_{j=1}^{m} M_{ij}\mu_j} + b_i, \tag{49}$$

where I_i is the blank scan count, assumed noiseless, and b_i the expected count of scattered photons.

Photon detection in emission and transmission tomography is a Poisson process. The measured data $\mathbf{y}$ are integer random variables that represent the number of detected events (X-rays, single gammas, or pairs of gammas in coincidence), and their probability distribution is a Poisson distribution. This model is not valid if the data $\mathbf{y}$ are preprocessed prior to reconstruction or photons detection is not an independent process. If the data are pre-corrected for background events, they do not follow a Poisson distribution. This means that for a valid Poisson noise model, all corrections should be included in the scanning model of the reconstruction and not applied prior to reconstruction like for analytical reconstruction algorithms. In practice, this is not always the case. The use of an inappropriate noise model can lead to degraded image quality (Comtat et al. 1998).

The Poisson distribution is a discrete probability distribution characterized by the probability mass function

$$P(n|\theta) = \frac{\mathrm{e}^{-\theta}\theta^n}{n!}, \tag{50}$$

where $n \in \mathbb{N}$. The Poisson distribution is defined by one parameter θ: its mean $\mathrm{E}\{n\}$ and its variance $\sigma^2\{n\}$ are equal to θ. In emission tomography, the mean is given by the forward model $\mathrm{E}\{\mathbf{y}\} = \mathbf{A}\boldsymbol{\lambda} + \mathbf{b}$ and depends on the unknown image $\boldsymbol{\lambda}$. The probability distribution for $\mathbf{y}$ is given by

$$P(\mathbf{y}|\boldsymbol{\lambda}) = \prod_{i=1}^{n} \frac{\mathrm{e}^{-(\langle \mathbf{A}_i \cdot \boldsymbol{\lambda} \rangle + b_i)} (\langle \mathbf{A}_i \cdot \boldsymbol{\lambda} \rangle + b_i)^{y_i}}{y_i!}, \tag{51}$$

where $\langle \mathbf{A}_i \cdot \boldsymbol{\lambda} \rangle = \sum_{j=1}^{m} A_{ij}\lambda_j$. For transmission tomography,

$$P(\mathbf{y}|\boldsymbol{\mu}) = \prod_{i=1}^{n} \frac{\mathrm{e}^{-(I_i \mathrm{e}^{-\langle \mathbf{M}_i \cdot \boldsymbol{\mu} \rangle} + b_i)} (I_i \mathrm{e}^{-\langle \mathbf{M}_i \cdot \boldsymbol{\mu} \rangle} + b_i)^{y_i}}{y_i!}. \tag{52}$$

5.2 Objective Function and Minimization Algorithm

Given the data $\mathbf{y}$, the probability distributions $P(\mathbf{y}|\boldsymbol{\lambda})$ and $P(\mathbf{y}|\boldsymbol{\mu})$ can be regarded as a function of the unknown images $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$. This is the *likelihood* function of $\boldsymbol{\lambda}$ or $\boldsymbol{\mu}$ given $\mathbf{y}$.

The most common objective function in statistical image-reconstruction algorithms is the likelihood. This was proposed already in 1976 by Rockmore and Macovski (Rockmore and Macovski 1976). Given an observation $\mathbf{y}$, the goal is to compute the most likely (ML) estimate of the parameter $\boldsymbol{\lambda}$ or $\boldsymbol{\mu}$. This is equivalent and more convenient to maximize the logarithm of

the likelihood or minimize the negative log-likelihood. For emission tomography

$$\widehat{\boldsymbol{\lambda}}_{\mathrm{ML}} = \arg\max_{\boldsymbol{\lambda}\geq 0} L(\boldsymbol{\lambda}) = \arg\max_{\boldsymbol{\lambda}\geq 0} \left\{ \sum_{i=1}^{n} \left(y_i \ln\left(\langle \mathbf{A}_i \cdot \boldsymbol{\lambda}\rangle + b_i\right) - \left(\langle \mathbf{A}_i \cdot \boldsymbol{\lambda}\rangle + b_i\right)\right)\right\} \tag{53}$$

and for transmission tomography

$$\widehat{\boldsymbol{\mu}}_{\mathrm{ML}} = \arg\max_{\boldsymbol{\mu}\geq 0} L(\boldsymbol{\mu}) = \arg\max_{\boldsymbol{\mu}\geq 0} \left\{ \sum_{i=1}^{n} \left(y_i \ln\left(I_i \mathrm{e}^{-\langle \mathbf{M}_i \cdot \boldsymbol{\mu}\rangle} + b_i\right) - \left(I_i \mathrm{e}^{-\langle \mathbf{M}_i \cdot \boldsymbol{\mu}\rangle} + b_i\right)\right)\right\}, \tag{54}$$

where the terms that do not depend on $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ have been dropped.

The maximization of the likelihood can be regarded as the minimization of some distance. For Poisson noise, the estimator $\widehat{\boldsymbol{\lambda}}_{\mathrm{ML}}$ coincides (Titterington 1987) with the minimizer of the following "distance" between $\mathbf{y}$ and $\mathbf{A}\boldsymbol{\lambda} + \mathbf{b}$, based on the Kullback–Leibler divergence (not a true distance metric):

$$\widehat{\boldsymbol{\lambda}}_{\mathrm{KL}} = \arg\min_{\boldsymbol{\lambda}} \left\{ \sum_{i=1}^{n} \left(y_i \ln\left(\frac{y_i}{\langle \mathbf{A}_i \cdot \boldsymbol{\lambda}\rangle + b_i} \right) - y_i + \langle \mathbf{A}_i \cdot \boldsymbol{\lambda}\rangle + b_i \right)\right\}. \tag{55}$$

Once an objective function has been defined, the most difficult part of iterative reconstruction algorithms is to derive a maximization (or minimization) algorithm. For the ML estimator, the *expectation-maximization* algorithm (EM-ML) of Dempster and colleagues (Dempster et al. 1977) is well suited for emission tomography. This algorithm was proposed for computed tomography independently by Lange and Carson (1984) and by Shepp and Vardi (1982).

5.2.1 The EM-ML Algorithm in Emission Tomography

The Hessian matrix $\nabla^2 L$ of the log-likelihood formed by the second derivatives $\frac{\partial^2 L}{\partial \lambda_k \partial \lambda_l}$ is negative semidefinite for all allowed images (Vardi et al. 1985): for any image $\boldsymbol{\lambda} \geq 0 \in \mathbb{R}^m$, $\boldsymbol{\lambda}^{\top} \nabla^2 L \boldsymbol{\lambda} \leq 0$. This shows that $L(\boldsymbol{\lambda})$ is concave and that a local maximum of $L(\boldsymbol{\lambda})$ will also be a global maximum.

In the expression of the log-likelihood (Eq. 53), the term $y_i \ln(\langle \mathbf{A}_i \cdot \boldsymbol{\lambda}\rangle)$ prevents an analytical maximization: the probability distribution $\ln P(y_i|\boldsymbol{\lambda})$ depends on several components of $\boldsymbol{\lambda}$. The data $\mathbf{y}$ are said to be *incomplete*. The basic idea of the EM algorithm is to postulate a so-called *complete* data random vector $\mathbf{z}$ made of independent variables such as no more than one component of $\boldsymbol{\lambda}$ contributes to the expectation of each element of $\mathbf{z}$.

Let $\mathbf{z}$ be a complete data random vector $\{\{x_{ij}|j = 1,\ldots,m\}, q_i|i = 1,\ldots,n\} \in \mathbb{R}^{n\times m+n}$ where, for detection element i, x_{ij} is defined as the number of events originating from voxel j and q_i the number of detected background events, with the many to one mapping $y_i = \sum_{j=1}^{m} z_{ij} + q_i$ (Politte and Snyder 1991). The expectations are given by $\mathrm{E}\{x_{ij}\} = A_{ij}\lambda_j$ and $\mathrm{E}\{q_i\} = b_i$. The data $\mathbf{z}$ follow a Poisson distribution with a log-likelihood with respect to $\mathbf{z}$ given by

$$\ln R(\mathbf{z}|\boldsymbol{\lambda}) = \sum_{i=1}^{n} \left(\sum_{j=1}^{m} \left(x_{ij} \ln(A_{ij}\lambda_j) - A_{ij}\lambda_j \right) + q_i \ln(b_i) - b_i \right), \tag{56}$$

where the terms $x_{ij}!$ and $q_i!$ have been dropped. As only the incomplete data $\mathbf{y}$ are known, the key principle of the EM algorithm is, instead of maximizing the log-likelihood with respect

to $\mathbf{z}$, to maximize the expected log-likelihood with the expectation taken with respect to the unobserved complete data using a guess $\boldsymbol{\lambda}^{(p)}$ of $\boldsymbol{\lambda}$ (Dempster et al. 1977). The EM algorithm is decomposed in two steps: the calculation of the expectation (*E-step*)

$$Q(\boldsymbol{\lambda}|\boldsymbol{\lambda}^{(p)}) = \mathrm{E}\left\{\ln R(\mathbf{z}|\boldsymbol{\lambda})|\mathbf{y}, \boldsymbol{\lambda}^{(p)}\right\}, \tag{57}$$

and its maximization (*M-step*)

$$\boldsymbol{\lambda}^{(p+1)} = \arg\max_{\boldsymbol{\lambda}\geq 0} Q(\boldsymbol{\lambda}|\boldsymbol{\lambda}^{(p)}). \tag{58}$$

If we take $\boldsymbol{\lambda}^{(p+1)}$ as a new guess and iterate over the E and M steps, the EM algorithm will produce a likelihood sequence $L(\boldsymbol{\lambda}^{(p)})$ that is monotonically increasing. As $L(\boldsymbol{\lambda})$ is concave, it will converge to its unique maximum.

For a random vector $\mathbf{x}$ made of independent Poisson variables x_i with mean a_i, its conditional expectation is given by

$$\mathrm{E}\left\{x_i|\sum_j x_j = N\right\} = N\frac{a_i}{\sum_j a_j}. \tag{59}$$

The *E-step* is then given by

$$Q(\boldsymbol{\lambda}|\boldsymbol{\lambda}^{(p)}) = \sum_{i=1}^{n}\left(\sum_{j=1}^{m}\left(-A_{ij}\lambda_j + y_i\frac{A_{ij}\lambda_j^{(p)}}{\sum_{k=1}^{m}A_{ik}\lambda_k^{(p)} + b_i}\ln(A_{ij}\lambda_j)\right)\right. \tag{60}$$
$$\left. -b_i + y_i\frac{b_i}{\sum_{k=1}^{m}A_{ik}\lambda_k^{(p)} + b_i}\ln(b_i)\right).$$

This expression can be analytically maximized using the Kuhn–Tucker conditions, leading for the *M-step* to

$$\lambda_j^{(p+1)} = \frac{\lambda_j^{(p)}}{\sum_{i=1}^{n}A_{ij}}\sum_{i=1}^{n}A_{ij}\frac{y_i}{\sum_{k=1}^{m}A_{ik}\lambda_k^{(p)} + b_i} \tag{61}$$

or, in vector–matrix notation,

$$\boldsymbol{\lambda}^{(p+1)} = \frac{\boldsymbol{\lambda}^{(p)}}{\mathbf{A}^{\mathrm{T}}\mathbf{1}}\mathbf{A}^{\mathrm{T}}\frac{\mathbf{y}}{\mathbf{A}\boldsymbol{\lambda}^{(p)} + \mathbf{b}}, \tag{62}$$

where multiplication and division between vectors are component wise, $\mathbf{1}$ is a vector of length n with all elements set to 1, and the matrix $\mathbf{A}^{\mathrm{T}}$ is the transpose of $\mathbf{A}$. By analogy to the analytical operators $\mathbf{P}$ and $\mathbf{P}^{\#}$, the matrix $\mathbf{A}$ corresponds to the forward projection of the image $\boldsymbol{\lambda}$ and the matrix $\mathbf{A}^{\mathrm{T}}$ to the backprojection of the data $\mathbf{y}$.

The *E-step* and the *M-step* are combined into a single step (Eq. 62). The EM-ML algorithm for emission computed tomography is

- Initialize $\boldsymbol{\lambda}^{(p=0)}$ with strictly positive values : $\lambda_j^{(0)} > 0 \;\forall\; j = 1, \ldots, m$.
- Compute the normalization image: $\mathbf{n} = \mathbf{A}^{\mathrm{T}}\,\mathbf{1}$.
- For iteration (p)
  - Compute the expected data given $\boldsymbol{\lambda}^{(p)}$: $\mathrm{E}\{\mathbf{y}|\boldsymbol{\lambda}^{(p)}\} = \mathbf{A}\boldsymbol{\lambda}^{(p)} + \mathbf{b}$.

- Compute the ratio between the measured data and their expectation: $\mathbf{r}^{(p)} = \mathbf{y}/\mathrm{E}\{\mathbf{y}|\boldsymbol{\lambda}^{(p)}\}$.
- Backproject the ratio: $\mathbf{c}^{(p)} = \mathbf{A}^{\mathrm{T}}\mathbf{r}^{(p)}$.
- Update the current image: $\boldsymbol{\lambda}^{(p+1)} = \boldsymbol{\lambda}^{(p)}\mathbf{c}^{(p)}/\mathbf{n}$.
- Go to next iteration.

The relative simplicity of this algorithm, in particular the fact that the E and M steps are combined, also explains its popularity. If the starting image $\boldsymbol{\lambda}^{(p=0)}$ is strictly positive, the algorithm ensures that the reconstructed images $\boldsymbol{\lambda}^{(p)}$ are positive as long as the data $\mathbf{y}$ are themselves positive (if the data are pre-corrected for background events, they can be negative; negative values are then set to zero).

The EM-ML algorithm belongs to the class of simultaneous techniques: it makes an update of the reconstructed image based on all data bins simultaneously. As a consequence, it suffers from a major drawback: it is slow to converge. To speed up convergence, accelerated algorithms have been proposed. The most popular, as it requires only minor implementation changes relative to the original algorithm, is the *Ordered-Subsets* EM (OSEM) algorithm from Hudson and Larkin (1994). The OSEM algorithm belongs to the class of block-iterative algorithms: the measured data are grouped in ordered disjoint subsets, and the standard EM-ML algorithm is applied to each of the subsets in turn. The resulting reconstructed image of one subset becomes the starting value for the next subset. The subsets are chosen in a balanced way so that all voxels contribute approximately equally to any subset. In such case, the convergence acceleration provided by the OSEM algorithm is roughly equal to the number of subsets for the early iterations. However, for noisy data, the algorithm does not converge to the ML estimate; it cycles through a number of distinct points. This is not a major limitation, as in practice the EM-ML algorithm is usually not run until convergence.

Another accelerated algorithm for the ML estimate is the row-action maximum-likelihood algorithm (RAMLA) of Browne and De Pierro (1996). A row-action algorithm updates the reconstructed image for each data bin separately, processing one row of the system matrix. RAMLA was inspired from the ART (algebraic reconstruction algorithm) technique of Herman and Meyer (1993), an algebraic row-action algorithm for minimizing a least-square objective function. Let l be the index of the sub-iterations over the n data bin and p the index for a complete cycle over the n indices l with $\boldsymbol{\lambda}^{(p,0)} = \boldsymbol{\lambda}^{(p-1)}$ and $\boldsymbol{\lambda}^{(p)} = \boldsymbol{\lambda}^{(p,n)}$. The expression of the RAMLA algorithm is given by

$$\boldsymbol{\lambda}^{(p,l+1)} = \boldsymbol{\lambda}^{(p,l)} + \epsilon_p \boldsymbol{\lambda}^{(p,l)} \mathbf{A}_{i_l} \left(\frac{y_{i_l}}{\langle \mathbf{A}_{i_l} \cdot \boldsymbol{\lambda}^{(p,l)} \rangle + b_{i_l}} - 1 \right), \tag{63}$$

where ϵ_p is a strictly positive relaxation parameter such that $\epsilon_p A_{ij} \leq 1 \ \forall \ (i,j)$, i_l is a permutation of the data bin, and $\mathbf{A}_i$ is the i^{th} column of $\mathbf{A}^{\mathrm{T}}$. The ordering i_l of the data bin is such that the consecutive vectors $\mathbf{A}_{i_l}$ are as orthogonal to each other as possible to accelerate convergence. The RAMLA algorithm avoids the limit cycle by using strong underrelaxation involving a decreasing sequence of relaxation parameters ϵ_p: under the conditions that $\lim_{p\to\infty} \epsilon_p = 0$ and $\sum_{p=0}^{\infty} \epsilon_p = +\infty$, the algorithm converges to $\widehat{\boldsymbol{\lambda}}_{\mathrm{ML}}$. If the numerical implementation of RAMLA is such that one complete cycle over the data bin is no more computationally intensive than one EM iteration, then it is about one order of magnitude faster than EM in convergence.

5.2.2 ML Algorithms in Transmission Tomography

For transmission tomography, an EM-ML algorithm has been derived by Lange and Carson (1984). The complete data are defined as $\mathbf{y}$, augmented by the number of photons entering (u_{ij}) and leaving (v_{ij}) each voxel j along each projection line i. The expectation (E-step) is given by

$$Q(\boldsymbol{\mu}|\boldsymbol{\mu}^{(p)}) = \sum_{i=1}^{n}\sum_{j=1}^{m}\Big(-\mathrm{E}\{v_{ij}|y_i,\boldsymbol{\mu}^{(p)}\}M_{ij}\mu_j + \left(\mathrm{E}\{u_{ij}|y_i,\boldsymbol{\mu}^{(p)}\} - \mathrm{E}\{v_{ij}|y_i,\boldsymbol{\mu}^{(p)}\}\right) \ln(1-\mathrm{e}^{-M_{ij}\mu_j})\Big). \quad (64)$$

Unlike for emission data, it does not yield a closed-form solution for the M-step. As a consequence, algorithms for resolving the M-step are slow to converge (Ollinger 1994).

Algorithms more efficient than EM-ML were developed based on direct maximization of the log-likelihood (Lange et al. 1987; Lange and Fessler 1995). One approach proposed originally by De Pierro for emission tomography is to use a surrogate function that approximates locally the objective function and is easier to minimize (De Pierro 1995). Given a current estimate $\boldsymbol{\mu}^{(p)}$ of the image, the surrogate $S(\boldsymbol{\mu};\boldsymbol{\mu}^{(p)})$ function is tangent to $-L(\boldsymbol{\mu}^{(p)})$ at $\boldsymbol{\mu}^{(p)}$ and lies above $-L(\boldsymbol{\mu})$ for other positive $\boldsymbol{\mu}$:

$$S(\boldsymbol{\mu};\boldsymbol{\mu}^{(p)}) \geq -L(\boldsymbol{\mu})\ \forall\boldsymbol{\mu}\geq 0, \quad S(\boldsymbol{\mu}^{(p)};\boldsymbol{\mu}^{(p)}) = -L(\boldsymbol{\mu}^{(p)}), \quad \nabla S(\boldsymbol{\mu}^{(p)};\boldsymbol{\mu}^{(p)}) = -\nabla L(\boldsymbol{\mu}^{(p)}). \quad (65)$$

If the surrogate is convex, then its minimization or diminution decreases the objective function. This procedure is repeated iteratively to minimize the objective function

$$\boldsymbol{\mu}^{(p+1)} = \arg\min_{\boldsymbol{\mu}\geq 0} S(\boldsymbol{\mu};\boldsymbol{\mu}^{(p)}). \quad (66)$$

The convergence rate of the algorithm depends on the curvature of the surrogate. The EM-ML algorithm for emission tomography can be interpreted as a surrogate-based technique, where the conditional expectation $Q(\boldsymbol{\lambda}|\boldsymbol{\lambda}^{(p)})$ is an inferior surrogate function for the log-likelihood $L(\boldsymbol{\lambda})$. Because the curvature of the conditional expectation at $\boldsymbol{\lambda}^{(p)}$ is large compared to $L(\boldsymbol{\lambda})$, the EM-ML algorithm is slow to converge.

Erdogan and Fessler developed an algorithm based on a surrogate function to minimize the negative log-likelihood $-L(\boldsymbol{\mu})$ (Eq. 52) in the presence of background events ($\mathbf{b} \neq 0$): the *separable paraboloidal surrogates* (SPS) algorithm (Erdogan and Fessler 1999). The surrogate is chosen quadratic and separable in the voxel elements j so that its minimization is reduced to the minimization of m 1D parabolas that depend on one voxel value μ_j only. The algorithm allows for simultaneous update, and the authors applied the ordered-subset principle of OSEM to SPS (OSTR) to accelerate convergence at the sacrifice of global convergence. Let k be the index of the sub-iterations over the K subsets and p the index for a complete cycle over the K subsets with $\boldsymbol{\mu}^{(p,0)} = \boldsymbol{\mu}^{(p-1)}$ and $\boldsymbol{\mu}^{(p)} = \boldsymbol{\mu}^{(p,k)}$. The expression of the ML-OSTR algorithm is given by

$$\mu_j^{(p,k+1)} = \max\left\{\mu_j^{(p,k)} - \frac{K\sum_{i\in S_k} M_{ij}h_i^{(p,k)}}{d_j};0\right\}, \quad (67)$$

where

$$h_i^{(p,k)} = \left(\frac{y_i}{I_i\mathrm{e}^{-\langle\mathbf{M}_i\cdot\boldsymbol{\mu}^{(p,k)}\rangle}+b_i}-1\right)I_i\mathrm{e}^{-\langle\mathbf{M}_i\cdot\boldsymbol{\mu}^{(p,k)}\rangle} \quad (68)$$

and $d_j = \sum_{i=1}^{n} M_{ij}(y_i - b_i)^2/y_i \sum_{j=1}^{m} M_{ij}$.

An alternative is to use the OSEM algorithm with $\ln(I_i/y_i)$ instead of y_i, but this is not optimal for low-count studies as the data $\ln(I_i/y_i)$ do not follow a Poisson distribution. For X-ray CT, another alternative is to assume Gaussian data. In this case, the maximization of the log-likelihood is equivalent to the minimization of a least-square function.

5.3 Regularization

Like the FBP algorithm, the ML estimation problem is ill-conditioned: small changes in the data might result in large variations in the reconstructed ML image. The algorithm attempts to recover the underlying distribution that best matches, in a statistical sense, noisy data (Snyder et al. 1987). When the data are very noisy, the ML estimate might be too noisy for an appropriate use. Like FBP, some regularization is needed to constrain the solution and get an "acceptable" image. The meaning of acceptable depends on the practical imaging task and the criteria might differ between feature detection or quantitative estimation of a physiological parameter. There are different ways to regularize an ML-based algorithm.

5.3.1 Early Termination

During the EM iterations, low frequencies are reconstructed first. For higher iteration numbers, the algorithm will essentially attempt to recover noise and yields a deteriorating "checkerboard effect" of high spatial variance in the reconstructed images (see ❯ *Fig. 10*). When starting from a smooth image, at some early iterations of the EM-ML algorithm, the image might "look" nicer than for higher iteration numbers. Early termination of the algorithm is one way of regularization that is commonly used in clinical studies. It has the major advantage of reducing the computation time of the image reconstruction. This also explains the popularity of the OSEM algorithm: the non-convergence of the subset algorithm is not a practical limitation. The major drawback of early termination is the nonuniform convergence of the algorithm and consequently the spatially varying resolution of the image. Some parts of the image converge slower, in particular lower-count regions as opposed to higher-count regions.

The choice of the number of iterations is empirical and task dependent. For detection, small features should be visible and not hidden by surrounding noise. For quantitative estimations, it is important to verify that the numerical value of the parameter of interest has reached a plateau. The balance between noise level and resolution is controlled by the number of iterations.

5.3.2 Post-Reconstruction Smoothing

In order to be less sensitive to the nonstationary convergence of the EM-ML algorithm, it might be appropriate to run the algorithm for a higher number of iterations and then smooth the reconstructed image with a Gaussian kernel. This approach is related to the Grenander's method of sieves (Snyder et al. 1987), which constrains the estimate of the image to be in a subset, called the sieve, of the space of nonnegative functions. An illustration is given in ❯ *Fig. 11* for the same data set as for ❯ *Fig. 10*. The trade-off between noise and resolution is controlled by the width of the Gaussian kernel.

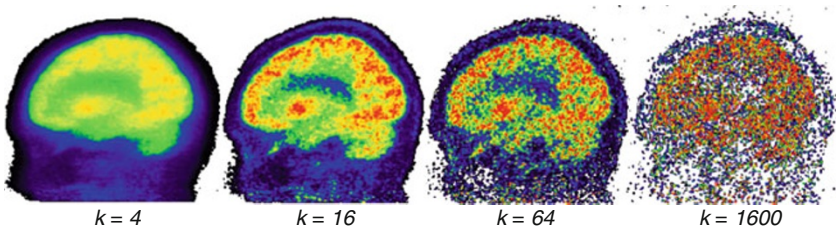


Fig. 10
Sagittal slice of the reconstructed images of a 2 min [^{18}F]-FDG brain PET scan after *k* iterations of the EM-ML algorithm

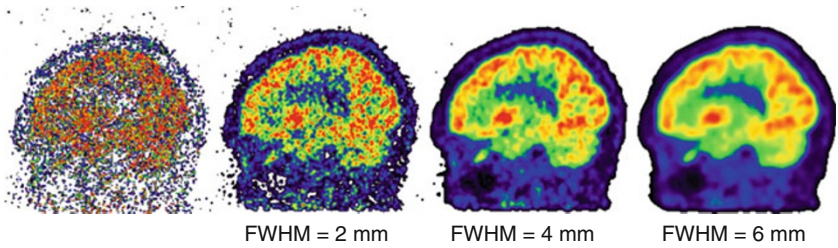


Fig. 11
Sagittal slice of the reconstructed images of a 2 min [^{18}F]-FDG brain PET scan after 1600 iterations of the EM-ML algorithm, followed by a smoothing with an isotropic 3D Gauss function of width FWHM

5.3.3 Penalized Objective-Function or MAP Reconstruction

An alternative solution to resolve the ill-conditioning in the ML consists in the addition of a penalty function $U(\boldsymbol{\lambda})$ directly in the objective function and running the iterative algorithm until convergence to compute a penalized ML estimate

$$\widehat{\boldsymbol{\lambda}}_{\mathrm{PML}} = \arg\max_{\boldsymbol{\lambda}\geq 0} L(\boldsymbol{\lambda}) - U(\boldsymbol{\lambda}). \tag{69}$$

The log-likelihood $L(\boldsymbol{\lambda})$ relates the estimated image to the data (data fitting term), whereas the penalty function $U(\boldsymbol{\lambda})$ penalizes excessive noise in the estimated image by adding a roughness penalty (regularization term).

The expression in Eq. 69 can also be viewed in a Bayesian framework: given some prior distribution on the image, $p(\boldsymbol{\lambda})$, the posterior density conditioned on the data, $p(\boldsymbol{\lambda}|\mathbf{y})$, is maximized. The relation between these terms is given by the Bayes rule

$$p(\boldsymbol{\lambda}|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\lambda})p(\boldsymbol{\lambda})}{p(\mathbf{y})}. \tag{70}$$

The *Maximum A Posteriori* (MAP) reconstruction algorithm consists in the computation of the maximum estimate of the log of the posterior density:

$$\widehat{\boldsymbol{\lambda}}_{\mathrm{MAP}} = \arg\max_{\boldsymbol{\lambda}\geq 0} L(\boldsymbol{\lambda}) + \ln p(\boldsymbol{\lambda}), \tag{71}$$

where the term $p(\mathbf{y})$ has been dropped.

The images are usually modeled as locally smooth with sharp transition between different types of tissues. Two types of prior are generally considered: spatially independent priors and Gibbs priors. In the first case, voxels are assumed statistically independent and optimizations can be based on extensions of the EM-ML algorithm. However, these priors require information on the expected mean voxel values distribution. It can be appropriate for transmission tomography, in particular with radionuclide sources, where the μ values of the tissues are known.

In the second case, more common, spatial interactions between voxels are modeled using a Gibbs distribution, with the generic form

$$p(\boldsymbol{\lambda}) = \frac{1}{Z} \mathrm{e}^{-\beta U(\boldsymbol{\lambda})}, \tag{72}$$

where $U(\boldsymbol{\lambda})$ is the Gibbs energy function. The MAP estimate is then given by

$$\widehat{\boldsymbol{\lambda}}_{\mathrm{MAP}} = \arg\max_{\boldsymbol{\lambda}\geq 0} L(\boldsymbol{\lambda}) - \beta U(\boldsymbol{\lambda}), \tag{73}$$

where the hyper-parameter β controls the weight of the prior relative to the data fidelity term.

The Gibbs energy function is generally defined as a sum of potential functions V on the absolute difference between pairs of neighboring voxels

$$U(\boldsymbol{\lambda}) = \sum_{j=1}^{m} \sum_{k\in N_j, k>j} \frac{V(|\lambda_j - \lambda_k|)}{d_{jk}}, \tag{74}$$

where d_{jk} is the Euclidean distance between the pair of voxels (j, k) and N_j the neighbor subset of voxel j. Typically, N_j consists in the 6 (first order) or 26 (second order) nearest voxels of voxel j. One example is the Huber prior which is quadratic for small values and switches to linear at a user-specified value δ to allow for sharp transitions

$$V_{\mathrm{H}}(t) = \begin{cases} t^2/2 & \text{if} \quad |t| \leq \delta \\ |t|\delta - \delta^2/2 & \text{else} \end{cases}. \tag{75}$$

In emission tomography, when anatomical information is available in the form of a co-registered MR or X-ray CT image, the smoothing prior can be deactivated between pairs of voxels belonging to different anatomical structures.

Many optimization procedures for MAP, or equivalently PML, estimation are proposed in the literature. They can be classified into the following methods (Qi and Leahy 2006):

- Gradient-based algorithms: the direction of the update is calculated from the gradient of the objective function. The preconditioned conjugate-gradient algorithm is particularly efficient, although it is difficult to enforce the non-negativity constraint on the image (Mumcuoglu et al. 1994).
- Coordinate-ascent algorithms: one voxel is updated in turn so as to ease the application of the non-negativity constraint. The objective function is maximized with respect to one voxel while holding the other fixed, leading to a one-dimensional problem. The algorithm is particularly efficient with a Gaussian noise model and a quadratic penalty term: the objective function is then quadratic (Fessler 1994).
- Functional substitution: the original objective function is replaced at each step by a surrogate function (De Pierro 1995). An example was given in Eqs. 65 and 66.

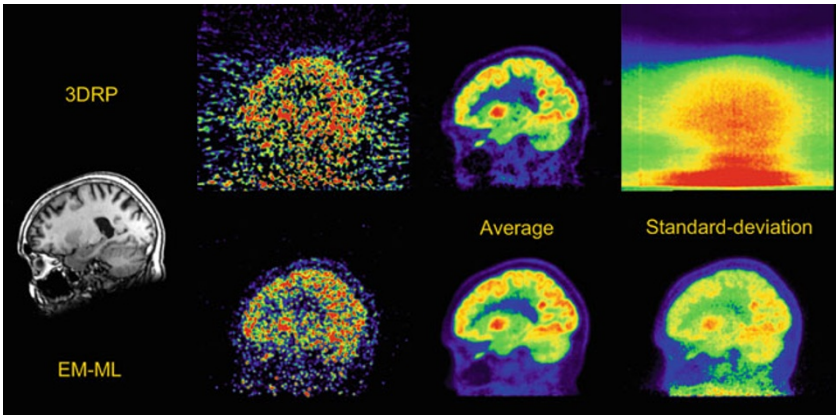


Fig. 12
Sagittal slice of the reconstructed images of a [^{18}F]-FDG brain PET scan. Three hundred and sixty independent acquisitions of 10 s were performed and reconstructed with the analytical 3DRP algorithm (Hann apodization window and Nyquist cutoff frequency) (*top row*) and with the iterative EM-ML algorithm (96 iterations, 2 mm FWHM post-reconstruction Gaussian smoothing) (*bottom row*). *Extreme left*: MR image. *Left*: one acquisition; *middle*: average across the 360 reconstructions; *right*: standard deviation across the 360 reconstructions

It should be pointed out that at the time of the writing of this chapter, MAP or PML reconstructions have not been adopted routinely in clinical applications. The favor still goes to the EM-ML algorithm or its accelerated variants (OSEM or RAMLA), with a post-reconstruction smoothing of the estimated image. This method is easy to implement and already produces satisfactory results in practice when compared to FBP.

6 Concluding Remarks

As an illustration, a comparison between the EM-ML and the FBP algorithms is presented in *Fig. 12* for very noisy PET data corresponding to a 10-s acquisition. The FBP image is characterized by a high level of noise all across the reconstructed field of view, while the noise is more concentrated on the brain for the EM-ML image. The same acquisition was repeated 360 times and reconstructed with both algorithms, allowing for the computation of a mean image and a standard-deviation image across the 360 statistically independent reconstructions. Although one replica is very noisy and difficult to examine, the average of the replicas looks similar between FBP and EM-ML, nicely showing the structures of the cortex. On the contrary, the standard-deviation images are very different; quite similar to the average image for EM-ML based on a Poisson noise model, and much more uniform spatially for FBP, based on no noise model.

The choice between analytical and iterative reconstruction techniques is not obvious, although iterative algorithms allow in principle for better modeling of the scanning process and could be preferred for this particular reason. This is definitely not the end of analytical methods. The latter are linear algorithms, unlike iterative algorithms, which is advantageous for tasks like

the quantitative estimation of physiological parameters. In addition, analytical techniques run much faster on a computer: one iteration of an optimization-based technique takes about the same time as one FBP reconstruction. Implementation simplicity is also a decisive advantage, in particular for the scanner manufacturers. A very accurate modeling of the scanning process will surely make its implementation more complex for an hypothetical gain on image quality.

This chapter presents a limited view on tomographic reconstruction techniques, focusing on some long-established techniques. Many methods and new fields were left out. Among many uncovered research fields, we can cite the active developments that are currently performed on dynamic reconstructions (4D reconstructions), including time as an additional variable to space to account for the change in the tracer kinetics or the morphological motions of patients, or even both (5D reconstruction). It is also interesting to note the new developments on analytical algorithms for X-ray CT to perform exact and stable 2D reconstruction of a region of interest with a limited data set that were thought to be intractable before (Defrise and Gullberg 2006).

The reader is encouraged to read the following chapters of the handbook related to image reconstruction in computed tomography:

7 Cross-References

- Chapter 5, "Statistics"
- Chapter 36, "CT Imaging: Basics and New Trends"
- Chapter 37, "SPECT Imaging: Basics and New Trends"
- Chapter 38, "PET Imaging: Basics and New Trends"
- Chapter 41, "Quantitative Image Analysis in Tomography"
- Chapter 43, "Evaluation and Image Quality in Radiation-Based Medical Imaging"

References

Browne J, de Pierro A (1996) A row-action alternative to the EM algorithm for maximizing likelihood in emission tomography. IEEE Trans Med Imag 15(5):687–699

Colsher JG (1980) Fully-three-dimensional positron emission tomography. Phys Med Biol 25(1):103

Comtat C, Kinahan P, Defrise M, Michel C, Townsend D (1998) Fast reconstruction of 3D PET data with accurate statistical modeling. IEEE Trans Nucl Sci 45(3):1083–1089

Daube-Witherspoon ME, Muehllehner G (1987) Treatment of axial data in three-dimensional PET. J Nucl Med 28(11):1717–1724

De Pierro AR (1995) A modified expectation maximization algorithm for penalized likelihood estimation in emission tomography. IEEE Trans Med Imag 14(1):132–137

Defrise M, Gullberg GT (2006) Review: image reconstruction. Phys Med Biol 51(13):R139–R154

Defrise M, Kinahan P, Townsend D, Michel C, Sibomana M, Newport D (1997) Exact and approximate rebinning algorithms for 3-D pet data. IEEE Trans Med Imag 16(2):145–158

Defrise M, Townsend DW, Clack R (1989) Three-dimensional image reconstruction from complete projections. Phys Med Biol 34(5):573

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J Roy Stat Soc B 39(1):1–38

Edholm PR, Lewitt RM, Lindholm B (1986) Novel properties of the fourier decomposition of the sinogram. In: International workshop on physics and engineering of computerized multidimensional imaging and processing, vol 671, Newport Beach, California, pp 8–18

Erdogan H, Fessler JA (1999) Ordered subsets algorithms for transmission tomography. Phys Med Biol 44(11):2835–2851

Feldkamp LA, Davis LC, Kress JW (1984) Practical cone-beam algorithm. J Opt Soc Am A 1(6): 612–619

Fessler J (1994) Penalized weighted least-square image reconstruction for positron emission tomography. IEEE Trans Med Imag 13(2): 290–300

Frese T, Bouman C, Sauer K (2002) Adaptive wavelet graph model for bayesian tomographic reconstruction. IEEE Trans Image Process 11(7): 756–770

Grangeat P (1991) Mathematical framework of cone beam 3D reconstruction via the first derivative of the radon transform. In: Mathematical methods in tomography, Lecture notes in mathematics, vol 1497. Springer, Berlin/Heidelberg, pp 66–97

Herman G, Meyer L (1993) Algebraic reconstruction techniques can be made computationally efficient positron emission tomography application. IEEE Trans Med Imag 12(3):600–609

Hu H (1999) Multi-slice helical CT: scan and reconstruction. Med Phys 26(1):5–18. Available at http://link.aip.org/link/?MPH/26/5/1

Hudson H, Larkin R (1994) Accelerated image reconstruction using ordered subsets of projection data. IEEE Trans Med Imag 13(4):601–609

Katsevich A (2002) Analysis of an exact inversion algorithm for spiral cone-beam CT. Phy Med Biol 47(15):2583–2597. Available at http://stacks.iop.org/00319155/47/i=15/a=302

Kinahan P, Rogers J (1989) Analytic 3D image reconstruction using all detected events. IEEE Trans Nucl Sci 36(1):964–968

Kudo H, Noo F, Defrise M (1998) Cone-beam filtered-backprojection algorithm for truncated helical data. Phys Med Biol 43(10):2885–2909. Available at http://stacks.iop.org/0031-9155/43/i=10/a=016

Lange K, Bahn M, Little R (1987) A theoretical study of some maximum likelihood algorithms for emission and transmission tomography. IEEE Trans Med Imag 6(2):106–114

Lange K, Carson R (1984) Em reconsturction algorithm for emission and transmission tomography. J Comp Assist Tomogr 8(2):306–316

Lange K, Fessler J (1995) Globally convergent algorithms for maximum a posteriori transmission tomography. IEEE Trans Image Process 4(10):1430–1438

Lewitt RM (1992) Alternatives to voxels for image representation in iterative reconstruction algorithms. Phy Med Biol 37(3):705–716

Liu X et al (1999) Exact rebinning methods for three-dimensional pet. IEEE Trans Med Imag 18(8):657–664

Mumcuoglu E, Leahy R, Cherry S, Zhou Z (1994) Fast gradient-based methods for bayesian reconstruction of transmission and emission pet images. IEEE Trans Med Imag 13(4): 687–701

Natterer F (2001) The mathematics of computerized tomography. SIAM, Philadelphia

Novikov RG (2002) On the range characterization for the two-dimensional attenuated x-ray transformation. Inverse Prob 18(3):677

Ollinger J (1994) Maximum-likelihood reconstruction of transmission images in emission computed tomography via the EM algorithm. IEEE Trans Med Imag 13(1):89–101

Orlov S (1975) Theory of three dimensional reconstruction. I. Conditions for a complete set of projections. Sov Phys Crystallogr 20(3): 312–314

Panin V, Kehren F, Michel C, Casey M (2006) Fully 3-D PET reconstruction with system matrix derived from point source measurements. IEEE Trans Med Imag 25(7):907–921

Parker D (1982) Optimal short scan convolution reconstruction for fan-beam CT. Med Phys 9(2):254–257

Politte D, Snyder D (1991) Corrections for accidental coincidences and attenuation in maximum-likelihood image reconstruction for positron-emission tomography. IEEE Trans Med Imag 10(1):82–89

Qi J, Leahy RM (2006) Iterative reconstruction techniques in emission computed tomography. Phys Med Biol 51(15):R541. Available at http://stacks.iop.org/0031-9155/51/i=15/a=R01

Qi J, Leahy RM, Cherry SR, Chatziioannou A, Farquhar TH (1998) High-resolution 3D bayesian image reconstruction using the micropet small-animal scanner. Phys Med Biol 43(4):1001–1013

Rafecas M et al (2004) Use of Monte Carlo based probability matrix for 3-D reconstruction of MADPET-II data. IEEE Trans Nucl Sci 51(5):2597–2605

Rockmore AJ, Macovski A (1976) A maximum likelihood approach to emission image reconstruction from projections. IEEE Trans Nucl Sci 23(4):1428–1432

Selivanov V, Picard Y, Cadorette J, Rodrigue S, Lecomte R (2000) Detector response models for statistical iterative image reconstruction in high resolution PET. IEEE Trans Nucl Sci 47(3):1168–1175

Shepp LA, Vardi Y (1982) Maximum likelihood reconstruction for emission tomography. IEEE Trans Med Imag 1(2):113–122

Snyder DL, Miller MI, Thomas LJ, Politte DG (1987) Noise and edge artifacts in maximum-likelihood reconstructions for emission tomography. IEEE Trans Med Imag 6(3):228–238

Titterington DM (1987) On the iterative image space reconstruction algorthm for ect. IEEE Trans Med Imag 6(1):52–56

Tuy HK (1983) An inversion formula for cone-beam reconstruction. SIAM J Appl Math 43(3): 546–552

Vardi Y, Shepp LA, Kafman L (1985) A statistical model for positron emission tomography. J Am Stat Assoc 80(389): 8–37

40 Motion Compensation in Emission Tomography

Jörg van den Hoff · Jens Langner
Helmholtz-Zentrum Dresden-Rossendorf, Dresden, Germany

1 Introduction 1008

2 Different Types of Motion and Their Effects 1009
2.1 Periodic Motion 1010
2.2 Irregular Motion 1010
2.3 Rigid Motion 1014
2.4 Nonrigid Motion 1014

3 Motion Detection 1014
3.1 Internal Motion Detection 1014
3.2 External Motion Detection 1016
3.2.1 Phase-Sensitive Motion Sensors 1016
3.2.2 Optical Motion Sensors 1016

4 Motion Correction 1019
4.1 Image-Based Techniques 1019
4.1.1 Image Registration 1020
4.1.2 Optical Flow 1022
4.1.3 Multiple Acquisition Framing 1025
4.1.4 Image Deblurring Approaches 1028
4.1.5 Correction of Breathing-Related Motion 1028
4.2 Projection-Based Techniques 1031
4.3 Event-Based Techniques 1031
4.3.1 Pre-correction of List Mode Data 1033
4.3.2 Incorporation of Motion Correction into the Image Reconstruction Process 1039

5 Conclusion 1040

References 1041

C. Grupen, I. Buvat (eds.), *Handbook of Particle Detection and Imaging*,
DOI 10.1007/978-3-642-13271-1_40, © Springer-Verlag Berlin Heidelberg 2012

Abstract: With the ever-improving spatial resolution available in single photon emission computed tomography (SPECT) and, especially, in positron emission tomography (PET), the unavoidable organ and subject motion is becoming one of the dominant factors limiting the practically achievable spatial resolution in the tomographic images. Moreover, uncorrected subject motion can lead to potentially severe image artifacts and compromise the quantitative integrity of the data. The latter is of special importance in PET where quantitative assessment of tracer concentrations is commonplace both in static investigations via so-called standardized uptake values (SUVs) and in dynamic studies aiming at tracer kinetic modeling and quantification of the corresponding transport constants. Correction of the heart-cycle-related motion in cardiac applications has a long tradition and is covered extensively in the literature. Correction of breathing-related organ motion in emission tomography, however, has drawn considerable interest only in recent years in the context of oncological PET. This is mainly due to the demands of therapy response monitoring and radiation treatment planning. The third important area is high-precision motion correction of random head motion in brain investigations. In this chapter, we give an overview of the methods employed to minimize – and possibly eliminate – the motion influence in emission tomography.

1 Introduction

Emission tomography using radioactively labeled tracers comes in two flavors: *Single Photon Emission Tomography* (Chap. 37, "SPECT Imaging: Basics and New Trends") and *Positron Emission Tomography* (Chap. 38, "PET Imaging: Basics and New Trends"). Both modalities have seen much technical progress during the last decades both in hardware (scintillation crystals, collimators, analog and digital signal processing electronics) as well as software (scatter correction algorithms, image reconstruction algorithms (Chap. 39, "Image Reconstruction"), etc.). As a result, today's machines offer a combination of improved spatial and temporal resolution, better quantitative accuracy regarding measurement of local radiotracer concentrations, and reduced acquisition times.

In PET, with state-of-the-art machines, one is able to reduce data acquisition times down to a few minutes for static investigations (i.e., those not investigating time-dependent tracer kinetic processes). This improvement is partly due to an increase of the axial field of view (FOV) in combination with 3D data acquisition. In addition, new scintillation materials and faster front-end electronics contribute to the improved sensitivity. A further reduction of acquisition times is principally limited by the requirement of maintaining a certain minimum level of statistical accuracy, i.e., signal-to-noise ratio (SNR) in the reconstructed images. The injected dose and thus the photon flux cannot be much increased due to radiation protection requirements as well as due to limitations of the count-rate performance of the hardware. Thus, acquisition times remain substantial, typically several minutes for a single image volume, and the obtained tomographic images represent "long exposure pictures" of the investigated subjects.

These relatively long acquisition times increase the probability of subject motion during the course of the investigation. It is obvious that uncorrected subject motion can lead to a

potentially severe image degradation and compromise the quantitative integrity of the data. The latter is of special importance in PET where quantitative assessment of tracer concentrations is commonplace both in static investigations (via so-called standardized uptake values (SUVs) (Thie 2004)) and in dynamic studies aiming at tracer kinetic modeling and quantification of the corresponding transport constants, see van den Hoff (2005) and ❯ Chap. 42, "Compartmental Modeling in Emission Tomography" in this book. The described problems caused by subject motion become increasingly serious considering the improved spatial resolution of recent PET devices.

In this chapter, we give an overview of currently employed methods to minimize – and possibly eliminate – the motion influence in the acquired data. In doing so, we will not be able to cover all aspects of the topic in equal depths due to limitations in available space. Therefore, we refer the reader to the literature where several review articles have appeared over the last years covering material partly complementary to that presented in this chapter, see, e.g., Rahmim et al. (2007), Nehmeh and Erdi (2008).

2 Different Types of Motion and Their Effects

There are a number of possible types of motion to be considered which differ in origin, their ultimate influence on the tomographic images, and the available options for correcting them. The latter range from straightforward approaches to quite elaborate and sophisticated algorithms.

One important criterion in distinguishing different motion types concerns the temporal characteristics, namely the question whether the motion is (semi-)periodic or irregular. The other one concerns the spatial characteristics, i.e., the distinction between rigid motions (e.g., in brain investigations) and nonrigid motion typically encountered in investigations of the body stem. In the following sections, we briefly discuss the most important aspects of the different motion types.

In doing so, we focus on the description of the alterations of the effective *Point Spread Function* (PSF) which characterizes the imaging process. To a good approximation, the imaging process, i.e., the translation from object space (the spatial radioactivity distribution in the case of emission tomography) to image space is a linear operation. The cumulative motion influence during a given time interval can be considered as contributing to the effective total PSF of the imaging process, H_t, by translating the resting activity distribution to a new one which then undergoes the actual imaging process. The finally measured signal S (the image) for a given activity distribution A can then be written as

$$S = H_t \otimes A = H_s \otimes H_m \otimes A\,, \tag{1}$$

where H_s is the PSF of the imaging system including the image reconstruction process, H_m describes the motion influence, and $\otimes$ denotes convolution.

While H_s is time-invariant and also approximately spatially invariant across the FOV (constant spatial resolution of the device), H_m obviously is dependent on time as well as on location within the FOV since all motions involving rotations and/or nonrigid deformations lead to a spatially variant H_m. Nevertheless, for a given motion within a given time interval, H_m is a well-defined function at each point in the FOV.

2.1 Periodic Motion

Periodic motions are encountered as a consequence of the cardiac and breathing cycles. While the heartbeat-related movement obviously is only a concern when investigating this organ, breathing-related movement affects not only the lung but rather extends over most of the thorax and abdomen. Correction of breathing-related movement thus is a very important problem in all types of whole-body investigations, notably in oncology. The dominant effect of periodic organ motion is deterioration of spatial resolution. For illustration, consider the following one-dimensional example where a point source executes harmonic oscillations around its origin with an amplitude a:

$$x(t) = a \cdot \sin(\omega t)\,.$$

The absolute value of the source velocity is

$$v(x) = |\dot{x}| = \omega \cdot a \cdot |\cos(\omega t)| = \omega \cdot \sqrt{a^2 - x^2}\,.$$

The probability $\mathrm{d}p$ of finding the object near position x is proportional to $\frac{1}{v(x)}\mathrm{d}x$. Including the correct normalization the resulting probability density $\frac{\mathrm{d}p}{\mathrm{d}x}$ is equal to the motion's PSF, H_m, and is given by

$$H_m(x) = \frac{1}{\pi \cdot \sqrt{a^2 - x^2}}\,, \tag{2}$$

which expresses quantitatively the obvious fact that the object is to be found most probably near the points of maximum elongation where the velocity is small. A "long exposure picture" of the moving source will then yield an image of this probability density, H_m, convolved with the PSF of the imaging system, H_s, which is identical to the image of an ideal point source at rest:

$$H_t = H_s \otimes H_m = \int_{-\infty}^{\infty} H_s(x')H_m(x - x')\,\mathrm{d}x'\,. \tag{3}$$

Figure 1 illustrates how such a harmonic motion modifies the total PSF, H_t, as a function of the motion amplitude. This is also demonstrated in *Fig. 2*, which shows results from measurements with a point source. The important point to note here is that the typical motion amplitudes occurring in patient investigations are of the same order as the reconstructed spatial resolution of current tomographs (breathing-related motion up to about 20 mm, corresponding to amplitudes of 10 mm), which leads to a significant reduction of the effective spatial resolution. While for most of the presently installed PET systems the reconstructed spatial resolution is about 6 mm, new tomographs achieve distinctly better performance, which leads to a correspondingly larger impact of breathing-related motion.

2.2 Irregular Motion

By "irregular motion" we designate all motions due to involuntary or voluntary random subject movement. Such movements can in principle affect all parts of the body but are usually largest for the head due to practical limitations of patient fixation and the relative ease with which head movements can be executed during the investigation. The image deteriorating effects of the movement are in this case aggravated by the fact that the spatial resolution in brain investigations is for certain reasons (related to a reduced size of the relevant FOV and reduced scatter and attenuation influence) somewhat higher than in whole-body investigations.

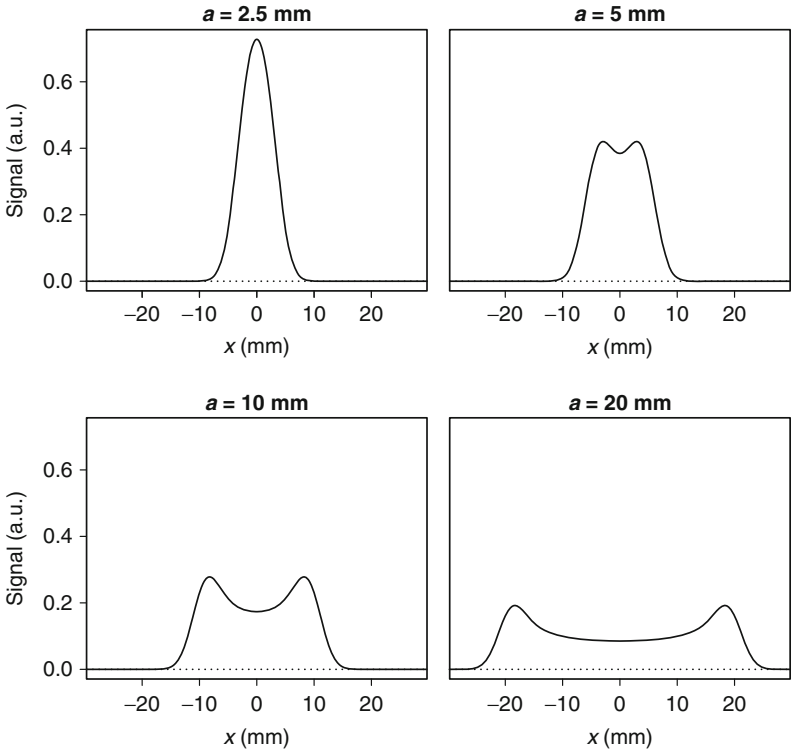


Fig. 1

Imaging a point source executing harmonic oscillations around the origin. The signal intensity is expressed relative to the maximum of the point source image at rest. The system's PSF, H_s, is assumed to be a Gaussian with a full width at half maximum (FWHM) of 5 mm. For motions where $2 \cdot a$ (the total range of the motion) is smaller than the FWHM the main effect is image blurring (increased width of the profile and decreased signal intensity, but only minor shape distortion). For larger motion amplitudes, shape distortions (elongation and heterogeneities) are observed as well

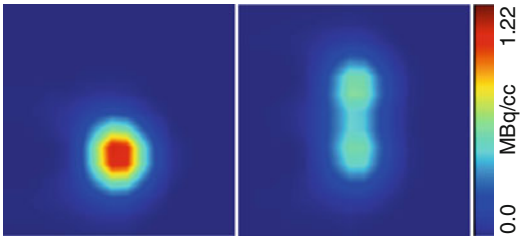


Fig. 2

PET image of a 1 mm point source (^{68}Ge , 5 min acquisition). *Left:* Source at rest. *Right:* Source oscillating along the vertical with a frequency of 0.1 Hz and an amplitude of 10 mm

The influence of irregular motions on the tomographic images is more difficult to predict than in the case of periodic motions simply because different irregular motions can have completely different characteristics. Considering a single point source, a few limiting cases of irregular motion might serve to illustrate this:

Abrupt motion ("stepping"). This phenomenon occurs when the patient readjusts his position spontaneously at one or several time points during the investigation, e.g., to find a more comfortable resting position on the patient bed. Despite measures regarding patient fixation, this occurs rather frequently, notably in brain investigations. The effect on the images is a sort of double exposure effect for larger motion amplitudes leading to ghosting artifacts, i.e., structures are duplicated in the images. The relative intensity of the "ghosts" is determined by the fractional time the respective position was occupied. The total PSF (point source image) in this case is the sum of two or more Gaussians with different means and amplitudes.

Random Walk around the origin ("walking"). This idealized type of motion (approximating, e.g., motion caused by unrest or neurological disorders of the patient) leads to a Gaussian probability density of the time-averaged source position with some standard deviation σ_m. This is the easiest possible motion type in terms of predicting its effect. The PSF of the tomograph itself is usually well approximated by a Gaussian with a standard deviation, σ_s, defining the spatial resolution of the system. The convolution of this PSF with the motion's Gaussian probability density yields a new Gaussian – the total PSF of the system including the random-walk motion – with a σ_t of

$$\sigma_t = \sqrt{\sigma_s^2 + \sigma_m^2}\,. \tag{4}$$

In this idealized case, the motion leaves the shape of the PSF unaltered. It is still a Gaussian but with increased width, corresponding to a reduction of the effective spatial resolution. Note that the variances, not the standard deviations, add up in Eq. 4, making the resolution reduction moderate as long as σ_m is somewhat smaller than the inherent resolution of the tomograph.

Linear motion ("creeping"). This type of motion occurs quite frequently in practice. Patients tend to slowly move out of the FOV along the patient bed leading sometimes to an overall displacement of a few centimeters. This type of idealized motion corresponds to a rectangular probability density which is centered not at the origin but rather at the midpoint of the traveled distance. The convolution with the PSF of the system in this case leads to an anisotropic blurring and a shift of the image position relative to the motion-free image. The resulting effective PSF of the imaging process along the given direction of the motion can be expressed with the help of the *Gaussian Error Function* but no longer by a simple Gaussian.[1] It rather resembles a trapezoidal shape with slopes whose width is determined by the machine's PSF, see *Figs. 3* and *4*.

[1] For a total traveled distance d, the resulting *PSF* is given by

$$H_t(x) = \frac{1}{2} \cdot \left(\mathrm{erf}\left(\frac{x}{\sigma_s \sqrt{2}} \right) - \mathrm{erf}\left(\frac{x-d}{\sigma_s \sqrt{2}} \right) \right),$$

where

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x \mathrm{e}^{-s^2}\,\mathrm{d}s\,.$$

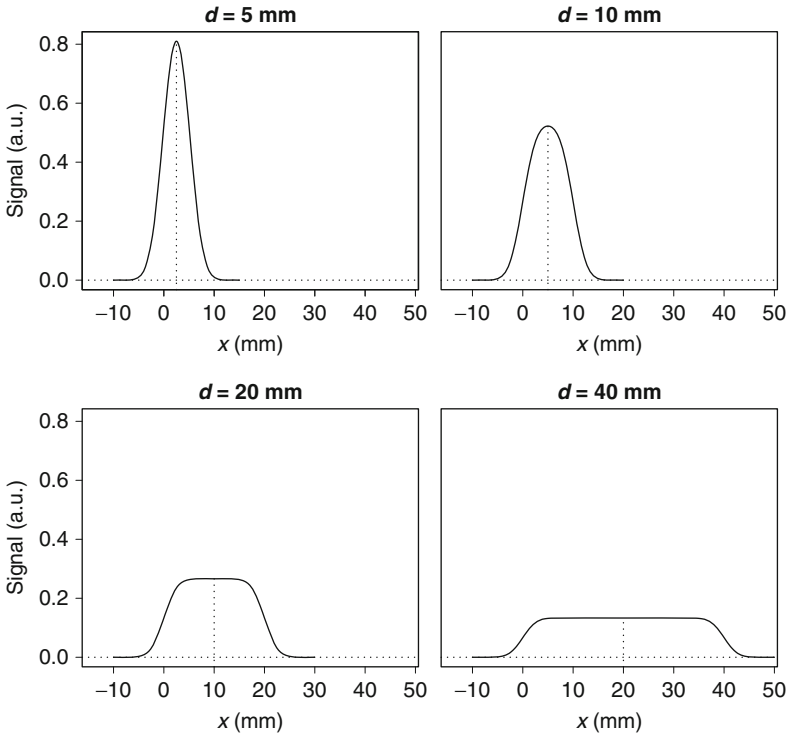


Fig. 3

Imaging a point source executing a uniform linear motion starting at the origin. The signal intensity is expressed relative to the maximum of the point source image at rest. The PSF is assumed to be a Gaussian with a FWHM of 5 mm. For motions where *d* is smaller than the FWHM, the dominant effect is image blurring (increased width of the profile, but only minor shape distortion). For larger motions, shape distortions (elongations) and a center-of-mass shift (equal to one half of the traveled distance) become apparent

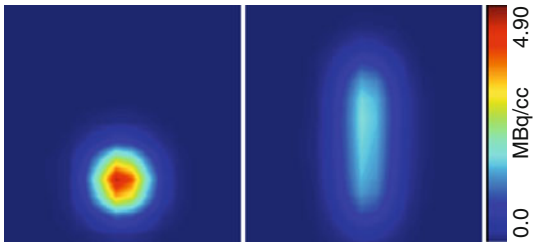


Fig. 4

PET image of a 1 mm point source (^{68}Ge , 5 min acquisition). *Left:* Source at rest. *Right:* Source moving a distance of 19 mm along the vertical during the acquisition

2.3 Rigid Motion

This type of motion is more or less strictly realized in brain investigations. Otherwise it can frequently be used as a reasonable (not necessarily very accurate) approximation of local motion in sufficiently small regions.

Rigid motions in 3D space are completely described by six parameters, the so-called *six degrees of freedom*. These include the three components of the translation ("displacement") vector **b** and the three angles of rotation around the three coordinate axes defining the orthogonal rotation matrix R. The complete transformation restoring all object points from their motion-affected positions $\mathbf{x}'$ back to the original positions **x** can be written as

$$\mathbf{x} = \mathsf{R} \cdot \mathbf{x}' + \mathbf{b}\,, \tag{5}$$

where the translation is assumed to be executed after the rotations and **b** and R are time-dependent functions. In the case of pure translations, all points are affected equally by the transformation (consider, e.g., the case of linear creeping motion described earlier). In the presence of rotations, however, the motion influence is position dependent, generally affecting off-center areas more than those close to the center of the FOV.

2.4 Nonrigid Motion

For investigations in the body stem, one quite generally faces the problem of nonrigid motions in the thorax and abdomen. This motion is mostly caused by the breathing-related organ motion which is quite strongly position dependent. In simple cases, these motions can be described by affine transformations of the image volume (Sect. 4.1). Otherwise, more general nonlinear deformation fields might be necessary, which in turn can in principle be used to correct for the motion if they can be determined with reasonable accuracy from the available motion sensor or tomographic image data. An alternative to specifying an explicit parametrization of the deformation field for the use with image registration algorithms (Sect. 4.1.1) is the optical flow technique (Sect. 4.1.2).

3 Motion Detection

In order to be able to perform any motion correction, it is of course necessary to first detect and analyze the motion. As far as the motion detection is concerned, there are two principal strategies:

Internal motion detection. By this we mean all methods deriving the motion correction based on sole analysis of the respective tomographic image data or the primarily acquired projection data underlying the tomographic reconstruction.

External motion detection. By this we mean all methods using motion detection via external sensors such as video cameras or dedicated motion tracking systems.

3.1 Internal Motion Detection

This approach is based on the idea of using some kind of time-resolved imaging. The subdivision of the imaging interval yields N tomographic data sets instead of a single one, thus allowing to

analyze the different data sets independently with the aim of determining adequate transformations for all of them in such a way that a common object position is achieved. Two main scenarios can be distinguished:

Dynamic studies. Dynamic studies are frequently used in PET for assessment of tracer kinetic transport processes, not as a means of performing solely motion correction. The chosen subdivision into so-called *frames* is thus usually determined not by the requirements of motion correction but rather by the properties of the tracer kinetics under investigation. Nevertheless, dynamic studies still offer the ability of analyzing the images of the different frames, determining the inter-frame motion and performing an automated mapping of all frames into a common reference frame. This approach is usually called *image registration*, and there exists a huge literature describing different algorithms, see, for instance, the reviews Maintz and Viergever (1998), Hill et al. (2001).

In short, the main idea is to use a certain *measure of similarity* between two 3D image volumes as a criterion in a nonlinear optimization. One such measure would be the covariance of the voxel intensities in both image volumes. Another one, which has proven especially valuable, is the so-called *mutual information*, see, e.g., Pluim et al. (2003). An overview of image registration methodology is given in ❯ Sect. 4.1.1.

Gated studies. Gated studies are used to investigate cyclic motions related to heartbeat and respiration. These are stroboscopic techniques mapping successive noncontiguous time intervals corresponding to a common phase of the cyclic motion into the same image volume. The information on the cyclic motion can be derived with systems such as an *electrocardiogram* (ECG) for cardiac studies or via a *respiration belt* (❯ *Fig. 5*). The result is a series of N image volumes ("*gates*"), which represent the cyclic motion time-averaged over the whole duration of the investigation. All registration techniques suitable for motion correction of dynamic studies can be applied here, too. The main difference is that in this case one quite generally is forced to use plastic, nonrigid transformations of the gates in order to achieve a common position, which can be challenging.

Internal motion detection has the distinct advantage of not requiring any additional devices/sensors in order to perform motion correction. However, in emission tomography, the

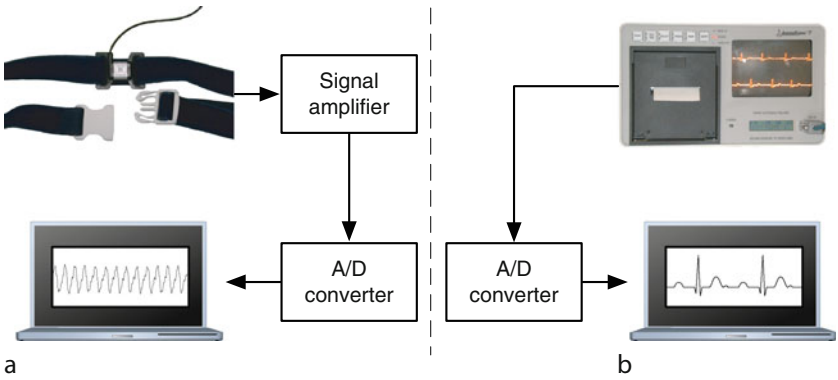


Fig. 5
Schematic depiction of external phase-sensitive sensors allowing to capture the different phases of the breathing cycle (a) or the heartbeat cycle (b)

major drawbacks are related to the generally limited count-rate statistics (low SNR) and modest spatial resolution of the tomographic images. Both factors impose limits on the available accuracy of the image registration (although it can be quite high under favorable conditions). Also affected is the achievable time resolution, i.e., the number of frames or gates which can be used without compromising tomographic image quality too much.

This situation might change with the advent of integrated PET/MRI systems, which potentially offer the possibility to utilize synchronously acquired MRI information for correction of the PET data, e.g., by using so-called *navigator sequences* (Ehman and Felmlee 1989; Sachs et al. 1994; Wang et al. 1996).

3.2 External Motion Detection

In contrast to motion detection directly in the projection or image data, motion can also be detected with external sensors. These sensors either register the motion directly (usually visually) or provide phase information for cyclic motions which can in turn be used for subsequent motion correction.

Originating in the field of computer animation, optical *motion tracking* devices allow to continuously monitor motion with high temporal (<0.1 s) and spatial (<1 mm) resolution. These systems can especially be applied to accurate tracking of head motion in emission tomography.

Another option are dedicated phase-sensitive sensors, which can be used to detect periodic motions. In the following section, we describe these devices in a bit more detail.

3.2.1 Phase-Sensitive Motion Sensors

Phase-sensitive sensors are often the first choice for detection of motions related to heartbeat and respiration. For the heartbeat-related motion, an *electrocardiogram* (ECG) is able to provide information regarding the different phases of the heartbeat cycle, e.g., the PR/QT intervals (*Fig. 5*).

Breathing sensors are used to derive information related to motion induced by respiration. One common device for providing such information is a so-called *respiration belt*. While earlier models where using a pneumatic electrical system, newer systems are frequently using a pressure-sensitive piezo crystal which provides a signal that is proportional to the pressure acting on it (due to the stretching of the belt).

After conversion of the analog signals to digital signals, analysis of the acquired data is performed by suitable software. This analysis finally provides the digital respiration or ECG curves and allows to identify the different phases of the cyclic motion in these curves from which the timing ("gating") information for decomposition of the emission tomography data is derived. This is illustrated for the example of a ECG-gated heart investigation in *Fig. 6*.

3.2.2 Optical Motion Sensors

The most obvious optical motion sensor is an ordinary video camera. Such cameras usually operate in the visible light spectrum and provide no differentiation per se between the monitored object and its surrounding (i.e., the image background). This necessitates substantial effort

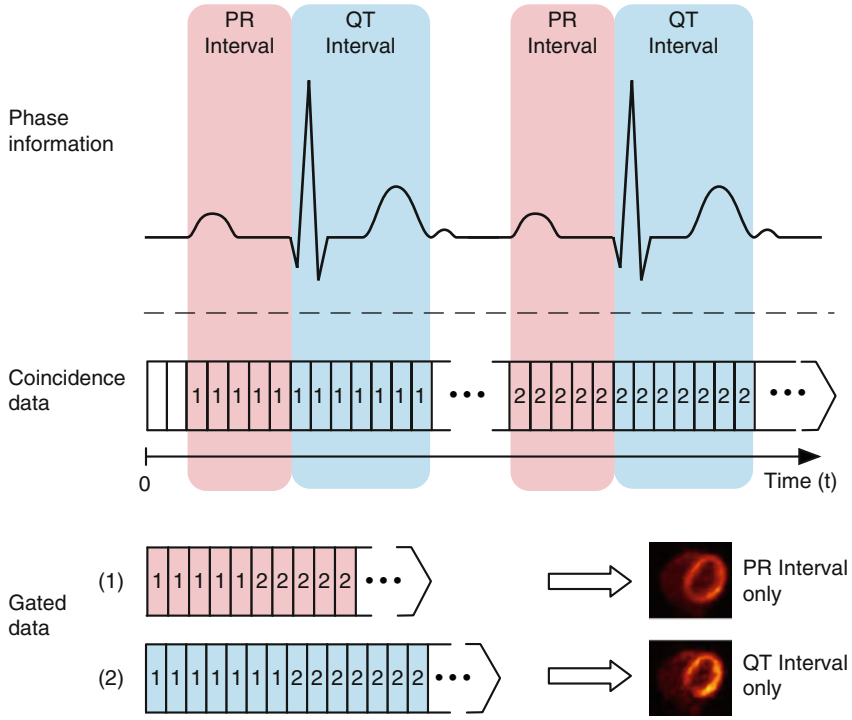


Fig. 6

Example of a gated cardiac study. By using the ECG-derived phase information (e.g., *PR* and *QT* phases), the coincidence data can be split into gates as indicated (only two gates are shown). As a result, images of the separate heart phases can be generated as shown in the *lower right corner*

during image analysis to distinguish the object of interest from the background. By attaching additional structures (which possess either a special color or shape) as motion targets to the object, the necessary image processing can be simplified by the facilitated identification of these targets within the captured image stream. The substantial computational burden for performing such image analysis in complex scenes on the large number of successive images necessary for a sub-second time resolution is one disadvantage of this approach. Sensitivity to variations in ambient light conditions is another problem.

Therefore, most optical high-precision motion sensors are deriving the motion information by monitoring an easily distinguishable *motion target* with dedicated video camera systems.[2] The motion target position is continuously calculated throughout the measurement where the target tracking is performed by capturing an image stream from the optical camera system.

[2] There also exist high-precision systems such as laser scanners which derive surface models of the monitored object without the need for special target structures. While these techniques are interesting, there are additional complications in deriving the relevant motion information related to the fact that no a priori known landmarks are present in the acquired data leading to the need for image registration in order to derive the actual subject motion. The future will show whether these systems will prove useful in the context of emission tomography.

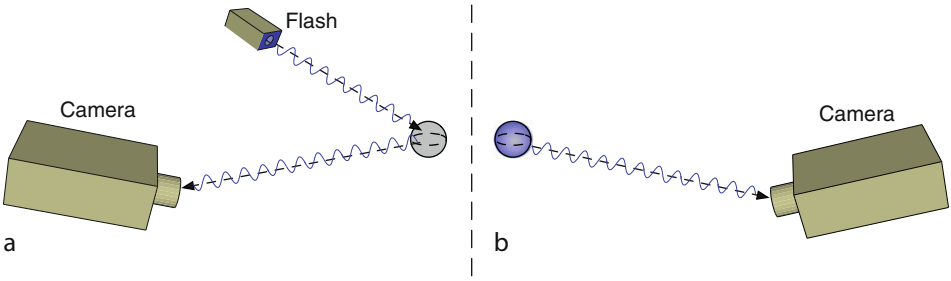


Fig. 7

The two types of motion tracking targets. In a passive motion tracking system (a) an external flash emits a stroboscopic light which is retro-reflected to the cameras by the target surface. In an active system (b) the stroboscopic light is emitted by the tracked target itself

The motion tracking system identifies the position of the target by performing a suitable image analysis in each captured image frame. For the calculation of the three-dimensional coordinates of the tracked target, stereoscopic setups with multiple camera systems monitoring the field of view are applied.

There are two choices for the type of tracking target which can be distinguished: *active* and *passive* targets. As shown in ❯ *Fig. 7*, passive targets are illuminated by separate light sources (often integrated into the cameras themselves). Stroboscopic illumination is used to discretize the motion detection process. Passive targets are designed in such a way that they strongly reflect the specific wavelength used in the illumination, while other structures are only poorly visible to the cameras at the respective wavelength. For these reasons, the illumination is usually performed in the infrared using pulsed LEDs for illumination. On the other hand, active targets carry themselves the stroboscopic light source whose emission is then detected by the tracking cameras.

While active systems have the advantage that the motion targets are brighter and thus more easily distinguishable, passive systems are often the preferred choice in the context of motion correction in emission tomography. This is due to the fact that active targets have to synchronize the flash frequency via cables between the cameras and the motion target. Moreover, active targets require LED light sources which have to be powered by batteries. Active targets are thus much more complicated and less suitable for routine use in patient investigations.

The motion tracking devices described above appeared only recently in medical applications (❯ *Fig. 8*) (Langner 2008). One major advantage is that they provide a good target-to-background discrimination due to the fact that the imaging sensors in these cameras are especially sensitive to the infrared frequencies used in the target illumination. Therefore, a high contrast in the captured images is achieved. This enables these systems to work effectively under conditions where a simple video-based solution might fail, e.g., due to changing ambient light conditions. Furthermore, the high target-to-background contrast facilitates the unambiguous identification of the target in the acquired images and thus allows fast determination of the spatial coordinates of all markers (repetition rates of more than 50 Hz) with submillimeter accuracy. Usually, tracking systems provide the derived motion parameters to the user rather than the raw image data. This eliminates the need to perform a separate data analysis.



Fig. 8
Passive infrared motion tracking system monitoring the head motion of a patient while undergoing a PET examination. The system monitors the position of several spherical markers attached to a pair of goggles and delivers the positional information (specified by the six degrees of freedom of a rigid transformation relative to the reference position)

4 Motion Correction

After having discussed the principal means of motion detection, we now address the question how to actually correct the motion and to obtain motion-free tomographic images.

Motion correction can be performed by manipulating the measured data at different levels. Which approach is most appropriate depends on the completeness of available motion information as well as on the available emission data. Last but not least, questions of the complexity of the correction procedure and the required processing time are relevant in deciding which approach to use. We present the principal possible approaches in turn, concentrating on typical applications of the respective techniques.

4.1 Image-Based Techniques

These techniques restrict themselves to correction of so-called *inter-frame* motion, where "frame" denotes a given tomographic image volume corresponding to the data acquired in a certain time interval.[3] This approach does not allow, however, to correct intra-frame motion. The implied assumption is that the dominant motion occurs between the frames and that the intra-frame motion can be neglected. The motion correction reduces in this case to applying an appropriate spatial transformation $\mathbf{T}$ to the image data which maps each point $\mathbf{x}'$ of the second (motion-affected) frame to its original position $\mathbf{x}$ in the first frame prior to the occurrence of motion. In the general case of arbitrary nonrigid motion, $\mathbf{T}$ is some invertible vector-valued function of the spatial coordinates:

$$\mathbf{x} = \mathbf{T}(\mathbf{x}') \,. \tag{6}$$

[3] We also include in this notation the case of stroboscopic (gated) acquisitions, where each frame corresponds to a certain phase of a cyclic motion (i.e., heartbeat or breathing cycle).

In many cases, it suffices, however, to assume that **T** is an affine transformation represented by an invertible 3 × 3 matrix A combined with a translation **b**:

$$\mathbf{x} = \mathsf{A} \cdot \mathbf{x}' + \mathbf{b}\,. \tag{7}$$

The most general affine transformation thus has 12 parameters, namely the 9 elements of the matrix A and the 3 elements of the translation vector **b**.

Affine transformations map straight lines onto straight lines (no warping). The transformation matrix A can be generated as a combination of an orthogonal rotation matrix R (parameters: three angles of rotation), a diagonal scaling matrix D (parameters: three scaling factors), and an upper triangular shearing matrix S with diagonal elements $s_{ii} \equiv 1$ (parameters: three shearing factors):

$$\mathsf{A} = \mathsf{R} \cdot \mathsf{D} \cdot \mathsf{S}\,. \tag{8}$$

In the special case of rigid motion, the transformation depends on the six degrees of freedom of this motion type, namely the three angles of rotation defining the orthogonal rotation matrix R and the three components of the translation ("displacement") vector **b**:

$$\mathbf{x} = \mathsf{R} \cdot \mathbf{x}' + \mathbf{b}\,. \tag{9}$$

Since we are dealing with digital image data, the components of the position vector $\mathbf{x}'$ take on only discrete values which can be chosen to be positive integers (i', j', k') representing the voxel coordinates in the given image volume. After transformation, the new coordinates will in general no longer lie on this predefined grid. This necessitates adequate interpolation to derive the transformed image volume. It is usually the best strategy to use the inverse transformation from **x** to $\mathbf{x}'$ and to determine for each destination voxel $\mathbf{x} = (i, j, k)$ its (in general non-integer) "source" coordinates (x', y', z') in the original volume. The corresponding value (image intensity) I_{ijk} can than easily be determined by linear interpolation in the given image data using the cube of eight voxels adjacent to (x', y', z').

The main challenge of performing motion correction in this way is of course the accurate determination of the correct transformation **T**. This information might be derived from the tomographic image data themselves or by external motion tracking devices. We discuss both possibilities in the next two sections.

4.1.1 Image Registration

In the absence of external motion information, the motion correction has to be derived from the two image volumes (frames) under consideration.

The most basic approach is to use so-called internal landmarks, i.e., structures which are identified interactively in both data sets. For rigid motion, using at least three different (not collinear) landmarks, the optimal transformation can be derived with the help of standard Least Squares methods. Shortcomings of this procedure are the need for user interaction and its usually only modest accuracy (given the limited spatial resolution, SNR, etc.).

Therefore, the usually preferred approach is to use fully automated registration algorithms. In the case of emission tomography, most algorithms currently used are not based on explicit feature extraction (edge detection, segmentation, etc.) but rather use integral statistical measures of image similarity.

One useful measure is the correlation coefficient of voxel intensities. Enumerating the voxels in the two image volumes, the image volumes are represented by the two vectors I_n, J_n defining the image intensity for each voxel. The image correlation is related to the covariance of the images, σ_{IJ}^2, and given by

$$C(I,J) = \frac{\sigma_{IJ}^2}{\sigma_I \cdot \sigma_J} = \frac{\sum_{n=1}^{N}(I_n - \bar{I}) \cdot (J_n - \bar{J})}{N \cdot \sigma_I \cdot \sigma_J}, \tag{10}$$

where the bar denotes the mean value over all voxels and the standard deviation σ_I is the square root of the voxel intensity variance defined by $\sigma_I^2 = \sum_n (I_n - \bar{I})^2/N$. Obviously, $C(I,J)$ becomes equal to one if the two image volumes are identical.

The image correlation is a sensible measure for intra-modality registration as is the case when performing motion correction. One of its shortcomings is that it is a quadratic measure that can react too sensitively to differences in regional image contrast (due, e.g., to dynamic redistribution of the tracer in successive frames).

Another measure of image similarity that is frequently superior[4] since it does not suffer from the mentioned shortcomings is the so-called *mutual information* which for our purpose can be defined as

$$M(I,J) = \sum_{i,j} P_{IJ}(i,j) \cdot \log\left(\frac{P_{IJ}(i,j)}{P_I(i)P_J(j)}\right). \tag{11}$$

Here, i, j enumerate the possible voxel intensities (values) in the two images (usually suitably binned to a discrete grid). $P_{IJ}(i,j)$ is the joint probability distribution. In practical terms, this is the normalized two-dimensional histogram of voxel intensities, where $P_{IJ}(i,j)$ represents the fraction of voxels having at the same time intensities in intensity bins i in image I and j in image J, respectively. $P_I(i), P_J(j)$ are the respective marginal distributions. In practical terms, these are the normalized voxel intensity histograms of the individual images. For a comprehensive review of the mutual information concept in the context of image registration, see Pluim et al. (2003).

In any case, the registration process can be visualized as operating on the two-dimensional histogram (or scatterplot) of the voxel intensities. For each set of parameters defining a transformation of the second ("test") image volume relative to the first ("reference") volume, the chosen measure of registration quality (e.g., the mutual information) is computed. Iterative algorithms such as a conjugate gradient search are then used to optimize the transformation parameters in such a way that the used measure of registration quality is improved until convergence is achieved. An example is given in ❯ *Figs. 9* and ❯ *10* in order to illustrate the difference in appearance of the histogram before and after registration.

❯ *Figure 11* demonstrates the different dependence of the correlation coefficient and the mutual information, respectively, on the degree of misregistration.

Registration is a very valuable technique and might be sufficient to perform adequate motion correction. Problems can (and frequently do) occur when the image characteristics differ too much or the information in one or both of the image volumes is insufficient due to a too low signal-to-noise ratio. Registration algorithms are rather computation-intensive and do require substantial time if applied to a large number of different image volumes, as can be the case in motion correction applications.

[4] and, contrary to the image correlation coefficient, is applicable to inter-modality registration (e.g. MRI vs. PET) as well

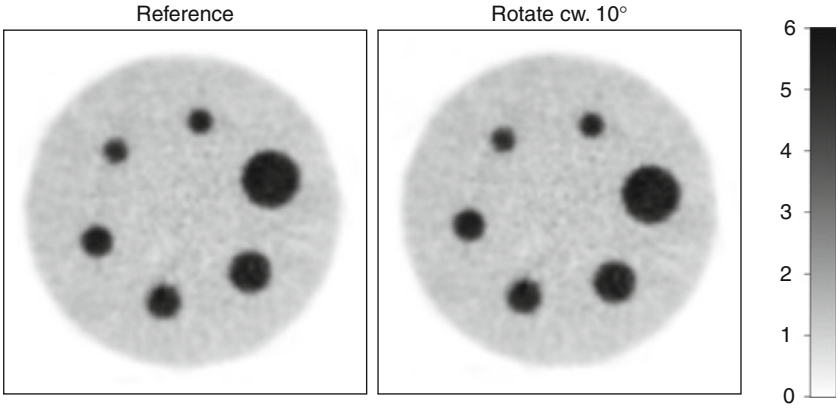


Fig. 9
A single transaxial PET image from a cylinder phantom study showing six hot spheres in a warm background. *Left:* original image. *Right:* the same image after clockwise rotation by 10 degrees. Corresponding two-dimensional histograms of the voxel intensities are shown in *Fig. 10*

Registration, however, is obviously not able to correct for intra-frame motion. By definition, image registration is a technique applied retrospectively to the given image data. The quality of the achieved motion correction therefore depends on a sufficient temporal resolution of the image data to be corrected. In the case of respiratory (or cardiac) gating, the number of different image frames necessary to resolve the motion is predictable and overall modest (about 5–10 frames usually suffice to reduce the residual breathing-related intra-frame motion to some 1–2 mm). In the absence of further substantial patient motion (unrelated to breathing or heartbeat), the intra-frame motion thus is essentially negligible and needs no longer to be considered.

For dynamic studies, the chosen number of frames usually is determined by the kinetics of the tracer and the necessity to include sufficient counts. Especially in the late phase of the investigation, frame durations usually are quite long, typically 5–15 minutes. It is obvious that during these long measurement times, substantial intra-frame motion might occur as well. Image registration in this case is at best able to correct the average displacement between successive frames.

The obvious strategy to reduce the motion influence is to decrease the frame duration sufficiently. But below a certain limit, image registration will no longer be possible due to the rapidly deteriorating image quality with decreasing frame duration. In order to follow this path, external motion tracking is the solution which is discussed in Sect. 4.1.3.

4.1.2 Optical Flow

An interesting alternative to specifying an explicit parametrization of the motion field for the use with image registration algorithms (Sect. 4.1.1) is the *optical flow* technique. It is particularly interesting for nonrigid motions.

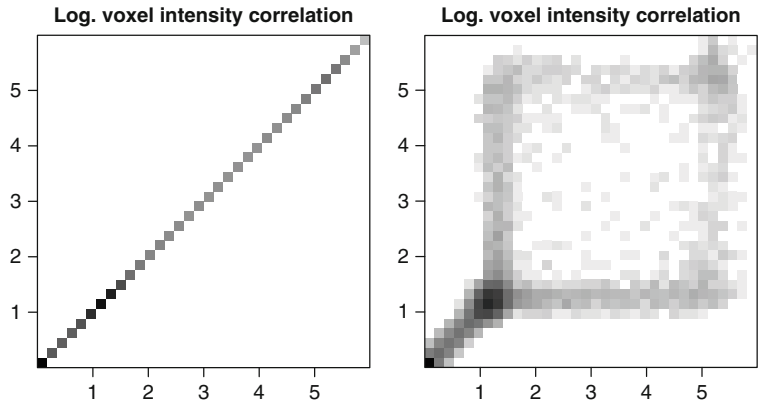


Fig. 10

Two-dimensional histograms of the voxel intensities for images shown in Fig. 9. Frequencies are color coded in a logarithmic inverse gray scale (white: empty bins). *Left:* Reference image versus itself. Obviously, in this case, one sees a perfect correlation and all off-diagonal bins are empty. *Right:* Reference vs. rotated image. Voxels with an intensity below a value of about one in both images still correlate strongly. This corresponds to background voxels (or voxels outside the object) which are mapped by the transformation to a new position which again lies in the background (or exterior) region, thus having essentially unaltered intensity (apart from statistical fluctuations). In this example, these regions do not carry much information regarding the relative orientation of both images. Voxels with higher intensities in one or both images, however, correspond to the hot spheres. Here, many voxels are mapped from the interior of the sphere to a point in the background (or the other way round), which leads to massive deviations from the ideal correlation. Exactly these regions contribute to the reduction of the correlation coefficient and the mutual information

In this approach, one estimates the deformation field from the incremental differences between subsequent images in a dynamic or gated series, see, e.g., Fleet and Weiss (2005). The central idea here is to treat the image intensities as a conserved scalar field $I(\mathbf{x})$ which obeys the continuity equation

$$\frac{\partial I}{\partial t} + \nabla \cdot (I\mathbf{v}) \equiv \frac{\partial I}{\partial t} + \mathbf{v} \cdot \nabla I + I \nabla \cdot \mathbf{v} = 0\,. \tag{12}$$

The flow of this quantity is assumed to be incompressible so that the divergence of the corresponding velocity field $\mathbf{v}$ vanishes:

$$\nabla \cdot \mathbf{v} = 0\,. \tag{13}$$

Together this yields

$$\frac{\partial I}{\partial t} + \mathbf{v} \cdot \nabla I = \frac{\mathrm{d}I}{\mathrm{d}t} = 0\,. \tag{14}$$

This is the well-known result that the total time derivative of the incompressible flow of a conserved quantity vanishes identically. Discrete approximations (in space and time) of this

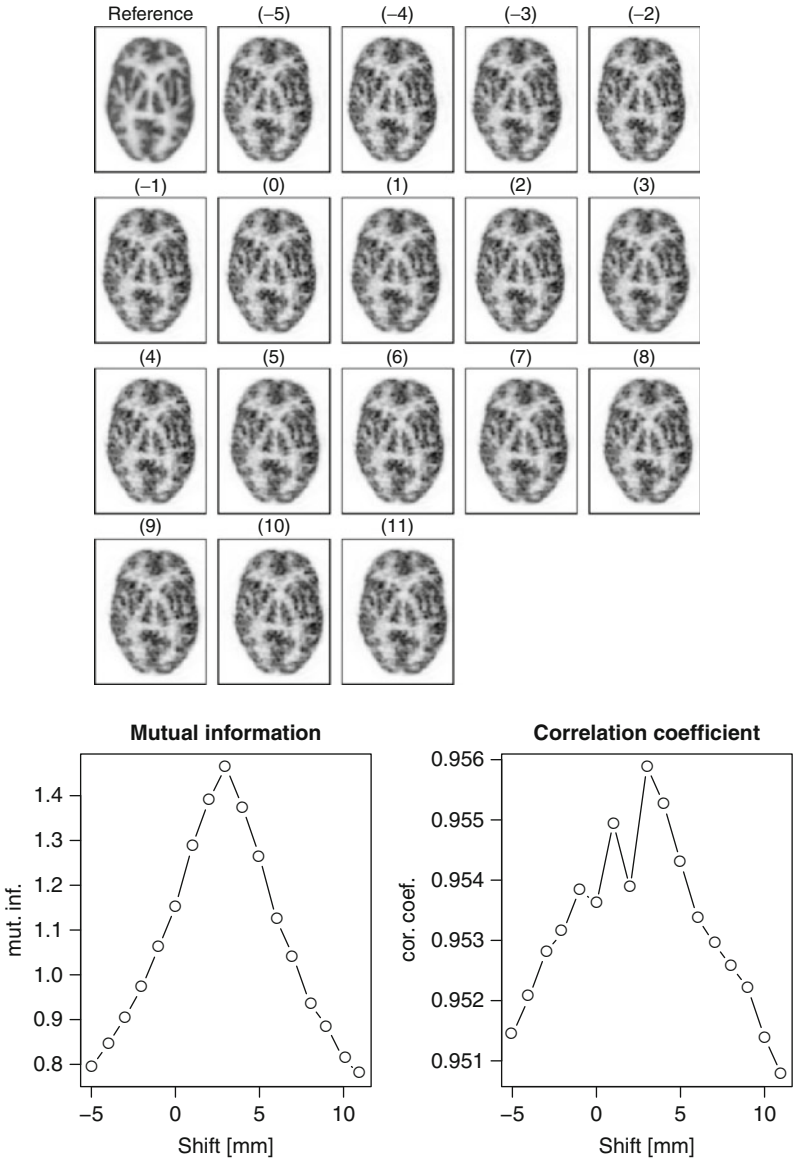


Fig. 11

A single transaxial PET image from two successive measurements with a brain phantom study where the alignment was not perfectly known between measurements. The images of the second measurement are labeled with the horizontal shift in mm against the first ("reference") image. Below the images, plots of the mutual information as well as the correlation coefficient between reference and shifted image are shown as a function of the shift. Mutual information is much more sensitive to the shift and identifies the optimal shift (3 mm in this example) more reliably than the correlation coefficient

equation plus further constraints are then used to derive the necessary transformations for the mapping between successive images (or, rather, image volumes) in a time series.

Numerically, optical flow algorithms are dependent on sufficiently low noise in the data as well as a sufficiently dense temporal sampling. Both assumptions are not always fulfilled in emission tomography. Nevertheless, this approach has proven to be a useful tool in this setting, see, e.g., Dawood et al. (2008). There are further limitations, however, which have to be kept in mind. An obvious one is the fact that the tracer redistribution occurring during the course of the measurement violates the assumed "conservation of image intensity." One example is the continuous accumulation of FDG ([^{18}F]2-fluoro-2-deoxy-D-glucose), the tracer ubiquitously used in oncological PET. Thus, the sum over all voxel intensities is not actually an invariant of a dynamic study. This does of course affect to some extent the ability of the algorithm to correct motion effects. For gated studies, on the other hand, this assumption can be considered to be fulfilled. Another issue is the assumption of incompressible flow, which will be a good approximation for soft tissue but not for the lung. Thorough validation of the algorithm in each special case is therefore mandatory.

4.1.3 Multiple Acquisition Framing

The *Multiple Acquisition Framing (MAF)* technique (Picard and Thompson 1997) uses an external motion tracking system to determine the time points where the cumulative motion relative to the previous time point has reached a certain threshold. The technique is essentially limited to the case of rigid motions and thus to brain investigations. This limitation is actually a limitation of the external motion tracking: only for rigid transformations, the detected motion will define the motion of all interior points in the patient. In the presence of nonrigid motion – especially breathing- or repositioning-related motion in the body stem – the external motion detection provides incomplete and thus at best only approximately correct information regarding local motion.

The motion threshold has to be chosen sufficiently small relative to the spatial resolution of the tomograph which in PET typically is about 3–7 mm nowadays. Every time this motion threshold is reached, the currently running frame is stopped and the next time frame is started.

Usually one cannot interface the motion tracking system to the tomograph's acquisition electronics or one faces hardware-related limitations in the number of allowed time frames. Therefore, the MAF technique is usually applied retrospectively to data acquired in list mode (Sect. 4.3) and the appropriate framing scheme is imposed retrospectively according to the motion tracking information. For older systems, list mode acquisition might not be routinely available, and major software and hardware modifications are required to use this approach (Langner et al. 2006). Recent scanners, however, support list mode acquisition routinely and are thus better suited for the MAF technique.

After framing the list mode data according to the selected motion threshold, all resulting frames are reconstructed in the standard way to obtain a set of image volumes. Motion blurring within each frame is then guaranteed to lie below the given motion threshold and will be small or negligible if the threshold has been selected sensibly.

In the next step, the motion tracking data are used to specify the necessary rigid transformation between the desired reference frame (e.g., the first or last one) and all other frames. This transformation is then executed in the reconstructed image space and maps all image volumes to a common orientation, (*Fig. 12*).

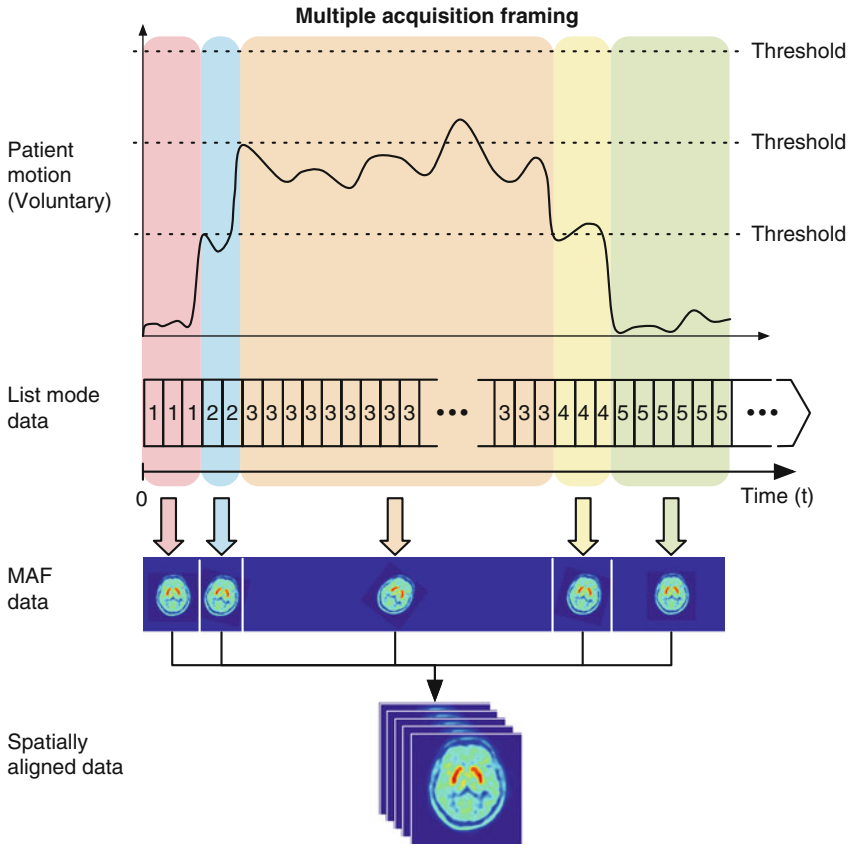


Fig. 12

Multiple Acquisition Framing (MAF) technique. External motion tracking is used to define a framing scheme for a list mode stream (Sect. 4.3) that guarantees that intra-frame motion remains smaller than the predefined threshold. After reconstruction, spatial alignment of all frames is performed according to the motion tracking data

The major advantages of the MAF technique in contrast to using retrospective image registration of a dynamic series with a fixed predefined framing scheme are twofold:

1. The ability to define an optimal framing scheme which reduces intra-frame motion below the preselected threshold
2. The ability to achieve good registration by using the motion tracking information even for low-statistics frames where coregistration algorithms would fail or provide only modest registration accuracy

The MAF technique is also attractive since it is computationally undemanding. For each image volume, it is only necessary to execute a single spatial transformation as defined by the motion tracking data. It has been applied successfully by different groups and has proven to be useful especially in the context of quantitative tracer kinetics studies (see, e.g., Herzog et al. (2005)).

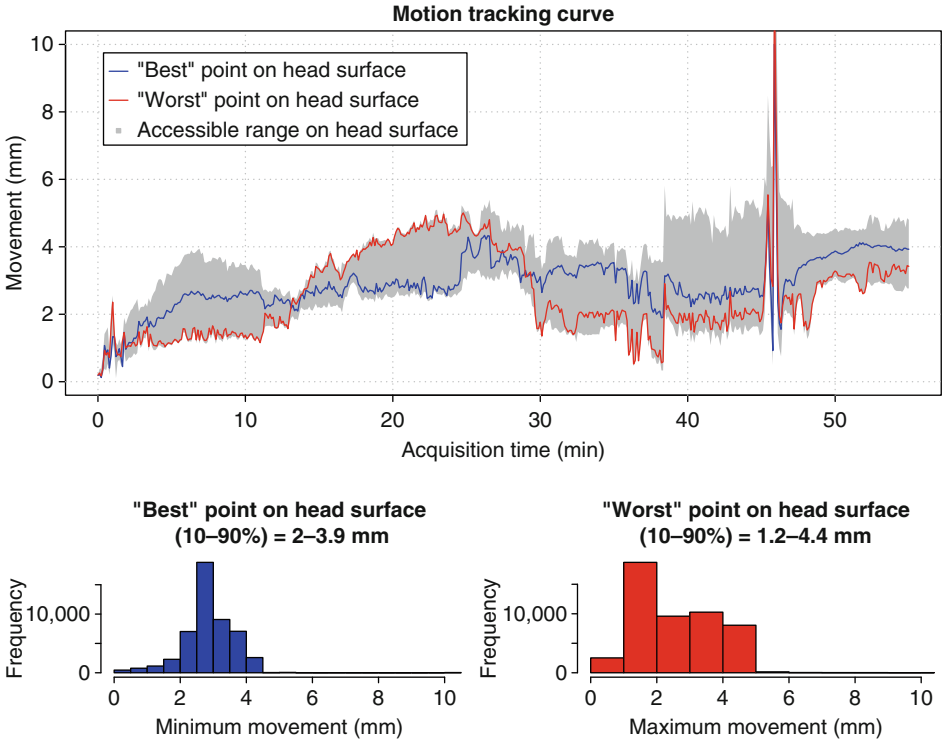


Fig. 13

Motion tracking data of time-dependent head motion. Motion is displayed as the distance of multiple grid points on the head surface from their respective position at some time chosen as zero point of the time axis (not necessarily the start of the acquisition). As a function of time, each individual grid point generates a trajectory lying within the *solid gray area* of the plot. In the presence of rotations, not all points are affected equally by the motion. The *blue* (*red*) line shows the trajectory of the grid point exhibiting least (most) motion (as defined by the extent of variation of the distance from the reference position). Histograms of the distances to the reference position for these points are given at the *bottom*

On the other hand, the technique has its limitations. The first one, namely restriction to rigid motion and thus to brain investigations has already been mentioned. Another one is the fact that the motion threshold which is used to discriminate between important ("large") and unimportant ("small") motions has to be chosen while respecting two potentially conflicting demands. For one, the threshold has to be distinctly smaller than the given spatial resolution of the tomograph. Otherwise the residual intra-frame motion would still be comparable to (or larger than) the given resolution and would thus lead to sizable motion artifacts (image blurring, etc.) within each frame. On the other side, the threshold cannot be chosen arbitrarily small since it has to stay above the amplitude of rapid "stochastic" motion. Superimposed on these rapid random motions of the patient slower and larger motions will usually occur. For illustration, an example of a typical motion tracking output is shown in ❯ *Fig. 13*.

Lowering the motion threshold below the amplitude of the rapid motion component would result in an unreasonable increase of the total number of frames – one could end up with literally hundreds or thousands of time frames – accompanied by a massive reduction of the count rates which ultimately prevents the tomographic reconstruction of acceptable images.[5] Even for higher thresholds, the MAF technique might produce several frames of very short duration and inadequate image quality thus necessitating suitable summation over frames to obtain a motion-corrected dynamic data set which than can be analyzed further.

4.1.4 Image Deblurring Approaches

Under favorable circumstances, another possibility for reducing the intra-frame motion effect can be applied, namely explicit or implicit deconvolution of the image. The blurring kernel used in the deconvolution, which generates the motion-affected image from the motion-free one, is defined by using a priori motion information.

As is well known, all deconvolution techniques are inherently amplifying noise so deconvolution is generally not an optimal strategy. Used conservatively, it nevertheless has its merits. Since we cannot get into the details of this technique, we refer the reader to the literature, e.g., Faber et al. (2009), Raghunath et al. (2009), Apostolova et al. (2010).

4.1.5 Correction of Breathing-Related Motion

Organ motion caused by the breathing cycle not only affects the lung but is substantial in most parts of the abdomen as well. With the ever-increasing importance of whole-body PET in the context of oncological investigations, notably for follow-up and treatment response monitoring, correction of breathing-related motion is attracting much interest in recent years.

The principal technique of motion detection is to use breathing triggers of different kinds. Most systems only provide phase information for a single phase of the breathing cycle, e.g., near end-inspiration for pressure sensors or midway between inspiration and expiration for air flow sensors. A few systems (e.g., optical motion tracking devices) allow to derive amplitude information as well and additionally allow to distinguish between deep and shallow breathing cycles.

The motion information is then used to create a gated series of image volumes, where each gate corresponds to a single phase or motion amplitude averaged over many breathing cycles. Like in ECG-gated cardiac studies, techniques are used to identify and exclude irregular cycles from this averaging in order to minimize the residual motion effect in the single gates. Investigations using some kind of respiratory gating are frequently labeled as "4D investigations" in the current literature.

The overwhelming majority of these studies is performed in the context of radiation therapy of pulmonary tumors and is, moreover, restricted to investigation of a single bed position. There is no principal difficulty which prevents the acquisition of gated whole-body (multi-bed) studies, but only very few groups have implemented this up to now. It seems safe to predict that this approach will draw increasing attention since it is well known that even in the lower

[5] Due to the noise and convergence characteristics of iterative image reconstruction, adding up the low-count images after reconstruction would not really solve this problem.

abdomen, breathing-related motion can lead to axial displacements of 1–2 cm and might thus be larger than in some lung areas.

Generally, a moderate number of gates (around 5 for amplitude-based gating or 10 for phase-based gating) is used in order to maintain a sensible SNR in the individual gates while still preserving a sufficient time resolution of the breathing cycle.

But even with only about 10 gates, the image quality of the individual gates is already seriously compromised and usually not sufficient for diagnostic evaluation. It is therefore necessary to use the complete gated data for generation of a single motion-corrected image volume. This can be done in many different ways, most of them operating in the image domain.

In order to achieve a globally correct registration of the individual gates, nonrigid transformations are obviously required. This, together with the mentioned limitations regarding the image quality in the single gates, causes serious problems for automatic coregistration algorithms. Several strategies have been investigated to improve this situation. One possibility is to use an anatomical motivated deformation model and impose this as a constraint on the coregistration. Interactive or automated identification of certain landmarks is required to make this work. Another possibility is the use of the *optical flow* technique described in ❯ Sect. 4.1.2.

A more modest approach is possible if only a few well-defined target structures, e.g., a lung tumor and/or a liver metastasis, are of interest. This will usually be the case in the context of radiation therapy. In this situation, it might be irrelevant to achieve a good average motion correction in the whole field of view. Rather, one is interested in getting optimal correction of the local motion effect in the vicinity of the target structure. This can be achieved much more easily than global correction by automatic or interactive tracking of the target structure over all gates and subsequent rigid translation. Simple nonrigid transformations of the test volume can also be used, such as a deformation depending only on the axial coordinate, i.e., on the position along the body axis, but not on the transaxial coordinates.

An example of the motion influence on the delineation of small lesions is given in ❯ *Fig. 14*. Motion correction in this case was performed by a very simple nonrigid transformation correcting only the dominant, cranio-caudal motion component. The parameters of the transformation were determined from interactive definition of the positions of a few landmarks (including the main target structures) in all gates.

❯ *Figure 15* demonstrates that the breathing-related motion not only affects the lung and upper abdomen but also the lower abdomen. Both studies were acquired as whole-body gated list mode studies, and a motion tracking system was used to provide amplitude-sensitive gating information.

The main obstacle that complicates accurate correction of respiratory motion is the limited statistical accuracy in the individual gates and the resulting deteriorated image quality. Optimized processing of the data is therefore especially important to achieve satisfactory results. An example is given in ❯ *Fig. 16*. Here the motion-averaged ungated data (a) are compared to a single gate at end-inspiration (b) and to a combination of principal component analysis (PCA) of the gated series – a technique previously applied successfully to gated cardiac PET (Narayanan et al. 1999) – with adaptive edge-preserving smoothing using a variant of so-called bilateral filtering (Aurich and Weule 1995). The latter yields the improved representation of the same single gate shown in (c). While the differences between ❯ *Figs. 16a* and *c* might seem slight, they actually have a substantial influence on quantitative measures, notably the lesion volume and the measured tracer uptake in the lesion (for a discussion of quantitative image analysis in the context of tomography see ❯ Chap. 41, "Quantitative Image Analysis in Tomography").

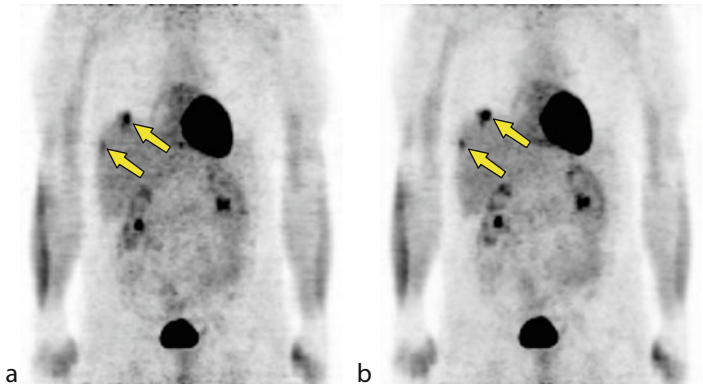


Fig. 14

Respiration-induced motion blurring of two liver lesions (a maximum intensity projection of the image data is shown). *Left*: without motion correction. *Right:* with motion correction. Note the better delineation and reduced partial volume effect after motion correction, especially for the small lesion

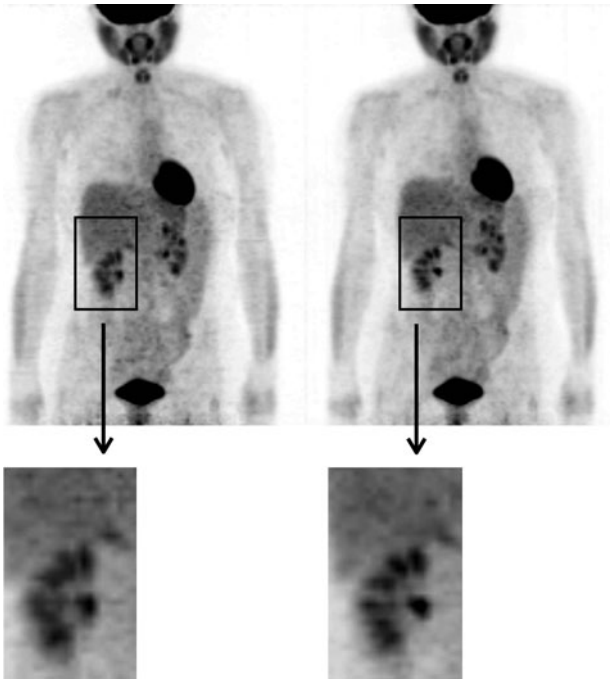


Fig. 15

Respiration-induced motion blurring in the lower abdomen (maximum intensity projection). *Left*: without motion correction. *Right:* with motion correction. Note the better delineation and reduced partial volume effect in the right kidney cortex after motion correction

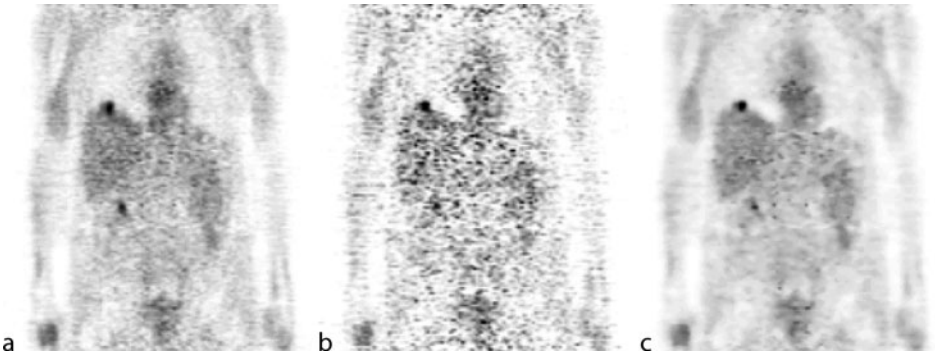


Fig. 16

Influence of image processing on obtainable image quality for single gates. A single coronal slice of the data used in *Fig. 14* is shown. (a) motion-averaged sum over all gates. (b) single gate (near maximal inspiration) (c) the same gate after analyzing and reconstructing the whole gated series with the help of principal component analysis and adaptive edge-preserving smoothing. This processing yields a very good delineation of the lesion at the top of the liver and concordance to the image in b while essentially maintaining a signal-to-noise ratio comparable to that of a

4.2 Projection-Based Techniques

Regarding inter-frame motion, one could try to perform the correction not in the reconstructed image space but rather in the space of the binned and histogrammed projection data (also called sinogram space) prior to reconstruction. Such techniques have been successfully applied to a number of special cases in emission tomography (mostly in SPECT, see, e.g., Fulton et al. (1994)) and more widely in X-ray computed tomography (CT). It seems fair to say that projection-based techniques are in general more suitable in CT than in emission tomography and are especially useful to correct for abrupt motions between sequentially acquired projections of a single scan. Another area concerns compensation for simple known motions such as center of mass shifts. However, for arbitrary motions correction in the histogrammed projections cannot be sensibly done. Rather, pre-reconstruction motion correction requires to perform all necessary operations on the level of the single coincidence event as described in the next section.

4.3 Event-Based Techniques

In the last sections, we discussed approaches of performing motion correction and elimination of the motion influence which operate on a time series of tomographic data sets either in projection space or, more often, in image space. All these approaches are limited to correction of the inter-frame motion between the successive image volumes resulting from the temporal decomposition of the available acquisition time. On the other hand, the intra-frame motion remains uncorrected by these techniques. We will now discuss approaches which can overcome

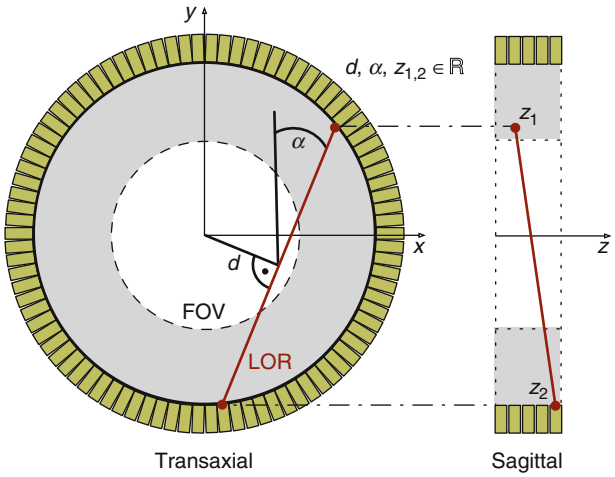


Fig. 17

Definition of LOR coordinates. *Left*: the azimuth angle α and the radial coordinate d define a plane parallel to the z-axis containing the LOR. The intersection of this plane with the x/y-plane (projection of the LOR onto the x/y-plane) is shown in *red*. *Right*: additionally specifying the two ring coordinates z_1, z_2, where the LOR intersects the detector rings, unambiguously defines the LOR in 3D space

this restriction. The central idea is to use the full time resolution of the acquired data by performing motion correction at the level of the individual detected event. This approach has been investigated especially in the context of PET.

A few remarks on terminology are necessary here: in positron emission tomography, each detected coincidence event is assigned to a certain *line of response* (LOR) connecting the two detectors which registered the event. All LORs together define the LOR space, where each discrete point defines a different LOR. A common choice for the coordinates in LOR space is the following (◉ *Fig. 17*): two coordinates are obtained by first projecting the LOR onto the transaxial (x/y)-plane. The LOR projection is then specified by the azimuth angle α (inclination relative to y-axis) and the radial coordinate d (distance to center of FOV). These two parameters thus define a plane parallel to the z-axis, which contains the LOR. Specifying in addition two ring coordinates z_1, z_2 along z, where the LOR intersects the detector rings, unambiguously identifies the LOR in 3D space.[6]

In a list mode acquisition, all events are stored individually in a common data stream. For each event, the coordinates of the LORs along which detection occurred are stored. Additionally, at regular time intervals (typically 1 ms) timing information and also status information (e.g., trigger signals) are inserted into the data stream. In this way, each single event is specified with a very high temporal resolution in contrast to the acquisition in histogram mode, where only the cumulative number of events registered along each LOR is stored.

[6] Four suitable coordinates unambiguously define a straight line in 3D space. This special choice of coordinates is convenient since α and d are also used in the 2D case. Moreover, the ring coordinates z_1, z_2 are directly provided by the front end electronics and are required anyway in further data processing.

In order to utilize the high time resolution of list mode data, availability of motion information with comparable time resolution is required. The only means to achieve this resolution is the use of external motion tracking systems as discussed in Sect. 3.2. These systems can provide a typical time resolution of about 20 ms, which – although inferior to the time resolution of the list mode data – is largely sufficient to correct the practically relevant motions.

Given a list mode data set together with the motion tracking information, two possibilities for motion correction exist:

1. Pre-correction of the list mode data followed by standard image reconstruction
2. Direct integration of the motion correction into the image reconstruction algorithm

These two approaches are discussed in the following sections.

4.3.1 Pre-correction of List Mode Data

This approach has been thoroughly investigated and applied successfully by several groups (Menke et al. 1996; Thielemans et al. 2003; Bühler et al. 2004; Livieratos et al. 2005; Langner et al. 2008).

In this approach, each registered coincidence event in the list mode data stream is assigned to a new LOR based on the motion information provided by the tracking device. After the corresponding spatial reorientation of all events, a new list mode stream results, and a motion-compensated image can then be generated via standard image reconstruction algorithms.

We will now explain in some detail how the event-based pre-correction of the list mode data is accomplished. The principal processing steps are illustrated in *Fig. 18*.

Step 1: Usually, the detectors registering a coincidence event at a certain time are stored in the list mode data stream by specifying two integer detector numbers (i, j). For further processing, it is necessary to convert these detector numbers to the Cartesian coordinates $\mathbf{p}_n$ $(n = i, j)$ of the respective detectors. This can easily be done, given the known geometry (ring diameter, detector size and number, etc.) of the scanner.

Step 2: The coordinates $\mathbf{p}_n$ are now spatially transformed according to the motion tracking data for the given time point. This yields new points $\mathbf{q}_n$, which in general are of course no longer located on the detector rings.

Step 3: By extending the straight line through $\mathbf{q}_i$, $\mathbf{q}_j$, one determines the intersection points with the detector rings yielding transformed detector coordinates $\mathbf{p}'_i$, $\mathbf{p}'_j$. The latter can also be interpreted as coordinates $\mathbf{p}_{i'}$ $\mathbf{p}_{j'}$ corresponding to some new detector numbers i', j'.

Step 4: Finally, the points $\mathbf{p}_{i'}$ $\mathbf{p}_{j'}$ are converted back to new detector numbers i', j'. These numbers are needed for defining the motion-corrected coincidence event in the resulting (motion-corrected) data stream.

The above-mentioned steps are repeated for all registered coincidence events where the number of events typically is of the order of 10^9 in patient investigations. The resulting motion-corrected events can then be further processed (histogram rebinning, image reconstruction) in the usual way.

Since this method is operating on the raw coincidence data, it does not interfere with the desired framing or gating of the data. It thus does not suffer from some of the previously

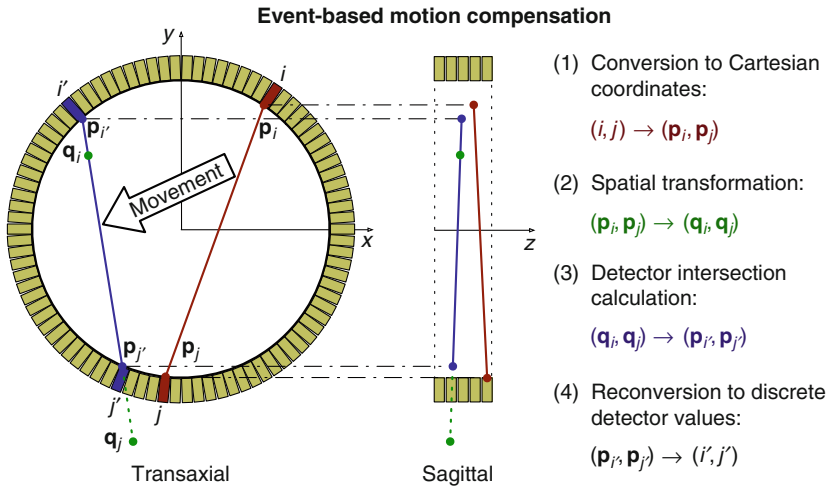


Fig. 18

Event-based motion compensation consists of four major steps: (1) Conversion of the discrete detector values (i, j) stored in the list mode format to points $(\mathbf{p}_i, \mathbf{p}_j)$ in 3D space using Cartesian coordinates. (2) Motion compensation by transformation of the points $(\mathbf{p}_i, \mathbf{p}_j)$ to $(\mathbf{q}_i, \mathbf{q}_j)$ according to the motion tracking data. (3) Calculation of the intersection points of the line through $(\mathbf{q}_i, \mathbf{q}_j)$ with the detector ring yielding the points $(\mathbf{p}_{i'}, \mathbf{p}_{j'})$. (4) Conversion of $(\mathbf{p}_{i'}, \mathbf{p}_{j'})$ to new discrete detector values (i', j') defining the motion-corrected coincidence event

explained disadvantages of the MAF technique. However, as should be obvious from the above description, it is significantly more complicated than MAF.

It is, moreover, necessary to take into account several refinements of the basic LOR transformation explained above: in the description given so far, ideal detectors with identical sensitivities have been assumed and the finite detector size has been ignored. Furthermore, the described transformation does in general move part of the LORs to new positions outside the FOV of the scanner. Neglecting these effects does lead to substantial artifacts and quantitative errors in the resulting motion-corrected tomographic images.

It is thus necessary to consider the following three corrections:

1. *Normalization correction* accounts for the detector efficiency differences when assigning an event to a different LOR.
2. *LOR discretization correction* accounts for the finite detector size by treating the LORs as a rectangular tubes rather than as lines.
3. *Out-of-FOV correction* accounts for LORs being transformed to new positions outside the FOV.

These corrections shall now be explained.

Normalization correction. Each detector crystal has a specific sensitivity which is determined by suitable calibration measurements and can be assumed to be known. This information is used

to perform a so-called normalization of the data prior to reconstruction. Essentially, each LOR (or rather the histogram bin accumulating all events detected along that LOR) is multiplied by a certain scaling factor resulting from the product of the relative sensitivities of the two detectors corresponding to the respective LOR.

Obviously, this procedure fails if the measured coincidence events are reassigned to new detector pairs by the motion correction since the normalization factor $\eta_{i'j'}$ of the target LOR will be different from the normalization factor η_{ij} of the source LOR. Since the event has actually been *measured* along the source LOR, it is necessary to use η_{ij} for normalization of the given event even if it is assigned to the target LOR (i', j') by the motion correction. With a suitable definition of the normalization factors, the normalization-corrected number of events in the histogram bin (i', j') is simply given by the sum of the normalization factors over all events assigned to this bin:

$$H_{i'j'} = \sum_{n=1}^{N} \eta_{ij}(n). \tag{15}$$

Here n enumerates the N events measured along different source LORs $(i(n), j(n))$ which are transformed to the common target LOR (i', j').

In order to be able to use the standard sinogram[7] -based image reconstruction available for the scanner, it is necessary to compensate for the unwanted implied multiplication by $\eta_{i'j'}$ by defining a "scaled" histogram

$$Z_{i'j'} = \frac{1}{\eta_{i'j'}} H_{i'j'} = \frac{1}{\eta_{i'j'}} \sum_{n=1}^{N} \eta_{ij}(n) = \sum_{n=1}^{N} \frac{\eta_{ij}(n)}{\eta_{i'j'}}. \tag{16}$$

In this way, when feeding the data into the rebinning and image reconstruction, the correctly normalized sum of events according to Eq. 15 is actually used. As can be seen from Eq. 16, the correct normalization can be obtained on the fly by histogramming the ratios of the normalization factors for the target and source LOR, respectively.

LOR discretization correction. In an idealized description, LORs are straight lines along which the γ quanta (those contributing to a coincidence event) propagate. Due to the finite detector size, there is no one-to-one correspondence between such an ideal LOR and a pair of detectors. When performing the spatial transformation of the LOR in order to achieve motion compensation, rather crude assumptions and approximations have to be made, usually treating the line connecting the centers of the detector faces as representing the LOR. This leads to discontinuous jumps of the LOR coordinates when crossing the border between adjacent detectors. These discretization errors lead to unacceptable artifacts in the resulting tomographic images.

The solution to this problem consists in using a much better geometric approximation by treating the LOR as a rectangular tube whose edges connect corresponding corners on the faces of the two contributing detectors. The motion-related spatial transformation of this "tube of response" then leads to the situation depicted in *Fig. 19*.

Figure 19a illustrates how a single LOR intersects the detectors after the motion-compensating transformation. Prior to this transformation, the LOR was connecting the faces of the two opposing detectors which detected the respective coincidence event. As can be seen, after transformation, the LOR in general intersects several neighboring detectors, each of which

[7] The distinction between histograms in LOR space and sinograms is not really important in the present context. Sinograms are generated from the histograms by a so-called re-binning process, i.e., a change to coordinates more suitable for image reconstruction.

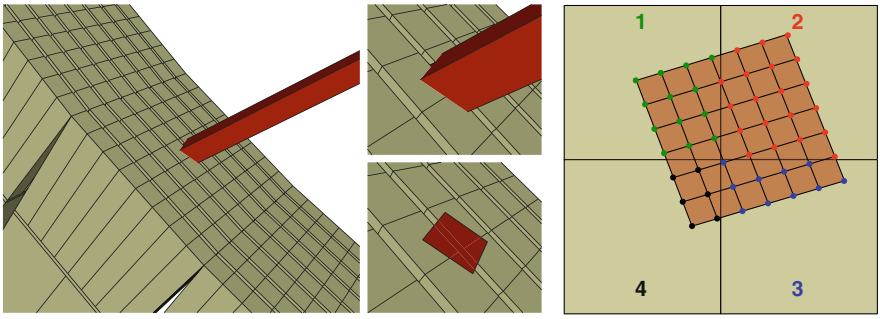


Fig. 19

Illustration of LOR discretization correction. (a) Intersection of a tube of response with multiple detectors after motion compensation. (b) Computationally efficient determination of intersection area by node counting on a superimposed grid

might have detected the event had no motion occurred. This localization uncertainty is handled by "distributing" the event over several target LORs in proportion to the respective intersection areas. However, an analytical computation of these areas is rather tedious and time consuming. *Figure 19b* illustrates a procedure for approximate determination of the intersection areas which still is sufficiently accurate but much faster than analytical computation.

Out-of-FOV correction. As explained previously, the basic idea of event-based motion correction is to reorient the LOR of a coincidence event registered at time $t = T$ (which we call LOR_T in the following) in such a way that LOR_T is moved back to its reference position at time $t = 0$. The problem might then arise that the motion-compensating transformation does reorient LOR_T to a new position which lies *outside* the FOV (*Fig. 20*).

Events which are reoriented to a position outside the FOV have to be discarded. They would have never been measured had the object not moved since no actual LOR within the FOV corresponds to such events. Discarding these events has the adverse effect of reducing the total number of events which ultimately contribute to the image reconstruction process. For practically relevant motions, however, the fractional count loss and the corresponding loss of statistical accuracy is usually modest and, therefore, tolerable.

There exists, however, a related and in a certain sense complementary problem which is more serious and needs to be addressed. It is also illustrated by *Fig. 20*.

Consider a single line of response LOR_0 defined in the reference position at time zero. LOR_0 can be regarded as being attached rigidly to the investigated object. In the presence of motion, the position of LOR_0 is a time-dependent function $\mathrm{LOR}_0(t)$. As long as LOR_0 does not leave the FOV, each coincidence event along LOR_0 is registered (with finite probability) in the usual way and can be reoriented to the reference position as described before. Differences in the efficiency of source and target LOR are accounted for by the normalization correction. If, however, LOR_0 leaves the FOV, further events along LOR_0 are no longer registered and are consequently lost from the cumulative counts assigned to the initial position of LOR_0 during the motion compensation process. This *Out-of-FOV effect* ultimately leads to inconsistent projection data and serious artifacts in the reconstructed images. Correction of this effect is therefore mandatory. It can be achieved as follows.

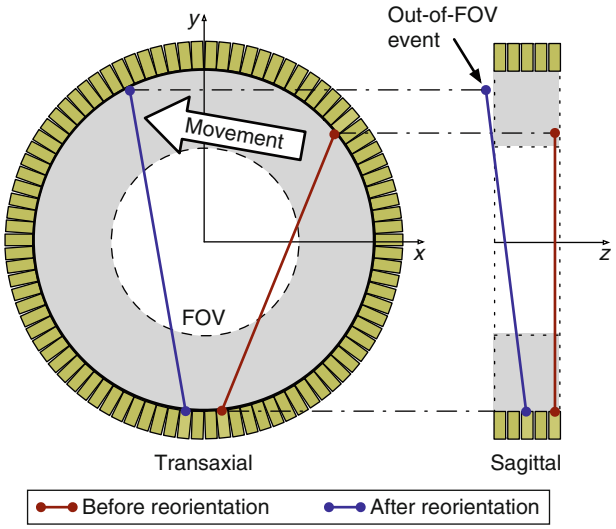


Fig. 20

Illustration of Out-of-FOV effects. (I): the motion-compensating reorientation of a line of response LOR_T corresponding to a registered coincidence event (*red*) might lead to a new position located outside the FOV (*blue*). The corresponding events have to be discarded but do not cause further problems. (II): alternatively, a valid line of response, LOR_0 defined in the reference position (*red*) might temporarily leave the FOV during the acquisition leading to unregistered "*Out-of-FOV events* (*blue*). The corresponding count loss along the LOR has to be corrected for

The correction is based on the fact that the decay-corrected count rate (the events per unit time) along LOR_0 can be considered as constant during the given time frame. This assumption is usually fulfilled to a very good approximation (otherwise the chosen framing scheme would be inadequate anyway). Therefore, the actually measured fraction of the total number of events occurring along $LOR_0(t)$ is simply given by the fraction f_{in} of the frame duration t_{frame} during which the LOR remained within the FOV:

$$f_{in} = \frac{t_{in}}{t_{frame}}. \tag{17}$$

In order to recover the correct total counts for LOR_0, it is then only necessary to divide retrospectively, i.e., after completion of the measurement, the cumulative number of measured counts which are assigned to LOR_0 by the motion correction by f_{in}.

In order to actually perform this so-called *Out-of-FOV correction*, it is necessary to track the position of *all* LORs during the frame duration and to compute the fraction f_{in} for each of them.[8] After Out-of-FOV correction, consistency of the projection data is restored and reconstruction of essentially artifact-free motion-corrected images is in principle possible.

[8] f_{in} is different for each LOR. Since the number of LORs typically is of the order of 10^8 and adequate time resolution is required, computation of the f_{in} is the most time-consuming step in the motion correction process.

A serious remaining problem of the described Out-of-FOV correction should be pointed out, however. The correction proceeds by upscaling the actually measured count rates along the different LORs. The scaling factors will become large for some fraction of the LORs if these happen to stay outside the FOV for most of the time. In this case, the statistical accuracy of the affected LORs is very small since the actually measured number of counts is small. The upscaling amplifies the absolute statistical errors as well. This heavily distorts the assumed Poisson statistics of the projection data and consequently leads to persisting problems during image reconstruction.

These effects can be minimized by using suitable frame boundaries and reference time points within the frames chosen in such a way that the f_{in} do not become too large. Suitable choices can be derived from an analysis of the motion tracking data alone. This approach is usually, though not always, sufficient to achieve satisfactory performance of the motion correction procedure (Langner 2008).

Another solution is discussed in ❯ Sect. 4.3.2 which outlines the possibility to directly integrate the motion compensation into the image reconstruction process.

Two examples now should serve to illustrate the achievable quality of event-based motion correction. ❯ *Figure 21* shows two sagittal PET images of a 1 mm ^{68}Ge point source. The source was moved randomly during a 5 min measurement. On the left, the image without motion correction is shown. On the right, the same data after motion correction are shown. As can be seen, the described correction is able to remove the motion influence completely in this case.

❯ *Figures 22* and ❯ *23* demonstrate the performance of event-based motion correction in a patient measurement investigating the time-dependent [^{18}F]DOPA tracer kinetics. ❯ *Figure 22* clearly shows the improved image quality after motion correction. In this example, the striatum (indicated by the arrows) is of special interest and delineated with much improved contrast.

❯ *Figure 23* displays time activity curves for three regions of interest (ROIs) positioned in the right and left *Nucleus caudatus* as well as in the occipital lobe. Without motion correction, there appear to be large differences between the left and right Nucleus caudatus, but these differences vanish after motion correction. Moreover, the shape of the time activity curves is heavily distorted without motion correction. These distortions would make a quantitative evaluation of the tracer kinetics impossible.

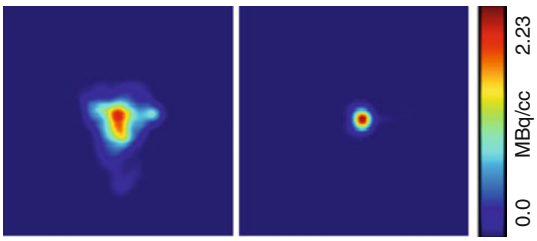


Fig. 21

Reconstructed images of a 1 mm point source (^{68}Ge , 5 min acquisition). The source was moved randomly during the measurement. *Left:* without motion correction, *Right:* with motion correction

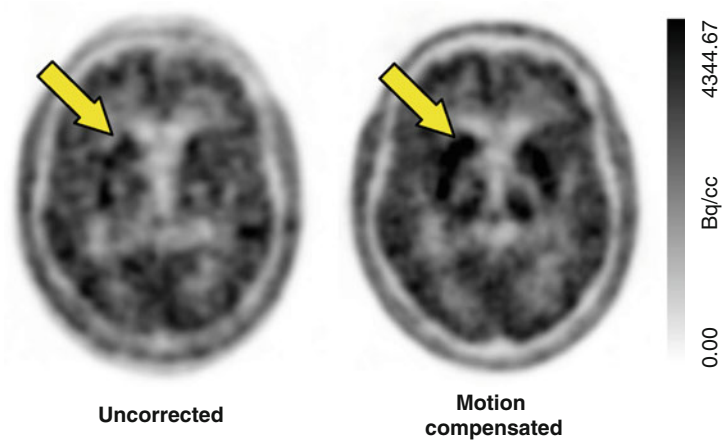


Fig. 22

Event-based motion correction in a patient measurement investigating the time-dependent [^{18}F]DOPA tracer kinetics. *Left:* without motion correction, *Right:* with motion correction. After motion correction, the image quality is generally improved and the target-to-background contrast in the striatum (indicated by the *arrows*) distinctly increased

4.3.2 Incorporation of Motion Correction into the Image Reconstruction Process

The ultimately most accurate and, in terms of achievable image quality, theoretically optimal approach is direct incorporation of the motion information into the iterative image reconstruction process, especially when using list mode reconstruction. This approach has been investigated more thoroughly only in recent years due to the considerable requirements in terms of computational resources required for list mode reconstruction.

In contrast to event-based pre-reconstruction correction, it is in principle very easy to incorporate all detected coincidence events into the list mode reconstruction even if their destination coordinates after motion correction are not part of the LOR space of the scanner. Otherwise these events have to be discarded. Moreover, certain problems related to rebinning and interpolation are reduced (which of course is true for list mode reconstruction in general, even in the absence of patient motion).

This technique has already been applied to ECG gating in cardiac investigations as well as to respiratory gating of oncological investigations of the thorax and abdomen. Here, it has been reported (Lamare et al. 2007a, b) that motion correction integrated into a one-pass list mode reconstruction is able to provide results superior to pre-reconstruction correction.

The improvement in this case has been attributed mainly to the ability to use the complete gated raw data during the reconstruction of any selected phase of the gated series instead of independently reconstructing the individual gates followed by motion correction and summing of the gates in image space.

It is obvious that this explicit consideration of the interdependence of the different gates will improve the noise characteristics of the resulting images. But the approach shares a central problem with the pre-correction approach: for nonrigid transformation, a considerable uncertainty

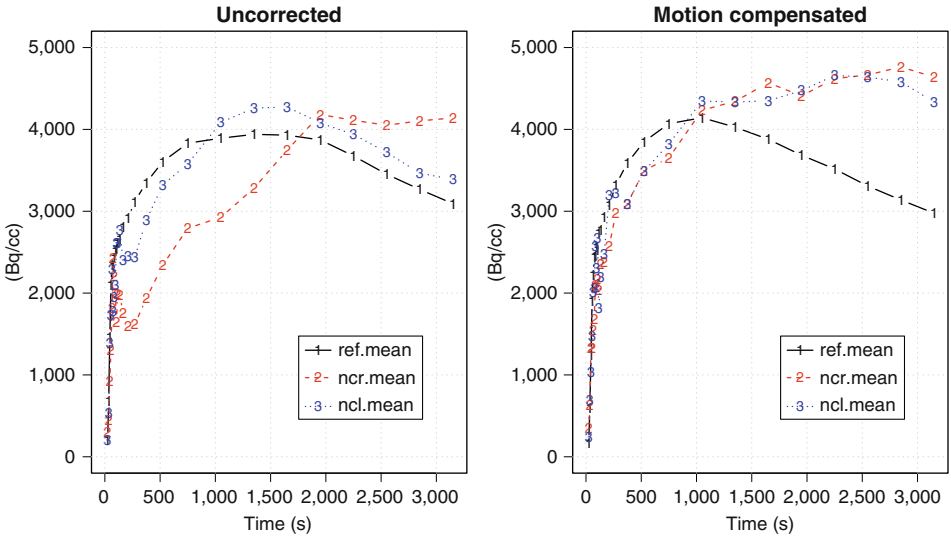


Fig. 23

Time activity curves (TAC) from the patient investigation shown in *Fig. 22*. Shown are the TACs for ROIs in the left and right *Nucleus caudatus* (labeled *ncl* and *ncr*, respectively) as well as in the occipital lobe (labeled *ref*). *Left*: without motion correction, *Right*: with motion correction. Without motion correction artefactual asymmetries between left and right Nucleus caudatus are visible. Moreover, the individual curve shapes are heavily distorted

regarding the actual time-dependent deformation field exists. The usual approach is to impose a fixed transformation field derived from the accompanying CT scan in PET/CT investigations. Since this motion information cannot be derived in real time during the PET investigation, such corrections can at best be approximate.

According to our knowledge, no comprehensive study exists which has compared pre- or in-reconstruction motion correction using an a priori determined deformation field with the much simpler image-based approach where the actual motion between gates is derived from the emission data themselves.

An interesting alternative to prescribing the deformation field is the use of a simple linear model for parametrization of the motion which is incorporated into a simultaneous reconstruction of all gates (Grotus et al. 2009).

Overall, the main obstacle, which so far has prevented serious use of reconstruction-integrated event-based motion correction in a clinical context, is the much higher computational burden associated with list mode reconstruction. The potential for progress in this area in the coming years is high, however.

5 Conclusion

The ubiquitous necessity for motion correction as a prerequisite for taking full advantage of the improved scanner hardware and image reconstruction software available today has been

recognized increasingly over the last years. Many approaches have already been investigated and new proposals are published constantly.

Existing motion detection techniques and correction algorithms differ markedly in their degree of accuracy and sophistication. While it is principally desirable to strive for the most accurate correction of any source of image degradation, from a practical point of view, the suitability for use in a clinical setting is obviously of paramount relevance.

In this context, it is certainly true that minimization of additional hardware requirements, such as motion tracking systems, and short processing times are important criteria. It has nevertheless been shown by several groups that these complications are perfectly manageable even when using sophisticated motion correction approaches for routine patient investigations. Therefore, integration of motion correction capabilities into commercially available tomographs (beyond the sole ability to perform gated studies) would be an attractive development.

Currently, the most important (and consequently most active) field seems to be improvement and general application of respiratory gating and motion correction in whole-body investigations. The advent of integrated PET/MRI systems might play a significant role in this context in the coming years.

Altogether, it seems therefore safe to predict that the next few years will see significant progress in establishing accurate motion correction as an integral and accepted part of clinical routine in emission tomography. Such a development will help to achieve a consistently high quality in diagnostic functional imaging.

References

Apostolova I, Wiemker R, Paulus T, Kabus S, Dreilich T, van den Hoff J, Plotkin M, Mester J, Brenner W, Buchert R, Klutmann S (2010) Combined correction of recovery effect and motion blur for SUV quantification of solitary pulmonary nodules in FDG PET/CT. Eur Radiol 20(8):1868–1877. doi:10.1007/s00330-010-1747-1

Aurich V, Weule J (1995) Non-linear Gaussian filters performing edge preserving diffusion. In: Mustererkennung 1995, 17, DAGM-symposium. Springer, London, pp 538–545

Bühler P, Just U, Will E, Kotzerke J, van den Hoff J (2004) An accurate method for correction of head movement in PET. IEEE Trans Med Imaging 23(9):1176–1185. doi:10.1109/TMI.2004.831214

Dawood M, Büther F, Jiang X, Schäfers KP (2008) Respiratory motion correction in 3-D PET data with advanced optical flow algorithms. IEEE Trans Med Imaging 27(8):1164–1175. doi:10.1109/TMI.2008.918321

Ehman RL, Felmlee JP (1989) Adaptive technique for high-definition MR imaging of moving structures. Radiology 173(1):255–263

Faber TL, Raghunath N, Tudorascu D, Votaw JR (2009) Motion correction of PET brain images through deconvolution: I. Theoretical develpment and analysis in software simulations. Phys Med Biol 54(3):797–811. doi:10.1088/0031-9155/54/3/021

Fleet DJ, Weiss Y (2005) Handbook of mathematical models in computer vision, 1st edn. Springer, Berlin. ISBN 978-0387263717 (Optical flow estimation, Chap. 15)

Fulton RR, Hutton BF, Braun M, Ardekani B, Larkin R (1994) Use of 3D reconstruction to correct for patient motion in SPECT. Phys Med Biol 39(3): 563–574. doi:10.1088/0031-9155/39/3/018

Grotus N, Reader AJ, Stute S, Rosenwald JC, Giraud P, Buvat I (2009) Fully 4D list-mode reconstruction applied to respiratory-gated PET scans. Phys Med Biol 54(6):1705–1721. doi:10.1088/0031-9155/54/6/020

Herzog H, Tellmann L, Fulton R, Stangier I, Rota Kops E, Bente K, Boy C, Hurlemann R, Pietrzyk U (2005) Motion artifact reduction on parametric PET images of neuroreceptor binding. J Nucl Med 46(6):1059–1065

Hill DLG, Batchelor PG, Holden M, Hawkes DJ (2001) Medical image registration. Phys Med Biol 46(3):R1–R45. doi:10.1088/0031-9155/46/3/201

Lamare F, Cresson T, Savean J, Cheze Le Rest C, Reader AJ, Visvikis D (2007ab) Respiratory

motion correction for PET oncology applications using affine transformation of list mode data. Phys Med Biol 52(1):121–140. doi:10.1088/0031-9155/52/1/009

Lamare F, Ledesma Carbayo MJ, Cresson T, Kontaxakis G, Santos A, Le Rest CC, Reader AJ, Visvikis D (2007b) List-mode-based reconstruction for respiratory motion correction in PET using non-rigid body transformations. Phys Med Biol 52(17):5187–5204. doi:10.1088/0031-9155/52/17/006

Langner J (2008) Event-driven motion compensation in positron emission tomography: development of a clinically applicable method. PhD thesis, Faculty of Medicine Carl Gustav Carus, University of Technology, Dresden, Germany, http://nbn-resolving.de/urn:nbn:de:bsz:14-qucosa-23509

Langner J, Bühler P, Just U, Pötzsch C, Will E, van den Hoff J (2006) Optimized list-mode acquisition and data processing procedures for ACS2 based PET systems. Z Med Phys 16(1):75–82. doi:10.1078/0939-3889-00294

Livieratos L, Stegger L, Bloomfield PM, Schafers K, Bailey DL, Camici PG (2005) Rigid-body transformation of list-mode projection data for respiratory motion correction in cardiac PET. Phys Med Biol 50(14):3313. doi:10.1088/0031-9155/50/14/008

Maintz JB, Viergever MA (1998) A survey of medical image registration. Med Image Anal 2(1):1–36. doi:10.1016/S1361-8415(01)80026-8

Menke M, Atkins MS, Buckley KR (1996) Compensation methods for head motion detected during PET imaging. IEEE Trans Nucl Sci 43(1):310–317. doi:10.1109/23.485971

Narayanan MV, King MA, Soares EJ, Byrne CL, Pretorius PH, Wernick MN (1999) Application of the Karhunen-Loeve transform to 4D reconstruction of cardiac gated SPECT images. IEEE Trans Nucl Sci 46(4):1001–1108. doi:10.1109/23.790811

Nehmeh SA, Erdi YE (2008) Respiratory motion in positron emission tomography/computed tomography: a review. Semin Nucl Med 38(3):167–176. doi:10.1053/j.semnuclmed.2008.01.002

Picard Y, Thompson CJ (1997) Motion correction of PET images using multiple acquisition frames. IEEE Trans Med Imaging 16(2):137–144. doi:10.1109/42.563659

Pluim JPW, Maintz JBA, Viergever MA (2003) Mutual-information-based registration of medical images: a survey. IEEE Trans Med Imaging 22(8):986–1004. doi:10.1109/TMI.2003.815867

Raghunath N, Faber TL, Suryanarayanan S, Votaw JR (2009) Motion correction of PET brain images through deconvolution: II. Practical implementation and algorithm optimization. Phys Med Biol 54(3):813–829. doi:10.1088/0031-9155/54/3/022

Rahmim A, Rousset O, Zaidi H (2007) Strategies for motion tracking and correction in PET. PET Clin 2(2):251–266. doi:10.1016/j.cpet.2007.08.002

Sachs TS, Meyer CH, Hu BS, Kohli J, Nishimura DG, Macovski A (1994) Real-time motion detection in spiral MRI using navigators. Magn Reson Med 32(5):639–645. doi:10.1002/mrm.1910320513

Thie JA (2004) Understanding the standardized uptake value, its methods, and implications for usage. J Nucl Med 45(9):1431–1434

Thielemans K, Mustafovic S, Schnorr L (2003) Image reconstruction of motion corrected sinograms. IEEE Nucl Sci Symp Conf Rec 4:2401–2406. doi:10.1109/NSSMIC.2003.1352379

van den Hoff J (2005) Principles of quantitative positron emission tomography. Amino Acids 29(4):341–353. doi:10.1007/s00726-005-0215-8

Wang Y, Rossman PJ, Grimm RC, Riederer SJ, Ehman RL (1996) Navigator-echo-based real-time respiratory gating and triggering for reduction of respiration effects in three-dimensional coronary MR angiography. Radiology 198(1):55–60

41 Quantitative Image Analysis in Tomography

Irène Buvat
UMR 8165 CNRS, Paris 7 and Paris 11 Universities, Orsay Cedex, France

1 Introduction . . . 1044

2 Motivation for Quantitative Image Analysis . . . 1044

3 Steps Required for Quantitative Image Analysis . . . 1046

4 Quantitation from Static Imaging . . . 1049
4.1 Preprocessing . . . 1049
4.1.1 Preprocessing in the Spatial Domain . . . 1049
4.1.2 Preprocessing in the Frequency Domain . . . 1051
4.1.3 Preprocessing Using Multiscale Methods . . . 1051
4.2 Measurements . . . 1051
4.2.1 Spatial Measurements . . . 1051
4.2.2 Intensity Measurements . . . 1054
4.2.3 Texture Measurements . . . 1055

5 Quantitation from Dynamic Images . . . 1056
5.1 Kinetic Modeling . . . 1057
5.2 Parametric Imaging . . . 1058

6 Conclusion . . . 1060

References . . . 1060

Further Reading . . . 1063

Examples of Free Software for Quantitative Imaging . . . 1063

C. Grupen, I. Buvat (eds.), *Handbook of Particle Detection and Imaging*,
DOI 10.1007/978-3-642-13271-1_41, © Springer-Verlag Berlin Heidelberg 2012

Abstract: In tomography, quantitative image analysis – or quantitation in short – is the extraction of parameters from an image or a set of images, as opposed to visual analysis. Quantitation is expected to provide objective, accurate, precise, reproducible, and efficient image interpretation, hence making the most of the signal delivered by the imaging device to the patient benefit. In this chapter, we explain the main steps required for quantitative image analysis, and give an overview of the class of methods appropriate for quantitation of static images, and also of dynamic image series.

1 Introduction

Generally speaking, image analysis is the extraction of information from pictures. In tomography, quantitative image analysis – or quantitation in short – is the extraction of parameters from an image or a set of images. Although the human eyes/brain combination is an extremely powerful image analyzer, the estimation of quantitative parameters characterizing the structure or phenomenon of interest presents undeniable advantages in the medical framework. However, accurate and precise quantitation of tomographic images remains extremely challenging in many cases, due either to the limited image quality (e.g., limited spatial resolution, noise or limited contrast), and/or to the complex relationship between the medically meaningful information and the signal displayed in the images.

Emission Tomography (single photon emission computed tomography – SPECT or positron emission tomography – PET) and transmission computed tomography (CT) have played a key role in the advent of quantitative medical analysis. Indeed, these imaging techniques directly provide digital images well suited to image processing and extraction of parameters, compared to radiographic films. Still, a digital image coded by a computer is "only" a big table of numbers, without any explicit anatomical or/functional information. It is therefore not straightforward to extract a medically meaningful parameter, or to design an algorithm able to do so. The aim of this chapter is to provide an overview of methods, tools, and bibliographic resources relevant for extracting meaningful parameters from tomographic images.

In ❯ Sect. 2, the role of quantitative image analysis in tomography is explained. In ❯ Sect. 3, the main steps required for image quantitation are presented. In ❯ Sect. 4, various approaches appropriate for quantitation of static images are described, while methods dedicated to quantitation of dynamic image series are described in ❯ Sect. 5.

2 Motivation for Quantitative Image Analysis

The motivation for designing quantitative image analysis methods is to achieve *objective*, *accurate*, *precise*, *reproducible*, and *efficient* image interpretation.

Objective means uninfluenced by possibly biased prior knowledge about the structure or phenomenon of interest. Objectiveness often goes with the absence of operator or observer dependency. It means that the parameter estimate should be similar, if not identical, for different observers, whatever their level of experience in the field.

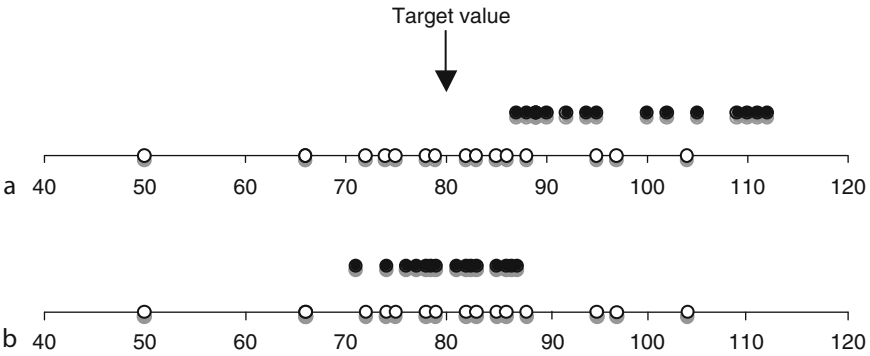


Fig. 1

Difference between accuracy and precision. (a) The *arrow* shows the target value (80). *Open circles* correspond to an accurate measurement method, meaning that on average, the measurement error is zero (or very close to zero). On the opposite, *black circles* correspond to an inaccurate measurement method, since the error in parameter estimate is always positive. (b) *Black circles* correspond to a more precise measurement than *open circles* that correspond to measurements presenting a high variability. Both measurement methods are accurate in (b)

The *accuracy* and *precision* notions should be clearly distinguished. Accuracy means that the interpretation is correct, on average. In quantitative image analysis, it means that if a measurement is repeated a large number of times on various images corresponding to different cases (like different patients), the average error in the estimate of the parameter of interest is zero (*Fig. 1a*: open circles, *Fig. 1b*: all circles). A precise measurement has a low variability, i.e., if the measurement is repeated a large number of times on different images, it will always yield a similar result (*Fig. 1*: black circles), although this result might be wrong. Ideally, one would like to get an accurate (i.e., exact) and precise (i.e., always the same) measurement. However, depending on the application, precision might actually be more important than accuracy. For instance, in image-based patient monitoring, if one always overestimates the volume of a tumor by about the same amount (i.e., poor accuracy but high precision), this might not prevent the radiologist from detecting a tumor regression and conclude at the treatment effectiveness.

Reproducibility is also called repeatability. Reproducibility is a frequent problem in visual interpretation of images: Visual interpretation is often significantly affected by the conditions in which the observer operates. For instance, the interpretation might differ if the image is not interpreted in the same environment (different lighting, whether the image is interpreted at the beginning of the day, or after a long series of image interpretation, etc). Reproducibility is a great advantage of image quantitation: Most algorithms output the same result when given the same input data and the same parameters. However, algorithms may still be sensitive to variations in the acquired data, such as different measurement noise, position of the patient, patient movement, etc. Ideally, if the same measurement is made on the same sample (or same patient), one would like to obtain the same result, that would be then qualified as perfectly reproducible or repeatable. Repeatability is usually assessed by what is called test-retest experiments, in which the same object of interest is scanned several times under the very same conditions, and the

parameter of interest is systematically estimated. The variability of the estimates, which can be measured by the coefficient of variation defined by the ratio of the standard deviation to the mean, gives an index of repeatability.

Finally, *efficiency* in image interpretation is always desirable, to reduce the time needed for the patient to get his/her scan results and to increase the throughput of the imaging department.

Many studies have shown the value of quantitative image analysis to improve differential diagnosis, prognosis, patient management, treatment planning, and patient monitoring, compared to visual analysis only. For all these applications, parameters measured from the images can be further analyzed using all sorts of statistical tools to determine the statistical significance of trends (for instance, signal intensity higher in diseased patients compared to healthy subjects) and deduce sound conclusions.

Two types of quantitation tasks should be distinguished, corresponding to relative and absolute quantitation.

Relative quantitation tasks refer to the estimate of dimensionless parameters. In CT for instance, measuring the ratio between the left and right kidney volumes is relative quantitation. On the contrary, absolute quantitation refers to the estimate of parameters expressed in a unit. Measuring the right kidney volume from a CT image is absolute quantitation. Absolute quantitation is often more demanding than relative quantitation, as it usually requires an addition calibration step. In emission tomography for instance, where the patient is administered a radioactive solution that the imaging device is supposed to detect and quantify, this calibration step consists in deriving the relationship between a count density (counts/voxel, measured by the tomograph) and a radioactivity concentration (Bq/ml), accounting for the imaging system sensitivity.

Relative quantitation does not yield anatomical or physiological parameters, unlike absolute quantitation, which is the appropriate approach to achieve noninvasive in-vivo measurements of such parameters. However, relative quantitation can often be sufficient for the medical interpretation of the data. Depending on the application, both types of quantitation are useful in the clinical or preclinical context.

3 Steps Required for Quantitative Image Analysis

Basically, medical image quantitation includes two steps. The first is obtaining images in which the voxel values (a voxel is a 3D pixel) do represent a well-understood physics quantity. We will call such images quantitative images thereafter. The second step is deriving medically meaningful parameters from these quantitative images.

The first step should not be overlooked. In emission tomography (SPECT and PET), a quantitative image is an image in which the voxel value is proportional to the activity concentration, expressed in Bq/ml for instance. Such a linear relationship is actually hard to achieve, and a large number of phenomena involved in the imaging process should be accounted for to approach it. These phenomena are summarized in ❯ *Table 1* (first two rows), and are explained in ❯ Chap. 37, "SPECT Imaging: Basics and New Trends" for SPECT and ❯ Chap. 38, "PET Imaging: Basics and New Trends" for PET. The relative importance of these different effects varies as a function of the application. A large number of corrections have been designed to correct for all these sources of nonlinearity, and references to papers describing the different correction strategies are provided in ❯ *Table 1* and in ❯ Chaps. 37, "SPECT Imaging: Basics and New Trends" and ❯ 38, "PET Imaging: Basics and New Trends." In short, current PET and

Table 1

Main effects to be accounted for in SPECT, PET, and CT to achieve a linear relationship between the voxel value in the images and the activity concentration (SPECT and PET) or between the voxel value and HU (CT)

SPECT	PET	CT
Attenuation through tissue of photons emitted by the radiotracer (Patton and Turkington 2008)	Attenuation through tissue of annihilation photons (Kinahan et al. 2003, Chap. 38, "PET Imaging: Basics and New Trends")	Ring artifacts (Abu Anas et al. 2010; Sadi et al. 2010)
Compton scattering of emitted photons in the patient and (to a lesser extent) in the detector (Hutton et al. 2011)	Compton scattering of annihilation photons in the patient and (to a lesser extent) in the detector (Chap. 38, "PET Imaging: Basics and New Trends")	Beam hardening artifacts (Kyriakou et al. 2010; Boas and Fleischmann 2011)
Spatial resolution variable across the field of view (Beekman et al. 2001)	Spatial resolution variable across the field of view (Alessio et al. 2006; Sureau et al. 2008; Stute et al. 2011)	Truncation artifacts (Kolditz et al. 2011)
Partial volume effect (Soret et al. 2006)	Partial volume effect (Soret et al. 2007)	Partial volume effect (Chap. 36, "CT Imaging: Basics and New Trends")
Patient physiological (cardiac and respiratory) motion (Chap. 40, "Motion Compensation in Emission Tomography")	Patient physiological (cardiac and respiratory) motion (Chap. 40, "Motion Compensation in Emission Tomography")	Patient motion (Chandler et al. 2011)

SPECT images are most often corrected for attenuation and scatter, and sometimes for variable spatial resolution across the field of view. The first two corrections ensure that in large and static structures, the voxel value is roughly proportional to the activity concentration, with possible errors of the order of 5–20% depending on the sophistication of the corrections. Partial volume and motion remain the two big issues that can still introduce large deviation from the proportionality relationship between signal level and activity concentration, but their impact strongly depends on the application. Partial volume effect (see below) mostly introduces significant biases in structures smaller than three times the spatial resolution in the reconstructed image (i.e., less than about 3 cm in diameter in SPECT and 2.5 cm in PET), while motion only introduces bias in areas strongly affected by respiratory or cardiac motions.

In CT, only quantitative images are used, as all CT images map the linear attenuation coefficient normalized by that of the water, in Hounsfield units (HU) defined as:

$$\mathrm{HU} = \frac{\mu_X - \mu_{\mathrm{water}}}{\mu_{\mathrm{water}}} \times 1{,}000, \tag{1}$$

where μ_X is the linear attenuation coefficient of material X, and μ_{water} is the linear attenuation coefficient of water for the X-ray energy range of the X-ray tube. When contrast media are used to enhance the organs or structures of interest, it is usually assumed that there is a linear relationship between the contrast medium concentration and measured HU value, although this should always be checked (Dawson 1999). As in emission tomography, a number of phenomena should be taken care of when interpreting CT values (Chap. 36, "CT Imaging: Basics

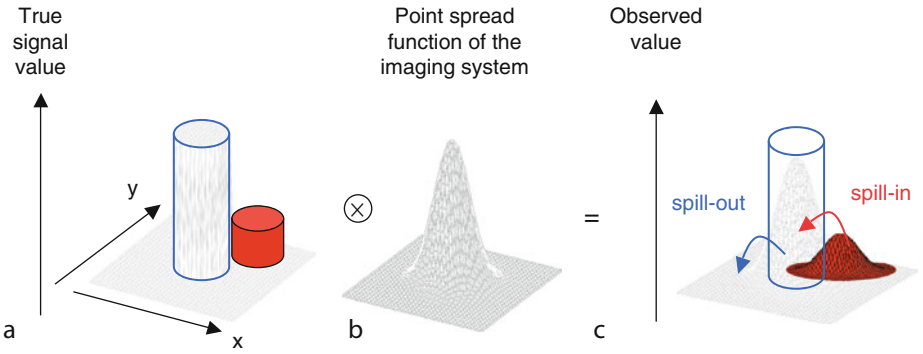


Fig. 2

Partial volume effect due to limited spatial resolution. The observed image (shown in c) is the result of the convolution of the real signal (in a) by the point spread function of the imaging system (in b). The result of the spread introduced by the imaging system is that signal coming from a structure of interest (in *blue*) is detected in the neighboring structure (in *red*), which is called spill-out. In addition, signal coming from the *red* structure is detected in the *blue* one (spill-in)

and New Trends"). These are summarized in *Table 1*, as well as associated references and in Chap. 36, "CT Imaging: Basics and New Trends."

A major issue affecting any medical imaging modality, including SPECT, PET, and CT, is what is often called "partial volume effect." This wording is actually confusing and includes two effects that should rather be distinguished: a spatial resolution effect and a sampling effect.

The spatial resolution effect is the blur resulting from the limited spatial resolution in all medical images. Spatial resolution is usually measured by the full width at half maximum (FWHM) of the point spread function obtained by imaging a point source. It can also be defined as the smallest distance that should separate two points so that they can be distinguished in the image. In current clinical images, spatial resolution is about 1 cm in SPECT, 6 mm in PET, and 1 mm in CT. These limited spatial resolution values imply that any structure of dimensions less than two to three times the FWHM will be affected by the so-called partial volume effect (Kessler et al. 1984; Soret et al. 2007). Consequently, signal coming from the structure will be detected in neighboring voxels (spill-out) and conversely (spill-in) (*Fig. 2*). The voxel value resulting from an unknown mixture of spill-in and spill-out will usually be a biased estimate of the real voxel value (be it activity concentration or HU).

The sampling effect is also often called "tissue-fraction effect." It comes from the fact that any image is sampled using voxels, and that the voxel borders do not follow the edges of the organ or structure of interest. A typical voxel side in PET or SPECT is 4 mm, and about 1 mm in CT (note that in CT, the voxel side is often larger in the axial direction, corresponding to the slice thickness, typically 3–5 mm). As a result, a voxel often includes a mixture of tissues with different signals (different activity concentrations or different HU), and the voxel value in the image is a weighted mixture of the real signal associated with each of the tissues present in the voxel (*Fig. 3*). This is especially true at the edge of structures. Given that it is often difficult to know the actual proportion of tissue in each and every voxel, signal specific to one tissue type should rather be estimated from voxels in which there is a high probability that only this tissue type is present.

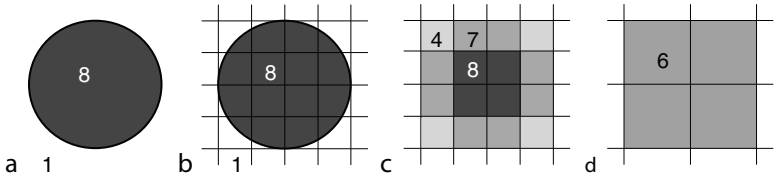


Fig. 3

Partial volume effect due to spatial sampling only. The tissue of interest is composed of values 8, surrounding by a background signal of 1 (a). After sampling (b) and with a perfect spatial resolution, the observed values are the weighted averaged of the signal coming from the tissues included in each voxel (c). For larger voxel size (d), the actual signal intensity of the structure of interest is underestimated (value of 6 instead of 8) because there is no more voxel containing only signal coming from the structure of interest

Overall, these two components of partial volume effect should always be kept in mind in image quantitation.

The second step of quantitative image analysis, consisting in deriving medically meaningful parameters from the quantitative images, is addressed in Sects. 4 and 5.

4 Quantitation from Static Imaging

In static imaging, quantitative parameters are extracted from a single image or image volume. To facilitate the estimation of relevant parameters, images might first undergo some preprocessing. The two main preprocessing steps used in medical image analysis are image enhancement and filtering. The motivation for these steps is to reduce noise and/or enhance contrast or spatial resolution so that the resulting image is better suited to the estimation of quantitative parameters than the original one. Image enhancement and filtering can fall into three categories: methods operating in the spatial domain, methods operating in the frequency domain, and multiscale methods which combine spatial and frequency data analyses.

4.1 Preprocessing

4.1.1 Preprocessing in the Spatial Domain

Preprocessing an image or image volume in the spatial domain consists in directly manipulating the voxel values. An extremely large variety of methods can be used to do so (Birkfellner 2010). The simplest approach consists in transforming the voxel values using linear, logarithmic, power-law, or piecewise linear transformation functions. Such transformations change the range of voxel values and the distribution of values within this range. Other approaches manipulate the histogram of the image, for instance performing histogram equalization. An image histogram plots the number of voxels for each voxel value, and the shape of the histogram gives a good description of the image contrast. Histogram equalization consists in modifying the voxel values so that each voxel value appears in the same number of voxels (thus yielding a

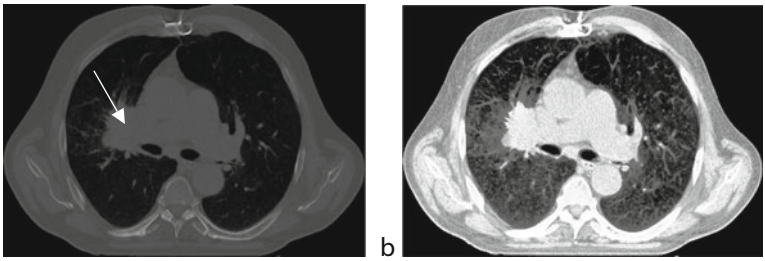


Fig. 4

Histogram equalization of the CT of a patient with a lung tumor (*arrow*). The original image is shown in (a), while the image after histogram equalization is shown in (b). Image (b) shows increased contrast compared with image (a): For instance, the vasculature of the lung tissue is clearly seen in image (b) unlike as in (a)

flat histogram). This operation redistributes the voxel values and stretches the dynamic range of the histogram, resulting in an overall contrast improvement. Note that after such transformations, the voxel value may lose its physical meaning. Yet, these transformations can be helpful to increase the contrast and facilitate subsequent image segmentation (*Fig. 4*).

Images can also be modified using all sorts of filters in the spatial domain, including smoothing filters and sharpening filters. Smoothing filters are largely used to reduce noise in SPECT and PET images. If smoothing filters can subjectively improve the image quality by reducing noise, the impact of the smoothing filter on subsequent quantitative analysis should be carefully evaluated. Indeed, smoothing filters reduce spatial resolution, hence increasing partial volume effect and associated biases (see Sect. 3). Sharpening filters are usually used to enhance borders between structures of interest in order to facilitate subsequent image segmentation. When properly normalized, filtering and sharpening filters can be designed so that they do not modify the overall signal intensity in the image, in which case the filtered image remains a quantitative image in which the signal has only been spatially redistributed.

If the point spread function PSF($\vec{x}$) affecting the image is known, where $\vec{x}$ is vector of 3 values representing the coordinate in the image volume, it is also possible to use a filter that would "undo" the action of PSF. This filter is called an inverse filter. The aim is to estimate the original image I that yielded the observed image I' with:

$$I' = I \otimes PSF. \tag{2}$$

The inverse filter F will be such that:

$$I' \otimes F = (I \otimes PSF) \otimes F \sim I. \tag{3}$$

So ideally, the inverse filter F should be defined by $PSF \otimes F = I$.

However, in tomography, I' always contains noise and the ideal F has to be approximated. Finding a good approximation for F requires modeling the noise present in I' and ultimately requires empirical adjustment of parameters. The approximate F will enhance the input image I', restoring some of the high-frequency content blurred by the PSF, and can thus be viewed as a particular enhancement filter whose derivation is driven by the knowledge of the PSF. However, it is practically impossible to design an approximate inverse filter that will perfectly restore I.

4.1.2 Preprocessing in the Frequency Domain

When preprocessing an image or image volume in the frequency domain, the idea is to modify the frequency content of the image instead of modifying the voxel values. Fourier analysis is a conventional way to manipulate the frequency content of an image. Values associated with each frequency in the Fourier domain reflect the pattern of intensity variations in an image. Low frequencies correspond to the slowly varying signal in an image, while the high frequencies correspond to fast signal change in an image, i.e., small structures, edges, but also possibly noise. Filters that reduce the high-frequency contents while preserving the low-frequency contents are called low-pass filters and are mostly used to reduce noise. Filters that attenuate the low-frequency contents while preserving the high-frequency contents are called high-pass filters and are used to enhance the edges of the structures present in an image. The impact of a frequency filter on the physical meaning of the voxel values in the filtered image should always be carefully analyzed before further processing.

4.1.3 Preprocessing Using Multiscale Methods

Preprocessing can also take advantage of multiscale methods, such as pyramid representation or wavelet transform, which can decompose an image into components, each describing features corresponding to a given spatial and/or frequency scale. In Fourier analysis, a coefficient of the Fourier transform describes the magnitude of this frequency in the whole image. When using a wavelet transform instead, a coefficient describes the magnitude of the frequency in a specific region of the image. Wavelet transforms are more appropriate than Fourier transform to describe sharp peaks and signal discontinuities that are rare in the image but that represent important signal.

Multiscale methods can be useful to enhance contrast, for instance by increasing the coefficients associated with the components corresponding to small details. Various wavelet analysis methods can be used to extract specific signal from a given image, manipulate it, and then perform the inverse wavelet transform (Aldroubi and Unser 1996). Unlike Fourier analysis that considers only the frequency content of a signal, wavelets are localized in both space and frequency. In addition, wavelet transforms have advantages over traditional Fourier transforms for representing functions that have discontinuities and sharp peaks.

4.2 Measurements

After optional preprocessing, images can be used to estimate parameters such as (1) dimensions and/or geometric characteristic of a structure of interest (spatial measurements), (2) intensity of the signal of interest in a given structure (intensity measurements), or (3) texture of the signal (texture measurements).

4.2.1 Spatial Measurements

Spatial measurements include measurements of dimensions (lengths, surfaces, volumes) and characterization of shapes (e.g., curvature). Examples of situations in which such parameters

Table 2

Tentative classification of the segmentation approaches that can be used in tomographic images to extract edges, surfaces, and volumes

Region-based approaches	Contour-based approaches	Pattern recognition approaches
Threshold-based on the voxel values	Methods based on derivation (Canny edge detector, Laplacian edge detector)	Parametric probabilistic classification (e.g., Gaussian mixture model)
Region growing	Active contour models	Supervised nonparametric classification (e.g., k-nearest-neighbor, artificial neural network)
		Unsupervised clustering (e.g., fuzzy C-mean algorithm)
		Approaches using contextual priors (Markov field-based approaches)
		Approaches based on mathematical morphology

are needed include tumor delineation in PET or CT for treatment planning in radiotherapy and left ventricle delineation in SPECT for calculating the stroke volume and ejection fraction in cardiac imaging.

Spatial resolution and contrast are key features that affect the accuracy and precision with which spatial measurements can be performed from tomographic images. Such measurements require either the identification of points of interest (for instance to measure the largest diameter of a tumor from a CT image), or the segmentation of the structure of interest. In both cases, manual, semiautomatic, or automatic methods can be used. Manual identification or delineation of structures can be time consuming (especially when the task is segmenting a volume from a stack of slices) and suffers from intraobserver and interobserver variability. This is why many efforts are dedicated to the design of semiautomatic and automatic methods. Such methods can be split into three categories: region-based approaches, contour-based approaches, and pattern recognition approaches. Each category itself includes a large number of methods, a list of which is provided in *Table 2*. We briefly describe the principles of the different methods and give examples of applications in PET, SPECT, and CT.

Region-based approaches are largely used in the context of emission tomography (SPECT and PET). Among the region-based approaches, threshold-based approaches are frequently used to segment tumors in PET (Zaidi and Naqa 2010). The idea is to consider that all voxels with a value greater (or smaller when looking for a low-uptake region) than a threshold (that can be updated during an iterative process) belong to the structure of interest. The chosen threshold is a key parameter, and additional constraints are often added to account for the spatial consistency of the resulting region (like obtaining a compact region, i.e., a region that cannot include two distinct parts).

Region-growing methods (Pohle and Toennies 2001) have been proposed for PET image segmentation. The segmentation process is started from a seed voxel, and adjacent voxels are

included in the region if they fulfill some clustering criteria (Green et al. 2008). Compared to threshold-based approaches, region-growing methods have the advantage of intrinsically resulting in compact regions.

Contour-based approaches basically look for signal discontinuities in the images to identify edges. Edges can be identified using derivative operators (e.g., Canny-Deriche, Laplacian) or active contours. The use of derivative operators has been proposed to segment structures in PET and SPECT, but sophisticated preprocessing is often needed to transform the original images and make them appropriate for contour-based segmentation. In the active contour approach, an initial contour matching the shape of the structure of interest is set and its location and shape are then adjusted to optimize a functional. The use of active contours is well suited to CT image segmentation (Truc et al. 2010), given the high resolution of the CT images compared to SPECT and PET images.

Pattern recognition approaches are most often based on classification methods. Here, the segmentation problem is seen as a labeling task, where each voxel is assigned to a given class based on the voxel features. These voxel features can be the voxel value only, but can also include additional properties, like the gradient of the voxel with respect to neighboring voxels. The voxel features are mathematically represented by a vector of values, each value corresponding to a feature. Several types of methods can be used to solve the classification problem based on the voxel features (see ❯ *Table 2*):

- Parametric probabilistic approaches can be used. As an example, Gaussian mixture models (GMM) have been used to segment tumors in PET (Aristophanous et al. 2007).
- When no statistical prior is available, supervised classification can be performed. Supervised classification is the process of deriving a mathematical function that can predict the membership of a class based on input data, here the voxel feature vector. The function is derived from voxel examples for which the segmentation results are known, also called training voxels. A simple example of supervised classification is the K-nearest-neighbor approach: A voxel is classified by comparison with the K voxels of the training voxels that are closest to the voxel under study. The distance between voxels is measured based on the Euclidian distance between the feature vectors corresponding to the voxels. The most frequent class among the K training voxels is then assigned to this voxel. Fuzzy segmentation, i.e., segmentation in which a voxel can belong to several classes with different weights, can also be achieved using such supervised classification, by assigning the weight associated with each possible class as a function of the proportion of training voxels belonging to that class. Fuzzy segmentation is especially appealing to account for partial volume effect, which results in mixture of signal inside voxels (cf. ❯ Sect. 3). Supervised classification can also be performed using a supervised artificial neural network (SANN) approach. In a SANN, the weights involved in the model are also determined based on training voxels during a learning phase. After training, the SANN is used to segment new data. The advantage of a SANN approach compared to simpler supervised classifications is that complex relationships between the input (i.e., the voxel features) and the class it belongs to can be modeled. In all supervised approaches, the training voxels should include all sorts of representative cases that should be accurately segmented for the subsequent segmentation algorithm to be accurate.
- Unsupervised voxel classification (also called clustering where a cluster corresponds to a class) can also be used. In this case, the features representative of each class are extracted from the classified voxels themselves. An example of such unsupervised algorithm used in

nuclear medicine (Belhassen and Zaidi 2010) and CT (Kakar and Olsen 2009) is the fuzzy C-mean algorithm (FCM). FCM algorithm minimizes the intra-cluster variation. The labeled voxels are assigned to the nearest cluster based on their weighted distances to the cluster centroids. New cluster centroids are then calculated and the voxels are reassigned. The process is iterated until all voxels are assigned to a fixed class.

- A fourth category of pattern recognition approaches includes those making use of contextual priors, i.e., of information in neighbor voxels, through the Markov field theory. Markov field theory has been used for image segmentation in PET (e.g., Hatt et al. 2007; Montgomery et al. 2007).
- Finally, segmentation approaches using mathematical morphology (Shih 2009) have been proposed. Mathematical morphology is the branch of mathematics dedicated to the analysis and nonlinear processing of geometrical structures, based on set theory, lattice theory, topology, and random functions. Like the active contour model approach, mathematical morphology uses priors regarding the shape of the structure of interest. Beyond basic mathematical morphology operators (closing, opening, erosion, dilation) that can be combined to extract specific geometric information about the image structures, more advanced algorithms such as the watershed algorithm (Beucher and Lantuéjoul 1979) are used in medical image segmentation, for instance in PET (e.g., Geets et al. 2007) and in CT (e.g., Ray et al. 2008). In the watershed algorithms, gray scale images are considered as reliefs and the gradient magnitude of each pixel is treated as elevation. Watershed lines are defined to be the voxels with local maximum of gradient magnitude. The segmentation procedure identifies the watershed by successive flooding of the gray value relief.

4.2.2 Intensity Measurements

Signal intensity measurements are usually performed within a bidimensional (2D) or tridimensional (3D) region previously segmented (Sect. 4.2.1). Most often, the mean in the region of interest is calculated. Ideally, an uncertainty should be associated with the mean to estimate the reliability of the measurement. In practice, the standard deviation associated with the mean is often calculated as an uncertainty index, based on the set of voxels included in the segmented region. However, this is not theoretically correct for two reasons:

- First, in tomographic images, noise is spatially correlated (due to the backprojection process involved in tomographic reconstruction, values and noise along a backprojection ray are correlated). Hence, the standard deviation calculated over a set of voxels is not identical to the standard deviation associated with the mean obtained when repeating the experiment a number of times (non-ergodicity). Replacing the sample standard deviation by the region standard deviation results in overestimating the reproducibility of the measurement.
- Second, the uncertainty associated with the measurement includes two components: the variability of the error (e.g., variation of the error from one patient to another, which depends on the robustness of the method in different operating conditions) and the lack of repeatability (e.g., the fact that the same measurements performed in the same patient in similar conditions will not give exactly the same value, see Sect. 2). The standard deviation associated with the mean in a given region is only a biased estimate of the lack of repeatability and disregards the variability of the error due to varying operating conditions.

The drawback of using the mean value in a region as an estimate of the signal intensity (radiotracer concentration or HU) is that this measurement is highly sensitive to partial volume effect in small structures. To reduce partial volume effect, a single voxel value is sometimes used to characterize tumor uptake in PET. This means that a single voxel value is assumed to be representative of the tumor metabolism. Actually, PET images used for such measurements are usually heavily smoothed during or right after reconstruction. Therefore, a single voxel value can also be viewed as a weighted average of surrounding activity, where the averaging process has been included during the reconstruction step or during post-smoothing. Yet, summarizing the metabolic activity of a tumor based on a single voxel might appear to be a reductive approach. To get a more complete description of signal intensity in a region, texture analysis can be considered (see ❯ Sect. 4.2.3).

In emission tomography, the tracer uptake is obviously a function of the injected activity and of the blood volume in which the activity is distributed. A simple way to normalize the measured uptake with respect to these two factors is using the so-called standardized uptake value (SUV), largely used in PET for characterizing tumor uptake. More precisely, SUV is usually defined by:

$$\mathrm{SUV} = \frac{\text{activity concentration (kBq/mL)}}{\text{injected dose (kBq)/patient weight (g)}}. \tag{4}$$

Strictly speaking, SUV should be expressed in g/ml. However, assuming that the patient density is 1 g/ml, SUV becomes dimensionless and can be considered as a relative quantitation index. In ❯ Eq. 4, activity concentration is corrected for radioactive decay, or alternatively, the decay affecting the injected dose at the scan time is considered. Assuming the tracer distributes uniformly throughout the patient, SUV should be 1 in each and every voxel within the body. Abnormalities in tracer uptake are thus detected by considering large deviations from 1. SUV is a convenient parameter to compare scans involving different injection protocols or different scans performed in patients. This is why it has found widespread acceptance in the clinics, although its relationship to the actual glucose metabolic rate when using the F18-Fluorodeoxyglucose tracer (FDG) (analogue of glucose widely used as a tracer in PET imaging, see ❯ Chap. 38, "PET Imaging: Basics and New Trends") is questionable (Visser et al. 2010). Although SUV is most often used in FDG PET, it can actually be used whatever may be the tracer, both in SPECT and PET as a normalized unit to facilitate image comparison between patients or patient scans. SUV calculation assumes that activity concentration in every voxel or in the region of interest is an unbiased estimate of the true activity concentration, which implies that all effects listed in ❯ *Table 1* should be carefully compensated for before SUV calculation and that the structure of interest is properly delineated from the image. The lack of common agreement regarding the best way to acquire and process PET data and then delineate tumors from PET images (Beyer et al. 2011; Buckler and Boellaard 2011) introduces significant variability in the resulting SUV estimate and subsequent interpretation.

4.2.3 Texture Measurements

Texture parameters can characterize local variations in signal intensity within a region of interest (e.g., El Naqa et al. 2009; Ganeshan et al. 2010, 2011; Tixier et al. 2011). They can thus describe a mixture of tissue types (in CT) or tissue functions (in SPECT or PET) that can by itself give relevant information about the structure under investigation. For instance, in PET, texture indices have been shown useful to determine whether a tumor will respond to treatment

Table 3

Examples of global texture parameters that can be derived from a single image. p(k) is the probability of occurrence of voxel value k derived from the histogram including a total of K bins

Texture parameter	Definition	Type of information
Entropy	$-\sum_{k=1,K} p(k)\log_2(p(k))$	Signal intensity and heterogeneity
Uniformity	$\sum_{k=1,K} p(k)^2$	Variation in signal intensity
Skewness	$\frac{\frac{1}{K}\sum_{k=1,K}\left(p(k)-\overline{p(k)}\right)^3}{\left(\frac{1}{K}\sum_{k=1,K}\left(p(k)-\overline{p(k)}\right)^2\right)^{3/2}}$	Asymmetry of the intensity values
Kurtosis	$\frac{\frac{1}{K}\sum_{k=1,K}\left(p(k)-\overline{p(k)}\right)^4}{\left(\frac{1}{K}\sum_{k=1,K}\left(p(k)-\overline{p(k)}\right)^2\right)^2}-3$	"Peakedness" of the intensity values

(Tixier et al. 2011). Global texture parameters characterize the histogram of voxel values in the region of interest in a more comprehensive way than just considering the mean voxel value. *Table 3* lists examples of global texture parameters used in medical images that can be derived from the histogram of voxel values, with the type of information provided by each parameter. More sophisticated texture parameters can be derived from texture matrices describing the relationship between neighboring voxel values as a function of their spatial arrangement (Haralick et al. 1973). An example of texture matrix is the co-occurrence matrix, describing how often pairs of voxels with specific values and in a specified spatial relationship occur in an image. Various parameters can be calculated from the co-occurrence matrix, including the entropy, correlation, homogeneity, contrast, and dissimilarity (Haralick et al. 1973). Other texture matrices (Haralick et al. 1973) can describe the alignments of voxels with the same intensity, or the local uniformity of signal, and specific texture parameters can be derived from them.

The use of texture analysis in SPECT, PET, and CT is still in its infancy, with promising results however (e.g., El Naqa et al. 2009; Ganeshan et al. 2010, 2011; Tixier et al. 2011). Care should be taken to distinguish between texture truly characterizing the tissue of interest and texture actually introduced by correlated noise affecting the reconstructed images. Some preprocessing, like rescaling and resampling, can thus be needed before applying texture analysis for increasing the robustness of the approach.

5 Quantitation from Dynamic Images

Static images can yield relative quantitation parameters (dimensionless), shape metrics or size, activity concentration, attenuation coefficients (HU), and indices characterizing the texture of the functional activity or tissue density. None of these precisely characterize a physiological or physiopathological process. To get further insight into the physiological process of interest, dynamic imaging is required, providing the time course of the tracer in the regions or voxels of interest. When image series are available from dynamic image acquisitions, more sophisticated quantitation procedures can be performed, including kinetic modeling and parametric imaging. All these procedures assume that the input image series consist of quantitative images, i.e., images in which the voxel value represents a well-understood physics parameter (most often

a tracer concentration). In short, kinetic modeling operates at a regional level, while parametric imaging operates at the voxel level. Only a short introduction is provided here, the reader is invited to refer to Chap. 42, "Compartmental Modeling in Emission Tomography" for a detailed presentation of these approaches.

5.1 Kinetic Modeling

The motivation for kinetic modeling is to extract some physiological meaningful parameters characterizing the fate of the tracer in tissues, including tracer exchange between tissues, based on the analysis of a dynamic image series (SPECT, PET, or CT). Indeed, when the tracer (contrast medium in CT) enters the body, it is usually transported through the circulation, then can travel across membranes (or not), undergo biochemical transformation (or not), and be eliminated. The idea of kinetic modeling is to characterize this "history" using parameters that can be physiologically or clinically relevant. A precise description of the fate of the tracer requires both spatial information (where is the tracer in the body?) and temporal information (when do changes occur?). Given that the tracer undergoes constant changes, a detailed description of the spatial and temporal changes is extremely challenging. As a result, the system under study (usually part of the body) is usually described using a very limited number of components that still describe the main physiological phenomena of interest.

Two types of kinetic modeling approaches can be used: stochastic (or non-compartmental) approaches and compartmental methods. In the stochastic approaches, only minimal assumptions about the underlying physiology of the tracer's uptake and metabolism are needed. There is no need for a detailed description of all the specific physiological regions (also called compartments) that the tracer may enter. An example of parameter that can be measured with a non-compartmental approach is the mean residence time (MRT). MRT is defined as the average time a tracer spends in a region of interest. It is given by

$$\mathrm{MRT} = \frac{\int_0^{+\infty} tA(t)\,\mathrm{d}t}{\int_0^{+\infty} A(t)\,\mathrm{d}t}, \tag{5}$$

where $A(t)$ is the signal intensity at time t in the region of interest. To derive MRT, only the time curve of signal within the compartment of interest is needed (time activity curve – TAC – in SPECT and PET for instance). This method does not rely on any model. Only a method for deriving an area under a sampled curve is needed, like using the trapezoid rule or fitting. Also, the investigator must define a method to extrapolate the measurements to time zero and time infinity. As a result, MRT will still depend on the way the investigator proceeds for these two steps (calculation of the area under a curve and extrapolation). This is true for any parameter derived from any modeling approach. It is thus important to always cast a critical eye to parameter estimates, bearing in mind that any measurement is prone to error or lack of reproducibility that should ideally be well characterized.

The second type of kinetic modeling approach is compartmental modeling (Chap. 42, "Compartmental Modeling in Emission Tomography"). This approach is the most commonly used to describe the uptake and clearance of tracers in tissues. These models assume that all molecules of injected tracer will, at any given time, exist in one of many compartments. A compartment defines both the physical location of the tracer (for instance, intravascular space, extracellular space) and its chemical state (typically bound or free). A compartment most often

actually describes a number of states lumped together for simplification. A compartmental model also describes the transportation of the tracer from a compartment to another, by means of "rate of changes" expressed in min^{-1}. The complete identification of a compartmental model thus describes at which rate the tracer moves from a compartment to another during the observation time. By fitting signal measurements performed in regions of interest corresponding to the compartments involved in the model, the rate constants can be determined (see Chap. 42, "Compartmental Modeling in Emission Tomography").

5.2 Parametric Imaging

When mentioning kinetic modeling, one usually implicitly means kinetic modeling performed at a region level. Kinetic modeling can actually also be performed at the voxel level, and is then called parametric imaging. The purpose of parametric imaging is to derive images of physiologically meaningful parameters. For instance, in F18-Fluorodeoxyglucose PET, quantitative images are usually expressed in Bq/ml, but parametric images of the glucose metabolic rate expressed in μmol/ml/min can also be obtained with a parametric imaging approach. The distinction between quantitative images and parametric images is actually tiny, but parametric imaging rather refers to images in which each voxel contains more than just a physics quantity, but a parameter with a straightforward physiological interpretation.

Parametric imaging can be based either on a model or not. For instance, it is conceivable to create a parametric image of the MRT, i.e., without relying on a particular model. When using a model, compartmental models similar to those used for kinetic modeling can be adapted (Chap. 42, "Compartmental Modeling in Emission Tomography"). As an example, the Patlak model presented in Chap. 42, "Compartmental Modeling in Emission Tomography" can be used to estimate the glucose metabolic rate, based on a two-tissue compartment model.

Parametric imaging can also be used using models other than compartmental models, such as Fourier models or linear decomposition models. In these approaches, the signal $A_i(t)$ detected in every voxel i is expressed as a function involving parameters of interest. For instance, in SPECT or PET cardiac imaging, cardiac acquisition can be gated based on an electrocardiograph measurement (Chap. 37, "SPECT Imaging: Basics and New Trends"). Each voxel i of the reconstructed gated image series is associated with a time activity curve $A_i(t)$. In voxels corresponding to the heart, $A_i(t)$ looks like a periodic function. The Fourier transform of $A_i(t)$ can be calculated, with

$$R_K(A_i) = \sum_{t=1}^{T} A_i(t) \cos(2\pi K t/T) \tag{6}$$

being the real value and

$$I_K(A_i) = \sum_{t=1}^{T} A_i(t) \sin(2\pi K t/T) \tag{7}$$

being the imaginary value.

In these expressions, T is the number of gates in the time series while K is the frequency ($K = 1$ for the fundamental frequency, $K = 2$ for the second harmonic, etc.). The amplitude of the transform is given by:

$$B_K(A_i) = \left[R_K(A_i)^2 + I_K(A_i)^2\right]^{1/2} \tag{8}$$

while the phase of the transform is given by:

$$P_K(A_i) = \arctan\left[I_K(A_i)/R_K(A_i)\right]. \tag{9}$$

The amplitude and phase images can then be represented to better assess the motion of the heart during the cardiac cycle. In particular, phase images have been shown to be useful to detect the left ventricle dyssynchrony (e.g., Boogers et al. 2009). Such Fourier-based parametric imaging is especially useful to interpret gated cardiac images.

Another model frequently used for analyzing data is a linear model, assuming that the signal curve $A_i(t)$ can be expressed as a linear combination of some curves, $f_q(t)$, often called factor, with weighting coefficients $C_i(t)$:

$$A_i(t) = \sum_{q=1}^{Q} C_q(i) f_q(t) + e_i(t). \tag{10}$$

In Eq. 10, Q represents the number of components in the linear decomposition, while $e_i(t)$ represents an error, including the noise present in the measured data and possibly a modeling error. In this model, the signal in each voxel is supposed to be described by the sum of functions f_q with a physiological meaning. For a given voxel, the weight $C_q(i)$ associated with function f_q gives the fraction of signal present in voxel i that is described by function f_q. The set of $C_q(i)$ coefficients associated with function f_q are represented as so-called factor images (there are as many $C_q(i)$ coefficients as voxels i in the analyzed image). Incidentally, this models accounts for partial volume. Indeed, if several types of tissues are present in voxel i, each with a distinct temporal behavior, then the weights $C_q(i)$ give the fraction of each tissue type in the voxel. The model described by Eq. 10 can be solved using a factor analysis approach (Frouin et al. 1993; Benali et al. 1994) or as an optimization problem (Sitek et al. 2002). The resulting factors, f_q, describe the kinetics of the tracer in the compartment described by the associated factor images C_q. Typical factors are shown in *Fig. 5* for a O^{15}-labeled water dynamic PET scan, where the three factors describe the time course of the tracer in the left ventricle, right ventricle, and myocardium. The associated factor images show the corresponding compartments (Frouin et al. 2001).

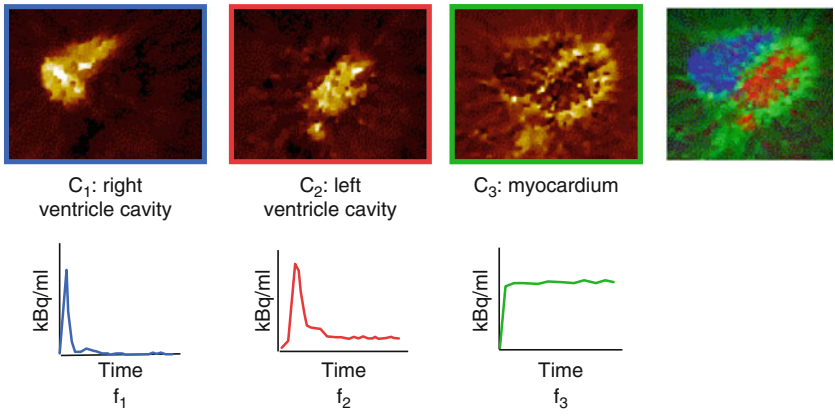


Fig. 5

Results of the factor analysis of a O^{15}-labeled water dynamic PET scan, where three compartments can be distinguished in the three factor images: right ventricle cavity, left ventricle cavity myocardium. The three-color superimposition of the three compartments in the image on the *right* shows the spatial consistency of the decomposition. The time course of O^{15}-labeled water in each compartment is given by the curves f_1, f_2, and f_3

The method has been shown to be useful for studying pharmacokinetics in CT (e.g., Chabriais et al. 1991), and has found many applications in SPECT and PET (e.g., Dowson et al. 2010; Klein et al. 2010; Rojas-Ordus et al. 2010), but also in ultrasound imaging (e.g., Frouin et al. 2002) and magnetic resonance imaging (Martel et al. 2003). Solving the linear model expressed as ❯ Eq. 10 using a factor analysis can also be viewed as a mean to perform image segmentation based on the temporal signal acquired in the different voxels, accounting or not for partial volume effect (forcing the $C_q(i)$ to be 0 for all q-values except one when solving the model makes it impossible to have a mixture of functions within the same voxels). In particular, it has been used to segment cardiac structures and derive the arterial input function in small animal PET (Kim et al. 2006).

6 Conclusion

Quantitative image analysis in tomography is a very broad area, involving a large number of signal and image processing methods, the choice of which should be guided by the medical application. Quantitative parameters can be derived either from static or dynamic imaging. In static imaging, quantitation makes it possible to determine the dimensions of abnormal structures (e.g., tumor volume), the magnitude of the abnormality (e.g., hypermetabolism or hypometabolism, hyperdensity or hypodensity), or the texture of the abnormality. In dynamic imaging, physiological parameters describing the function of organs or abnormalities can be derived, either at a regional level, or at a voxel level when performing parametric imaging. Accurate quantitation of clinically relevant parameters always assumes that the signal intensity in the image is at least proportional to a well-understood physics quantity (activity concentration in nuclear medicine, tissue attenuation coefficient in CT). Given that advances in PET, SPECT, and CT images now make it possible to acquire images that are quantitatively reliable, it is expected that their quantitative interpretation will become more and more widespread in a near future, mostly through the use of automatic or semiautomatic analysis tools. Many challenges like accurate segmentation and accounting for partial volume effect and noise still need to be coped with, but the effort is definitely worthwhile to make the most of increasingly accurate tomographic images to the patient benefit.

References

Abu Anas EM, Lee SY, Hasan MK (2010) Removal of ring artifacts in CT imaging through detection and correction of stripes in the sinogram. Phys Med Biol 55:6911–6930

Aldroubi A, Unser M (1996) Wavelets in medicine and biology. CRC Press, Boca Raton

Alessio A, Kinahan P, Lewellen T (2006) Modelling and incorporation of system response functions in 3-D whole body PET. IEEE Trans Med Imaging 25:828–837

Aristophanous M, Penney BC, Martel MK, Pelizzari CA (2007) A Gaussian mixture model for definition of lung tumor volumes in positron emission tomography. Med Phys 34:4223–4235

Beekman FJ, Kamphuis C, King MA, van Rijk PP, Viergever MA (2001) Improvement of image resolution and quantitative accuracy in clinical single photon emission computed tomography. Comput Med Imaging Graph 25:135–146

Belhassen S, Zaidi H (2010) A novel fuzzy C-means algorithm for unsupervised heterogeneous tumor quantification in PET. Med Phys 37:1309–1324

Benali H, Buvat I, Frouin F, Bazin JP, Di Paola R (1994) Foundations of factor analysis of medical image sequences: a unified approach and some practical implications. Image Vis. Comput. 12:375–385

Beucher S, Lantuéjoul C (1979) Use of watersheds in contour detection. In: International workshop on image processing, real-time edge and motion detection/estimation, Rennes, September 1979

Beyer T, Czernin J, Freudenberg LS (2011) Variations in Clinical PET/CT operations: results of an international survey of active PET/CT users. J Nucl Med 52:303–310

Birkfellner W (2010) Applied medical image processing: a basic course. Taylor & Francis, Boca Raton

Boas FE, Fleischmann D (2011) Evaluation of two iterative techniques for reducing metal artifacts in Computed Tomography. Radiology (in press)

Boogers M, Van Kriekinge SD, Henneman MM, Ypenburg C, Van Bommel RJ, Boersma E, Dibbets-Schneider P, Stokkel MP, Schalij MJ, Berman DS, Germano G, Bax JJ (2009) Quantitative gated SPECT–derived phase analysis on gated myocardial perfusion SPECT detects left ventricular dyssynchrony and predicts response to cardiac resynchronization therapy. J Nucl Med 50:718–725

Buckler AJ, Boellaard R (2011) Standardization of quantitative imaging: the time is right, and 18F-FDG PET/CT is a good place to start. J Nucl Med 52:171–172

Chabriais J, Lebo NK, Helenon O, Chourroute Y, Di Paola R, Moreau JF (1991) Iodinated contrast renal pharmacokinetic study by factor analysis dynamic computed tomography in the rabbit. Invest Radiol 26:S80–S82

Chandler A, Wei W, Herron DH, Anderson EF, Johnson VE, Ng CS (2011) Semiautomated motion correction of tumors in lung CT-perfusion studies. Acad Radiol 18:286–293

Dawson P (1999) Functional and physiological imaging. Textbook of contrast media. Informa Healthcare, Oxford, pp 75–94

Dowson N, Bourgeat P, Rose S, Daglish M, Smith J, Fay M, Coulthard A, Winter C, MacFarlane D, Thomas P, Crozier S, Salvado O (2010) Joint factor and kinetic analysis of dynamic FDOPA PET scans of brain cancer patients. Med Image Comput Comput Assist Interv 13:185–192

El Naqa I, Grigsby PW, Aptea A, Kidd E, Donnelly E, Khullar D, Chaudhari S, Yang D, Schmitt M, Laforest R, Thorstad WL, Deasy JO (2009) Exploring feature-based approaches in PET images for predicting cancer treatment outcomes. Pattern Recognit 42:1162–1171

Frouin F, Cinotti L, Benali H, Buvat I, Bazin JP, Millet P, Di Paola R (1993) Extraction of functional volumes from medical dynamic volumetric data sets. Comput Med Imaging Graph 17:397–404

Frouin F, Merlet P, Bouchareb Y, Frouin V, Dubois-Randé JL, De Cesare A, Herment A, Syrota A, Todd-Pokropek A (2001) Validation of myocardial perfusion reserve measurements using regularized factor images of H(2)(15)O dynamic PET scans. J Nucl Med 42: 1737–1746

Frouin F, Delouche A, Abergel E, Raffoul H, Diebold H, Diebold B (2002) Value of factor analysis in a wall motion study: preliminary application in the detection of left ventricular segmental contraction ischemic disorders by echocardiography. J Radiol 83:1835–1841

Ganeshan B, Abaleke S, Young RC, Chatwin CR, Miles KA (2010) Texture analysis of non-small cell lung cancer on unenhanced computed tomography: initial evidence for a relationship with tumour glucose metabolism and stage. Cancer Imaging 10:137–143

Ganeshan B, Burnand K, Young R, Chatwin C, Miles K (2011) Dynamic contrast-enhanced texture analysis of the liver: initial assessment in colorectal cancer. Invest Radiol 46:160–168

Geets X, Lee JA, Bol A, Lonneux M, Grégoire V (2007) A gradient-based method for segmenting FDG-PET images: methodology and validation. Eur J Nucl Med Mol Imaging 34:1427–1438

Green AJ, Francis RJ, Baig S, Begent RH (2008) Semiautomatic volume of interest drawing for (18)F-FDG image analysis-method and preliminary results. Eur J Nucl Med Mol Imaging 35:393–406

Haralick RM, Shanmugam K, Dinstein I (1973) Textural features for image classification. IEEE Trans Syst Man Cybern 3:610–621

Hatt M, Lamare F, Boussion N, Turzo A, Collet C, Salzenstein F, Roux C, Jarritt P, Carson K, Cheze-Le Rest C, Visvikis D (2007) Fuzzy hidden Markov chains segmentation for volume determination and quantitation in PET. Phys Med Biol 52:3467–3491

Hutton B, Buvat I, Beekman F (2011) Review and current status of SPECT scatter correction. Phys Med Biol (in press)

Kakar M Olsen DR (2009) Automatic segmentation and recognition of lungs and lesion from CT scans of thorax. Comput Med Imaging Graph 33:72–82

Kessler RM, Ellis JR, Eden M (1984) Analysis of emission tomographic scan data: limitations

imposed by resolution and background. J Comput Assist Tomogr 8:514–522

Kim J, Herrero P, Sharp T, Laforest R, Rowland DJ, Tai YC, Lewis JS, Welch MJ (2006) Minimally invasive method of determining blood input function from PET images in rodents. J Nucl Med 47:330–336

Kinahan PE, Hasegawa BH, Beyer T (2003) X-ray-based attenuation correction for positron emission tomography/computed tomography scanners. Semin Nucl Med 33:166–179

Klein R, Beanlands RS, Wassenaar RW, Thorn SL, Lamoureux M, DaSilva JN, Adler A, deKemp RA (2010) Kinetic model-based factor analysis of dynamic sequences for 82-rubidium cardiac positron emission tomography. Med Phys 37:3995–4010

Kolditz D, Meyer M, Kyriakou Y, Kalender WA (2011) Comparison of extended field-of-view reconstructions in C-arm flat-detector CT using patient size, shape or attenuation information. Phys Med Biol 56:39–56

Kyriakou Y, Meyer E, Prell D, Kachelriess M (2010) Empirical beam hardening correction (EBHC) for CT. Med Phys 37:5179–5187

Martel AL, Fraser D, Delay GS, Morgan PS, Moody AA (2003) Separating arterial and venous components from 3D dynamic contrast-enhanced MRI studies using factor analysis. Magn Reson Med 49:928–933

Montgomery DW, Amira A, Zaidi H (2007) Fully automated segmentation of oncological PET volumes using a combined multiscale and statistical model. Med Phys 34:722–736

Patton JA, Turkington TG (2008) SPECT/CT physical principles and attenuation correction. J Nucl Med Technol 36:1–10

Pohle R, Toennies KD (2001) Segmentation of medical images using adaptive region growing. Proc SPIE 4322:1337–1346

Ray S, Hagge R, Gillen M, Cerejo M, Shakeri S, Beckett L, Greasby T, Badawi RD (2008) Comparison of two-dimensional and three-dimensional iterative watershed segmentation methods in hepatic tumor volumetrics. Med Phys 35:5869–5881

Rojas-Ordus D, Jiménez-Angeles L, Hernández-Sandoval S, Valdes-Cristerna R (2010) Factor analysis of ventricular contraction using SPECT-ERNA images. Conf Proc IEEE Eng Med Biol Soc 5732–5735

Sadi F, Lee SY, Hasan MK (2010) Removal of ring artifacts in computed tomographic imaging using iterative center weighted median filter. Comput Biol Med 40:109–118

Shih FY (2009) Image processing and mathematical morphology: fundamentals and applications. CRC Press, Boca Raton

Sitek A, Gullberg GT, Huesman RH (2002) Correction for ambiguous solutions in factor analysis using a penalized least squares objective. IEEE Trans Med Imaging 21:216–225

Soret M, Koulibaly PM, Darcourt J, Buvat I (2006) Partial volume effect correction in SPECT for striatal uptake measurements in patients with neurodegenerative diseases: impact upon patient classification. Eur J Nucl Med Mol Imaging 33:1062–1072

Soret M, Bacharach SL, Buvat I (2007) Partial-volume effect in PET tumor imaging. J Nucl Med 48:932–945

Stute S, Benoit D, Martineau A, Rehfeld NS, Buvat I (2011) A method for accurate modelling of the crystal response function at a crystal sub-level applied to PET reconstruction. Phys Med Biol 56:793–809

Sureau FC, Reader AJ, Comtat C, Leroy C, Ribeiro MJ, Buvat I, Trébossen R (2008) Impact of image-space resolution modeling for studies with the high-resolution research tomograph. J Nucl Med 49:1000–1008

Tixier F, Le Rest CC, Hatt M, Albarghach N, Pradier O, Metges JP, Corcos L, Visvikis D (2011) Intratumor heterogeneity characterized by textural features on baseline 18F-FDG PET images predicts response to concomitant radiochemotherapy in esophageal cancer. J Nucl Med 52: 369–378

Truc PT, Kim TS, Lee S, Lee YK (2010) A study on the feasibility of active contours on automatic CT bone segmentation. J Digit Imaging 23:793–805

Visser EP, Boerman OC, Oyen WJ (2010) SUV: from silly useless value to smart uptake value. J Nucl Med 51:173–175

Zaidi H, El Naqa I (2010) PET-guided delineation of radiation therapy treatment volumes: a survey of image segmentation techniques. Eur J Nucl Med Mol Imaging 37:2165–2187

Further Reading

Dougherty G (2009) Digital image processing for medical applications. Cambridge University Press, Cambridge

Wolfgang B (2010) Applied medical image processing: a basic course. Taylor & Francis, Boca Raton

Zaidi H (ed) (2010) Quantitative analysis in nuclear medicine imaging. Springer

Examples of Free Software for Quantitative Imaging

Amide: http://amide.sourceforge.net/

Mipav: http://mipav.cit.nih.gov/

Osirix: http://www.osirix-viewer.com/

Pixies: http://imaging.apteryx.fr/pixies

42 Compartmental Modeling in Emission Tomography

Adriaan A. Lammertsma
VU University Medical Center, Amsterdam, The Netherlands

1 Introduction ***1066***

2 Compartment Models ***1066***
2.1 Single Tissue Compartment Model: Blood Flow 1068
2.2 Two Tissue Compartment Model: Receptor Studies 1069

3 Reference Tissue Models ***1073***

4 Weighting Factors ***1075***

5 Arterial Input Functions ***1076***

6 Comparison of Fits ***1078***

7 Parametric Methods ***1078***

8 Conclusions ***1080***

9 Cross-References ***1080***

References ***1080***

C. Grupen, I. Buvat (eds.), *Handbook of Particle Detection and Imaging*,
DOI 10.1007/978-3-642-13271-1_42, © Springer-Verlag Berlin Heidelberg 2012

Abstract: This chapter provides an overview of the basic principles of compartmental modeling as it is being applied to the quantitative analysis of positron emission tomography (PET) studies. Measurement of blood flow (perfusion) is used as an example of a single tissue compartment model and receptor studies are discussed in relation to a two tissue compartment model. Emphasis is placed on the accurate measurement of both arterial whole blood and metabolite-corrected plasma input functions. Reference tissue models are introduced as a noninvasive tool to investigate neuroreceptor studies. Finally, parametric methods are introduced in which calculations are performed at a voxel level.

1 Introduction

Both single photon emission tomography (SPECT) and positron emission tomography (PET) are tomographic imaging techniques that enable measurements of regional tissue radioactivity concentrations. Although other routes of administration are possible (e.g., inhalation), in general, accumulation in tissue follows intravenous injection of very small (tracer) amounts of molecules labeled with single photon or positron emitters, respectively. Clearly, uptake depends on how tissues handle these labeled molecules. As a result, both SPECT and PET images provide in vivo insight in tissue function (Phelps 2004). The specific function being imaged depends on the specific molecule that has been labeled. For example, tracers are available for imaging perfusion, metabolism, pre- and postsynaptic receptor density and affinity, neurotransmitter release, enzyme activity, and drug delivery and uptake.

Based on the principle of coincidence detection, in case of PET it is possible to perform accurate corrections for tissue attenuation. This is combined with unrivaled sensitivity, allowing for quantification at a picomolar level. In fact, PET represents the most selective and sensitive (up to picomolar range) method for measuring molecular pathways and interactions in vivo (Jones 1996). Therefore, this chapter will focus on tracer kinetic modeling in PET. Clearly, the same principles apply to SPECT and indeed to any imaging technique that makes use of externally administered tracers.

2 Compartment Models

PET only provides for accurate measurements of regional tissue concentrations of radioactivity. As schematically shown in ❯ *Fig. 1*, appropriate tracer kinetic models are needed to translate these measurements of tissue tracer concentrations into quantitative values of the tissue function under study. Such a tracer model is a mathematical description of the fate of the tracer in the human body, in particular the organ under study (i.e., within the field of view of the scanner). Although other types of models have been proposed, essentially all models used in practice are compartment models (Gunn et al. 2001). In these models, the possible distribution of a tracer is divided into a limited number of discrete compartments.

Examples of the most common compartment models are shown in ❯ *Fig. 2*. The simplest model is a zero tissue compartment model (❯ *Fig. 2a*). This would be an appropriate model for a blood volume tracer, i.e., a tracer that is not taken up by tissue and remains in the vascular space (Phelps et al. 1979). The next model is a single tissue compartment model (❯ *Fig. 2b*),

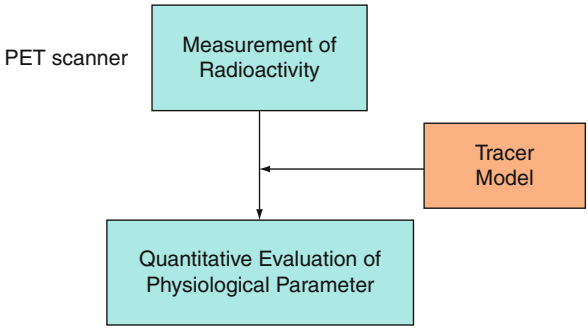


Fig. 1
Schematic diagram illustrating the need for a tracer kinetic model to translate PET measurements of radioactivity into a quantitative evaluation of a physiological parameter of interest

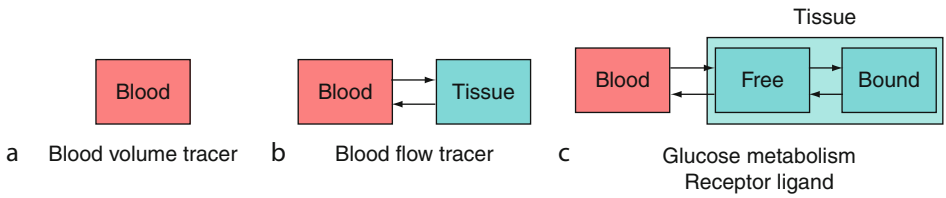


Fig. 2
Structure of the most common compartment models used in PET with (a) zero, (b) one, and (c) two tissue compartments

where the tracer is taken up by tissue, but has no further interactions within that tissue, i.e., no additional tissue compartments can be identified. A single tissue compartment model would be appropriate for a blood flow (perfusion) tracer (Frackowiak et al. 1980). A further compartment is introduced in the two tissue compartment model (*Fig. 2c*) with exchange between the two tissue compartments (Huang et al. 1980). This model can be used for both metabolic tracers and receptor ligands, where the first compartment represents free tracer in tissue and the second metabolized or bound tracer, respectively.

As illustrated by the two tissue compartment model, compartments do not represent physical compartments, but rather they constitute a convenient way of describing tracer kinetics. It should be noted that, when reading the literature, different nomenclature has been used for these models. For example, in older literature, the single tissue compartment model has been referred to as both a single (tissue) and a two (blood and tissue) compartment model. In recent years, there appears to be consensus to classify models according to the number of tissue compartments involved. Apart from the nomenclature for the models themselves, the parameters associated with these models have received many different names. A few years ago, however, consensus was reached about the nomenclature related to reversibly binding radioligands (Innis et al. 2007). As far as possible, in this chapter, this nomenclature will be followed even for models that do not relate to receptor ligands.

2.1 Single Tissue Compartment Model: Blood Flow

As mentioned above, the model describing kinetics of a perfusion tracer is a typical example of a single tissue compartment model and this application will be used as an example. The most important perfusion tracer is oxygen-15 labeled water, which in most organs is freely diffusible, at least in case of normal or reduced flow, conditions that are most frequently seen in pathology. In addition, water has no metabolic interactions in tissue, i.e., it is metabolically inert. This actually is the reason that a single tissue compartment suffices. A disadvantage is the short half-life of oxygen-15 (2 min), so that it needs to be (cyclotron) produced in close proximity to the scanning site. On the other hand, an advantage of this short half-life is the possibility to repeat measurements in a single scanning session, e.g., following an intervention. In addition, the short half-life permits acquisition of a $[^{15}O]H_2O$ scan immediately prior to a metabolic or receptor scan.

The basic differential equation of the single tissue compartment model, illustrated in *Fig. 3*, is given by

$$\frac{dC_T(t)}{dt} = K_1 \cdot C_A(t) - k_2 \cdot C_T(t), \tag{1}$$

where C_T and C_A are concentrations as function of time t for tissue and arterial blood, respectively, and K_1 and k_2 represent rate constants for influx into and efflux out of tissue, respectively. This equation simply reflects the fact that the rate of change of the tissue concentration is the difference between total influx and efflux. In general, the rate constant K_1 is given by

$$K_1 = E \cdot F, \tag{2}$$

where E represents first-pass extraction fraction and F blood flow (perfusion). Note that E itself depends on F (Renkin 1959; Crone 1963):

$$E = 1 - \exp\left(\frac{-PS}{F}\right), \tag{3}$$

where PS represents the permeability surface area product. This equation illustrates that E essentially is constant for lower flow values, but that it starts to decrease for higher flows, depending on the magnitude of PS.

Another important parameter is the volume of distribution (or partition coefficient) V_T, which reflects the equilibrium distribution of the tracer:

$$V_T = \frac{C_T}{C_A}, \tag{4}$$

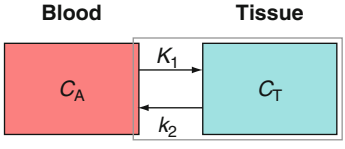


Fig. 3
Schematic diagram of the general single tissue compartment model. C_A and C_T represent arterial and tissue concentrations, K_1 and k_2 influx and efflux rate constants

where C_T and C_A now represent equilibrium (independent of time t) tissue and arterial blood concentrations, respectively. By imposing equilibrium conditions on Eq. 1, i.e., by setting $dC_T(t)/dt$ to 0, the following expression of V_T in terms of rate constants is obtained:

$$V_T = \frac{K_1}{k_2}. \tag{5}$$

For $[^{15}O]H_2O$ the extraction fraction is 100%, i.e., $E = 1$ and consequently $K_1 = F$. From Eq. 5 it follows that $k_2 = K_1/V_T$, which for water reduces to $k_2 = F/V_T$. In other words, the differential equation for $[^{15}O]H_2O$ studies becomes

$$\frac{dC_T(t)}{dt} = F \cdot C_A(t) - \left(\frac{F}{V_T}\right) \cdot C_T(t) \tag{6}$$

which has the following solution:

$$C_T(t) = F \cdot C_A(t) \otimes \exp\left\{-\left(\frac{F}{V_T}\right) \cdot t\right\} \tag{7}$$

where $\otimes$ represents a convolution.

Equation 7 illustrates a few basic principles of quantification. Firstly, the equation contains two entities that can be measured, the tissue concentration C_T and the arterial blood concentration C_A. C_T is measured using the PET scanner and, in general, C_A using arterial blood sampling. In other words, for full quantification of PET studies, arterial sampling is required (with a few exceptions – see later). This is logical, as the tracer is administered intravenously and subsequently delivered to the tissues through the arterial circulation. Note, that this implies that studies in which uptake is (semi)-quantified by normalizing to injected dose inherently assume that there is a constant relationship between injected dose and $C_A(t)$, i.e., that uptake and clearance of the tracer in the entire body (including all other organs than the one under study) is constant. The second principle illustrated by Eq. 7 is that C_T at a given time does not depend on C_A at that same time point, but rather on the history of C_A (represented by the convolution). Again, this is logical, as clearance from tissue is not instantaneous. Finally, it can also be seen from Eq. 7 that even for this simple single tissue compartment model there are two tissue-related parameters that are unknown, in this case F and V_T. This implies that it is impossible to quantify a study from a single PET scan, thereby illustrating the need for dynamic scanning.

In practice, a series of consecutive scans (so-called frames) is acquired, starting at the time of injection. This allows for measurement of C_T as a function of time. At the same time, using continuous or repetitive arterial blood sampling, C_A is measured as a function of time. Using nonlinear regression analysis, estimates of $C_T(t)$ for all frames are derived using $C_A(t)$ as input function and varying F and V_T in Eq. 7 until the optimal correspondence (i.e., lowest sum of squared differences) between estimated and measured $C_T(t)$ is obtained. F and V_T values corresponding to this optimal fit are then used as the parameter values characterizing the tissue under study.

2.2 Two Tissue Compartment Model: Receptor Studies

One of the most important PET applications is the study of various cerebral receptor systems using labeled ligands that bind to specific receptors. Again, such a ligand is introduced into the

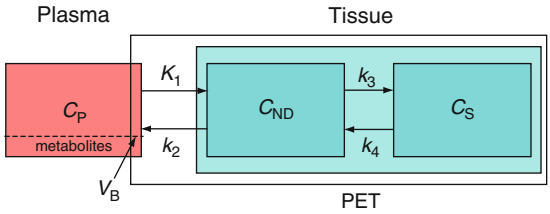


Fig. 4

Schematic diagram of the two tissue compartment model used for receptor studies. C_P, C_{ND}, and C_S represent arterial plasma, and non-displaceable and specific tissue concentrations. K_1 to k_4 represent rate constants characterizing transport between compartments and V_B blood volume within the PET region

human body by intravenous injection and is then transported to the brain through the arterial vasculature. If the specificity is high, i.e., if the ligand binds to only one receptor type, the cerebral signal will be the sum of free ligand, nonspecifically bound ligand (i.e., binding to various proteins), and ligand specifically bound to the receptor under study. This would require a three tissue compartment model, but in practice it is usually assumed that kinetics of the nonspecific compartment are very fast, so that free and nonspecific compartments can be combined into a single non-displaceable compartment. The corresponding two tissue compartment model is shown in ❯ *Fig. 4*. In this case, four rate constants are needed to describe kinetics of the tracer.

❯ *Figure 4* also illustrates some practical issues. Firstly, in this case input is not given by the arterial (whole blood) concentration C_A, but by the metabolite-corrected plasma concentration C_P, as, in contrast to water, only the tracer in plasma can exchange with tissue. A receptor ligand is a more complex molecule that circulates through the body and that will be taken up by the liver, where it can be metabolized. Subsequently, in most cases, these resulting radiolabeled metabolites will enter the circulation again. As a PET scanner only can measure radioactivity, quantification becomes extremely difficult, if not impossible, when these labeled metabolites would also be taken up by the brain. An important aspect of PET radiochemistry research is, therefore, to design ligands for which the labeled metabolites do not cross the blood–brain barrier. A second issue illustrated by ❯ *Fig. 4* is that a PET scanner measures radioactivity concentrations within a certain volume. This volume will not contain just tissue, but also blood. The concentration in this (fractional) blood volume V_B will not depend on C_P, but on C_A (not corrected for metabolites). In other words, for quantification both metabolite-corrected plasma and non-corrected whole blood concentrations are required.

The two tissue compartment model of ❯ *Fig. 4* leads to a measured concentration C_{PET} given by

$$C_{PET}(t) = (1 - V_B) \cdot C_T(t) + V_B \cdot C_A(t) \tag{8}$$

with

$$C_T(t) = C_{ND}(t) + C_S(t), \tag{9}$$

where C_{ND} and C_S represent concentrations in non-displaceable and specific compartments, respectively.

Tissue kinetics are described by the following differential equations:

$$\frac{dC_{ND}(t)}{dt} = K_1 C_P(t) - k_2 C_{ND}(t) - k_3 C_{ND}(t) + k_4 C_S(t), \tag{10}$$

$$\frac{dC_S(t)}{dt} = k_3 C_{ND}(t) - k_4 C_S(t). \tag{11}$$

As for the single tissue compartment model, these equations can be solved resulting in a nonlinear equation with two convolutions (Lammertsma et al. 1996), which can then be fitted using nonlinear regression analysis to obtain values of the four rate constants and V_B.

It should be noted that all these parameters need to be determined from a single time-activity curve, which can be noisy for small regions of interest or low uptake. In addition, there can be a high correlation between parameters, especially between k_2 and k_3. In many cases, tissue time–activity curves can be fitted, but with very low precision of fitted rate constants, sometimes called micro-parameters. Usually, in those cases it still is possible to obtain valuable information by looking at combinations of parameters, so-called macro-parameters. The most interesting macro-parameter for receptor studies is the non-displaceable binding potential BP_{ND} which, for tracer experiments, is given by

$$BP_{ND} = \frac{k_3}{k_4}. \tag{12}$$

BP_{ND} is of interest because of its direct relationship with parameters known in pharmacology. In pharmacological terms, the parameters k_3 and k_4 are given by (Innis et al. 2007)

$$k_3 = f_{ND} \cdot k_{on} \cdot B_{avail}, \tag{13}$$

$$k_4 = k_{off}, \tag{14}$$

where f_{ND} represents the free fraction of the non-displaceable compartment, B_{avail} the number of available receptors, and k_{on} and k_{off} the rate constants for association and dissociation of the ligand–receptor complex, respectively. Note, that if receptor occupancy by endogenous neurotransmitters is low, B_{avail} will approach the pharmacologically well-known parameter B_{max} (maximum number of receptors). Another important pharmacological parameter is the equilibrium dissociation constant K_d, given by (Innis et al. 2007)

$$K_d = \frac{k_{off}}{k_{on}}. \tag{15}$$

Combining ❯ Eqs. 12–15 results in the following relationship (Mintun et al. 1984):

$$BP_{ND} = f_{ND} \cdot \frac{B_{avail}}{K_d}. \tag{16}$$

Assuming that f_{ND} is constant, this equation demonstrates that BP_{ND} depends on the number of available receptors (B_{avail}) and the affinity of the ligand for the receptor (K_d). In clinical studies, it often is assumed that the affinity is relatively constant. In that case, BP_{ND} primarily depends on the number of available receptors. If occupancy of the receptor by endogenous ligands is negligible, it follows that BP_{ND} reflects receptor density. It is important, however, to verify that K_d is indeed constant in order to avoid erroneous conclusions. Note, that it is not possible to measure B_{avail} and K_d separately from a single tracer experiment, as they only enter the equations as a ratio. Separate measurements of B_{avail} and K_d are possible, but they involve multiple measurements in the same subject using tracer injections with different specific activities. In such a design, B_{avail} is modified by co-injection of a cold (nonradioactive) ligand.

In many cases, even BP_{ND} cannot be estimated reliably. This is especially the case when kinetics between both tissue compartments are relatively fast, making it difficult to distinguish these compartments from each other. In that case, analogues to the single tissue compartment model, the volume of distribution V_T, can be used. Again, this represents the equilibrium distribution of ligand in tissue relative to that in plasma:

$$V_T = \frac{C_T}{C_P}. \tag{17}$$

At equilibrium, both $dC_{ND}(t)/dt$ and $dC_S(t)/dt$ are 0. From Eq. 11 it follows that

$$C_S = \left(\frac{k_3}{k_4}\right) \cdot C_{ND}, \tag{18}$$

where C_S and C_{ND} are now independent of time. Substituting Eqs. 9 and 18 in Eq. 17 results in

$$V_T = \frac{C_T}{C_P} = \frac{C_{ND} + C_S}{C_P} = \left(1 + \frac{k_3}{k_4}\right) \cdot \frac{C_{ND}}{C_P}. \tag{19}$$

In addition, from

$$\frac{dC_{ND}(t)}{dt} + \frac{dC_S(t)}{dt} = K_1 C_P - k_2 C_{ND} = 0 \tag{20}$$

it follows that

$$C_{ND} = \left(\frac{K_1}{k_2}\right) \cdot C_P. \tag{21}$$

Finally, combining Eqs. 19 and 21 results in

$$V_T = \frac{K_1}{k_2} \cdot \left(1 + \frac{k_3}{k_4}\right) = \frac{K_1}{k_2} \cdot (1 + BP_{ND}). \tag{22}$$

It can be seen from Eq. 22 that V_T is related to the specific signal represented by BP_{ND}. It is also clear, however, that it is "contaminated" with a signal that is related to non-displaceable uptake, which usually is dominated by nonspecific binding.

Equation 22 describes V_T for a region with receptors that are targeted by the ligand. If there is another region in the brain that is devoid of these receptors, its volume of distribution can, analogues to Eq. 5 be described by

$$V_T' = \frac{K_1'}{k_2'}. \tag{23}$$

If the blood–brain barrier is symmetric (i.e., increased transport into the brain is matched by a similar rate of transport out of the brain) and the level of nonspecific binding is constant across the brain, the following equality holds:

$$\frac{K_1}{k_2} = \frac{K_1'}{k_2'}. \tag{24}$$

By combining Eqs. 22–24 it follows that (Lammertsma et al. 1996)

$$BP_{ND} = \frac{(V_T - K_1/k_2)}{(K_1/k_2)} = \frac{(V_T - V_T')}{V_T'} = \frac{V_T}{V_T'} - 1. \tag{25}$$

In other words, if a reference region devoid of receptors exists, it still is possible to derive BP_{ND} indirectly from V_T measurements, even if BP_{ND} cannot be fitted directly with sufficient precision.

3 Reference Tissue Models

As mentioned in the previous section, direct fitting of BP_{ND} often results in unstable estimates. Fortunately, BP_{ND} can still be obtained indirectly from V_T fits, provided a reference tissue devoid of receptors is available. This still requires a metabolite-corrected arterial plasma input function for both target and reference tissue V_T fits. Obtaining such an input function requires (invasive) arterial sampling together with labor-intensive estimation of plasma metabolites. If, however, a reference tissue exists it is also possible to derive BP_{ND} directly without arterial sampling. The differential equations describing the (full) reference tissue model illustrated in *Fig. 5* are given by

$$\frac{dC_{ND}(t)}{dt} = K_1 C_P(t) - k_2 C_{ND}(t) - k_3 C_{ND}(t) + k_4 C_S(t), \tag{26}$$

$$\frac{dC_S(t)}{dt} = k_3 C_{ND}(t) - k_4 C_S(t), \tag{27}$$

$$\frac{dC_R(t)}{dt} = K_1' C_P(t) - k_2' C_R(t), \tag{28}$$

where $C_R(t)$ represents the tissue concentration as function of time in the reference tissue.

Equation 28 can be used to express $C_P(t)$ in terms of $C_R(t)$, which can then be substituted in Eq. 26. Using the same assumption as for the indirect plasma estimation of BP_{ND}, expressed by Eq. 24, $C_T(t)$ can be fitted for $R_1 \left(= K_1/K_1'\right)$, k_2, k_3, and BP_{ND} using $C_R(t)$ as (indirect) reference tissue input function (Lammertsma et al. 1996). Note, that K_1 cannot be estimated but only the ratio R_1, as curves are expressed relative to the reference tissue. K_1 and K_1', however, do not have to be the same, i.e., this method does not assume that delivery to target and reference tissues is the same.

Looking at *Fig. 5* it may be tempting to subtract $C_R(t)$ from $C_T(t)$ to obtain $C_S(t)$ and to subsequently fit $C_S(t)$ with $C_R(t)$ as input. This would have the advantage that only two parameters (k_3 and k_4) need to be fitted. Unfortunately, this approach is not valid, as the time courses of $C_{ND}(t)$ and $C_R(t)$ are not the same, despite the fact that C_{ND} and C_R would be the

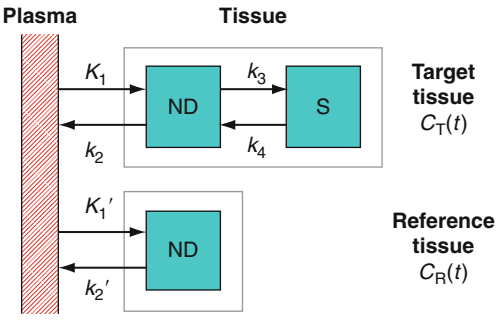


Fig. 5
Schematic diagram of the full reference tissue model. ND and S represent non-displaceable and specific compartments. In the operational equation, the target tissue concentration C_T is expressed in terms of the reference tissue concentration C_R

same at equilibrium (same level of nonspecific binding), simply because the environment of both compartments is different. In the target tissue, early after injection, the ligand will be transferred from the non-displaceable to the specific compartment, which does not happen in the reference tissue.

The full reference tissue model given above provides a full description of all kinetic parameters, but also has some limitations. The main problem is that parameters may be correlated and, in practice, often reasonable fits can be obtained with either high k_3, low k_2 or low k_3, high k_2. Especially in the presence of noise, it may be difficult to predict which solution would provide the lowest residual sum of squares. In line with this, the operational (fitting) equation appears to be sensitive to local minima, requiring repetitive fits with different parameter starting values in order to find the best overall solution.

A simple way of increasing stability of fits is to reduce the number of parameters that need to be estimated. The full reference tissue model contains four parameters. This can be reduced to three by making the assumption that the exchange between non-displaceable and specific compartments is relatively fast. The resulting simplified reference tissue model (SRTM) is shown in ❯ *Fig. 6* and kinetics can be described by the following differential equations:

$$\frac{\mathrm{d}C_\mathrm{T}(t)}{\mathrm{d}t} = K_1 C_\mathrm{P}(t) - \frac{k_2 C_\mathrm{T}(t)}{(1+\mathrm{BP_{ND}})}, \tag{29}$$

$$\frac{\mathrm{d}C_\mathrm{R}(t)}{\mathrm{d}t} = K_1' C_\mathrm{P}(t) - k_2' C_\mathrm{R}(t). \tag{30}$$

It can be seen that the complexity of these equations is significantly reduced compared with ❯ Eqs. 26–28. Note that the target tissue is now treated as a single tissue compartment with an apparent rate of clearance that is reduced by a factor $1 + \mathrm{BP_{ND}}$ compared with the situation that no receptors would have been present.

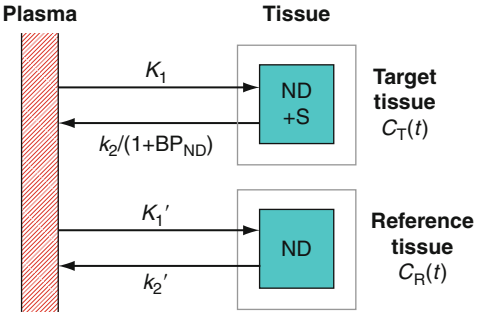


Fig. 6

Schematic diagram of the simplified reference tissue model (SRTM). The non-displaceable binding potential $\mathrm{BP_{ND}}$ results in a reduced apparent k_2 for the target tissue, where non-displaceable (ND) and specific (S) compartments are lumped together. In the operational equation, the target tissue concentration C_T is expressed in terms of the reference tissue concentration C_R

Similarly as for the full reference tissue model, Eq. 30 can be substituted into Eq. 29 and, again using the assumption of Eq. 24, $C_T(t)$ can be expressed as function of $C_R(t)$ (Lammertsma and Hume 1996):

$$C_T(t) = R_1 \cdot C_R(t) + \left\{ k_2 - \frac{R_1 \cdot k_2}{(1 + BP_{ND})} \right\} \cdot C_R(t) \otimes \exp\left\{ \frac{-k_2 \cdot t}{(1 + BP_{ND})} \right\}. \quad (31)$$

It can be seen that in this case only three parameters need to be fitted, R_1 $\left(= K_1/K_1'\right)$, k_2, and BP_{ND}, leading to much better stability.

It is of interest to note that although SRTM assumes fast exchange between non-displaceable and specific compartments, this assumption does not seem to be too critical. For example, originally SRTM was developed and validated for analyzing [^{11}C]raclopride kinetics in striatum. Using plasma input models, it is known that these kinetics are best described by two tissue compartments. Yet BP_{ND} values obtained with SRTM were identical to those obtained with the full reference tissue model (Lammertsma and Hume 1996). Nevertheless, for new applications (tracers), the validity of SRTM should always be checked by comparing with the full reference tissue model. In addition, as plasma input models are considered to be the gold standard, results of any reference tissue model should be compared with those of the most appropriate plasma input model.

4 Weighting Factors

In the sections above, both plasma input and reference tissue input compartment models have been described. In both cases, the general principle of using these models to estimate relevant tissue parameters is the same. It requires both an input (plasma or reference tissue) and a tissue response time–activity curve. Using standard nonlinear regression techniques and the measured input function, the tissue curve is then fitted to the operational equation corresponding to the model that best describes tracer kinetics.

In general, it can be assumed that statistical uncertainty of the input function is small (external sampling or derived from larger structures within the scan, e.g., cerebellum). For tissue curves, however, noise levels can be much higher, especially if small volumes of interest (VOI) are used. Clearly, noise levels will change over the period of a dynamic scan due to uptake and clearance of the tracer, and physical decay of the radionuclide. In addition, just after injection, changes in tissue concentrations are much faster than at later times. It is therefore customary to start a dynamic scan with short frames (e.g., as short as 5 s) and progressively increase the duration of the frames (possibly to 10 min at the end of the scan). As time–activity curves are expressed as concentrations, it will be clear that the statistical uncertainty of a given measured concentration also will depend on frame duration. Therefore, it is not appropriate to fit the data as if all data points had the same precision and, consequently, some form of weighting is required.

Although different weighting schemes have been proposed, only one will be presented here. In this scheme, it is assumed that relative noise levels do not depend on actual counts within a VOI, but are proportional to the total counts acquired in a frame. Note that only relative noise from frame to frame is required, as this relates to the precision of successive data points.

Suppose T total counts (not corrected for dead-time and decay) are acquired in a frame of duration L. Then the total count rate R for that frame equals T/L. For the total counts a Poisson

distribution can be assumed, leading to

$$\mathrm{Var}\,(T) = T, \tag{32}$$

$$\mathrm{SD}\,(T) = \sqrt{T}, \tag{33}$$

$$\mathrm{COV}\,(T) = \frac{\sqrt{T}}{T} = \frac{1}{\sqrt{T}}, \tag{34}$$

where Var represents variance, SD standard deviation, and COV coefficient of variation.

As there will be no error in the frame duration L, it follows that

$$\mathrm{COV}\,(R) = \frac{1}{\sqrt{T}}, \tag{35}$$

$$\mathrm{SD}\,(R) = \frac{R}{\sqrt{T}}, \tag{36}$$

$$\mathrm{Var}\,(R) = \frac{R^2}{T}, \tag{37}$$

$$\mathrm{Weight}\,(R) = \frac{T}{R^2} = \frac{L}{R} = \frac{L^2}{T}. \tag{38}$$

These weights are valid for non-decay-corrected data. In practice, it is customary to correct time–activity curves for decay. In that case, R is not equal to T/L, but needs to be modified to

$$R = f \cdot \frac{T}{L}, \tag{39}$$

where f is the decay correction factor for the frame. This is given by

$$f = \lambda \cdot \frac{T}{\{\exp\,(-\lambda T_s) - \exp\,(-\lambda T_e)\}}, \tag{40}$$

where T_s and T_e are start and end times of the frame, and λ is the decay constant of the radionuclide used. The corresponding weighting factors for decay-corrected data are then given by

$$\mathrm{Weight}\,(R) = \frac{L^2}{(f^2 \cdot T)}. \tag{41}$$

5 Arterial Input Functions

As mentioned earlier, in those cases where a reference tissue model is not available or valid, data can only be quantified if an arterial blood input function is available. For $[^{15}\mathrm{O}]\mathrm{H_2O}$ perfusion measurements only whole blood concentrations are needed. For most other applications, both metabolite-corrected plasma and non-corrected whole blood data are required. In general, a tracer is administered using a bolus intravenous injection. This implies that arterial concentrations are changing continuously during a scan, with rapid changes early after injection. In order to characterize such an arterial curve, multiple arterial blood samples are required and these samples have to be collected at high frequency early during a study. This, of course, is especially true for a $[^{15}\mathrm{O}]\mathrm{H_2O}$ study, where most perfusion information is present at the beginning of the study. High temporal resolution is best achieved by using an online automatic blood sampler (Boellaard et al. 2001). This is a system where blood is withdrawn continuously from

the radial artery using, e.g., an infusion pump. The line is led through or over a detector and the count rate is monitored continuously, e.g., in 1 s bins.

An online blood sampler saves a lot of time and manpower. In addition, it does not suffer from potential timing errors that can easily occur with very fast manual sampling schemes. It should be realized, however, that there will be a finite delay between withdrawal from the radial artery and counting by the detector. For a fixed speed this delay can be measured. However, during the delay, i.e., the transit time through the tubing, there will also be dispersion. Again, this can be measured beforehand, but it should be measured using active and cold blood, as dispersion for water will be different. In addition, although delay and dispersion in the tubing are important, relevant delay and dispersion may actually be higher. After intravenous administration, radioactivity will enter the arterial circulation through the heart. From the heart it takes a finite time to reach the organ of interest, e.g., the brain. It also takes a finite, but longer, time to reach the arterial cannula. So, the relevant delay (and associated dispersion) is given by delay in the tubing plus delay from heart to the cannula site minus delay from heart to organ. Therefore, rather than correcting curves using previously measured values, a better approach is to incorporate delay and dispersion in the fitting equation (Lammertsma et al. 1989). For a perfusion study, this will double (from two to four) the number of parameters to be fitted, which would result in reduced precision of parameter estimates. For the brain, it is therefore customary to first generate a whole brain curve (very good statistics) and to fit this curve for all parameters. Subsequently, delay and dispersion are fixed for smaller regions of interest. This approach is valid, as the difference in arrival time between different brain structures is negligible compared with the fitted delay. Similar approaches can be used for other organs. Note that although it is possible to correct for delay and dispersion, every effort should be made to limit both effects as much as possible, as large corrections will lead to large uncertainties. In addition, for optimal settings dispersion can usually be described by a single exponential. If no care is taken, this may no longer be valid. To limit dispersion, a single tube (i.e., as few connectors as possible) should be used that is as short as possible and has a small internal diameter (e.g., 1 mm).

Because of the rapid exchange between plasma and red cells, for $[^{15}O]H_2O$ perfusion studies only the arterial whole blood concentration is required. For receptor studies, however, both metabolite-corrected plasma and non-corrected whole blood concentrations are needed. A receptor ligand is dissolved in plasma and therefore the plasma concentration dictates the exchange with tissue. Unfortunately, almost all ligands are metabolized resulting in the ingrowth of radioactive metabolites. In general, at least for brain studies, only those tracers are accepted for which the radioactive metabolites do not cross the blood–brain barrier. That also implies that the arterial plasma curve needs to be corrected for radioactive metabolites, as only the parent compound constitutes the input function. In practice, deriving such a curve can be done in the following way. During scanning the whole blood curve is measured using an online automatic sampler. At set times (e.g., seven times during a 1 h study), online sampling is interrupted to withdraw manual samples. It is recommended to flush the line in front of the detector with heparinized saline. This clearly marks the timing of the sample in the online curve (of course these flushing periods should be removed from the online curve), but most importantly, it also provides a check on the stickiness of the tracer. If counts during a flushing period do not return to zero, the tracer sticks to the tubing and a correction for this stickiness should be considered (Lammertsma et al. 1991) or, better, other tubing should be used. Next, plasma and whole blood concentration of the manual samples should be measured after which a mathematical function (usually an exponential) can be fitted to the plasma to whole blood ratios over time. The whole blood curve should then be multiplied with this mathematical function to generate a (total) plasma curve as function of time. The same discrete arterial (plasma) samples

should also be used to measure the parent fraction, i.e., the fraction of total radioactivity in the plasma that is due to the originally injected tracer. This fraction over time should also be fitted to a mathematical function (e.g., a Hill function (Gunn et al. 1998)). Finally, the total plasma curve derived above should be multiplied with this mathematical function, resulting in the final metabolite-corrected plasma input function.

6 Comparison of Fits

Ideally, all receptor ligand studies should be analyzed using the two tissue compartment model described above. Unfortunately, as indicated previously, this is not always possible. Due to the specific kinetics of a ligand or higher noise levels (e.g., low uptake), it is possible that the second compartment is not identifiable. In developing a tracer kinetic model for a new radioligand it is necessary to compare the performance of several models with different numbers of parameters. The goal is then to identify the model that describes the observed kinetics adequately, but that is not too complex for reliable estimation of the parameter of interest. Note that this can be a macro-parameter (e.g., BP_{ND}). Given the research question, it often is not necessary to have very high precision of individual rate constants. Fits should, however, not be compared on the basis of residual sum of squares only, as this would favor models with more parameters. In practice, three different tests are commonly used for (statistical) comparison of various fits, the Akaike and Schwarz criteria, and the F-test.

The Akaike information criterion (AIC) is given by (Akaike 1974)

$$\mathrm{AIC} = N \cdot \ln(\mathrm{SS}) + 2P, \tag{42}$$

where N is the number frames, P the number of parameters, and SS the residual sum of squares. The fit with the lowest AIC is considered to be the "best" fit. It can be seen from Eq. 42 that additional parameters are penalized. This is also the case in the Schwarz criterion, which is given by (Schwarz 1978)

$$\mathrm{SC} = N \cdot \ln(\mathrm{SS}) + P \ln(N). \tag{43}$$

Again the fit with the lowest SC is considered to be the best. Finally, the F-test is given by (Cunningham 1985)

$$F = \frac{\{(\mathrm{SS}_1 - \mathrm{SS}_2)/(P_2 - P_1)\}}{\{\mathrm{SS}_2/(N - P_2)\}}, \tag{44}$$

where 1 and 2 stand for the fits with the lowest and highest number of parameters, respectively. An F-statistic table is required to assess significance.

It is difficult to indicate which is the better test. In practice, all three tests are often used to check whether they provide consistent results. A major advantage of the Akaike and Schwarz tests is the fact that they can easily be implemented in an (automatic) program (e.g., fitting of many regional curves for several patients). This is more difficult for the F-test.

7 Parametric Methods

As indicated earlier, the model equations describing kinetics of compartment models are nonlinear. Tissue curves are fitted using nonlinear regression algorithms. Usually, this is performed on a region of interest level, where regions are defined prior to kinetic analysis using a summed

tracer accumulation image or a co-registered MRI scan. There is, of course, the risk that an abnormality exists in only part of one or more regions. In that case it is not unlikely that the given abnormality is overlooked, as fitted parameters for the whole region will be "diluted" with normal kinetics for the rest of the region and therefore changes may not reach statistical significance.

To fully exploit the spatial resolution of a scanner, fitting at a voxel level is required. This is not practical when using nonlinear regression, as the method simply is too slow to fit a few million voxels. In addition, nonlinear regression suffers from noise amplification and the high noise levels associated with individual voxels would prohibit reliable results. To avoid noise amplification some form of linearization is required and indeed several linear methods have been proposed.

The best known linearized method is the Patlak plot (Patlak et al. 1983), which can be used to derive the net rate of influx K_i for irreversible tracers, notably FDG. When k_4 equals 0, it can be shown that, after an equilibration period, the plot of C_T/C_P against $\int C_P\,dt/C_P$ becomes a straight line:

$$\frac{C_T(t)}{C_P(t)} = K_i \cdot \frac{\int_0^t C_P(\tau)\,d\tau}{C_P(t)} + V_i, \tag{45}$$

where V_i is the initial volume of distribution (of free tracer) and K_i is given by

$$K_i = K_1 \cdot \frac{k_3}{(k_2 + k_3)}. \tag{46}$$

Another well-known linearization is the Logan plot for reversible tracers (Logan et al. 1990). In this case a linear transformation is used, such that after an equilibration time a straight line is obtained with a slope that represents V_T. Later the method was adapted to allow for a reference tissue input (Logan et al. 1996). A third method that has been used regularly involves linearization of the equations describing reference tissue models (Ichise et al. 2002, 2003).

In principle, the linear methods given above derive a macro-parameter of interest from the time–activity curves using a linear transformation of the data. Another, more general, approach is the basis function method, which in theory can be used for all compartment models. The method is best known for its implementation of SRTM, which is also known as RPM (Gunn et al. 1997). As can be seen from ❯ Eq. 31, the equation for the target tissue concentration contains both a linear and a convolution term. The basic idea of the method is to precalculate the convolution for a series of basis functions that cover the whole range of physiological values. For a single (precalculated) convolution, ❯ Eq. 31 becomes linear and the coefficients can be obtained by linear regression. This process is repeated for all basis functions and, finally, the linear fit with the lowest residual sum of squares is kept. From the coefficients of that specific linear fit, the actual model parameters can be calculated.

Using SRTM at a voxel level (i.e., RPM) provides the opportunity to reduce the number of parameters even further (Wu and Carson 2002), as R_1 and k_2 are obtained for all voxels. Given R_1 and k_2, k_2' can be calculated using ❯ Eq. 24. However, k_2' should be a constant as it refers to k_2 of the reference region. This can be utilized by running RPM twice. After the first run, k_2' is calculated and in the second run k_2' is fixed to the median value (not mean, as this would be sensitive to outliers) from the first run, effectively reducing the number of parameters to two.

Clearly, parametric methods need to be validated against full compartmental analysis. In fact, for a new tracer, it is advisable to test all parametric methods by comparing results with nonlinear regression of the compartment model (at a region of interest level). In addition, simulation studies should be performed to test sensitivity to both noise and deviations

from underlying assumptions. Experience with the parametric methods mentioned above has shown that no single method is ideal for all applications, i.e., for each new ligand the optimal method needs to be determined again.

8 Conclusions

Quantitative analysis of PET data requires dynamic scanning and measurements of both arterial whole blood and metabolite-corrected plasma input functions. Blood data should be collected with sufficient temporal resolution to "catch" the peak after a bolus injection. For a new tracer, several plasma input models should be tested to assess which model is optimal for that particular tracer. In case of neuroreceptor studies, reference tissue models should also be investigated, provided a region devoid of these receptors can be identified in the brain. Finally, to fully utilize the spatial resolution of the scanner, parametric methods should be investigated. It should be realized, however, that all parametric methods contain some form of linearization and, ideally, abnormalities seen in parametric images should be confirmed by full compartmental analysis.

9 Cross-References

- Chapter 4, "Data Analysis"
- Chapter 35, "Radiation-Based Medical Imaging Techniques: An Overview"
- Chapter 37, "SPECT Imaging: Basics and New Trends"
- Chapter 38, "PET Imaging: Basics and New Trends"
- Chapter 40, "Motion Compensation in Emission Tomography"
- Chapter 41, "Quantitative Image Analysis in Tomography"
- Chapter 45, "High-Resolution and Animal Imaging Instrumentation and Techniques"

References

Akaike H (1974) A new look at the statistical model identification. IEEE Trans Automat Contr 19:716–723

Boellaard R, Van Lingen A, Van Balen SCM, Hoving BG, Lammertsma AA (2001) Characteristics of a new fully programmable blood sampling device for monitoring blood radioactivity during PET. Eur J Nucl Med 28:81–89

Crone C (1963) The permeability of capillaries in various organs as determined by use of the 'indicator diffusion' method. Acta Physiol Scand 58:292–305

Cunningham VJ (1985) Non-linear regression techniques in data analysis. Med Inform 10: 137–142

Frackowiak RSJ, Lenzi GL, Jones T, Heather JD (1980) Quantitative measurement of regional cerebral blood flow and oxygen metabolism in man, using ^{15}O and positron emission tomography: theory, procedure, and normal values. J Comp Assist Tomogr 4:727–736

Gunn RN, Lammertsma AA, Hume SP, Cunningham VJ (1997) Parametric imaging of ligand-receptor binding in PET using a simplified reference region model. Neuroimage 6:279–287

Gunn RN, Sargent PA, Bench CJ, Rabiner EA, Osman S, Pike VW, Hume SP, Grasby PM, Lammertsma AA (1998) Tracer kinetic modelling of the 5-HT_{1A} receptor ligand [*carbonyl*-^{11}C]WAY-100635 for PET. Neuroimage 8:426–440

Gunn RN, Gunn SR, Cunningham VJ (2001) Positron emission tomography compartmental models. J Cereb Blood Flow Metab 21:635–652

Huang SC, Phelps ME, Hoffman EJ, Sideris K, Selin CJ, Kuhl DE (1980) Noninvasive determination of local cerebral metabolic rate of glucose in man. Am J Physiol 238:E69–E82

Ichise M, Toyama H, Innis RB, Carson RE (2002) Strategies to improve neuroreceptor parameter estimation by linear regression analysis. J Cereb Blood Flow Metab 22:1271–1281

Ichise M, Liow JS, Lu JQ, Takano A, Model K, Toyama H, Suhara T, Suzuki K, Innis RB, Carson RE (2003) Linearized reference tissue parametric imaging methods: application to [11C]DASB positron emission tomography studies of the serotonin transporter in human brain. J Cereb Blood Flow Metab 23:1096–1112

Innis RB, Cunningham VJ, Delforge J, Fujita M, Gjedde A, Gunn RN, Holden J, Houle S, Huang SC, Ichise M, Iida H, Ito H, Kimura Y, Koeppe RA, Knudsen GM, Knuuti J, Lammertsma AA, Laruelle M, Logan J, Maguire RP, Mintun MA, Morris ED, Parsey R, Price JC, Slifstein M, Sossi V, Suhara T, Votaw JR, Wong DF, Carson RE (2007) Consensus nomenclature for in vivo imaging of reversibly binding radioligands. J Cereb Blood Flow Metab 27:1533–1539

Jones T (1996) The role of positron emission tomography within the spectrum of medical imaging. Eur J Nucl Med 23:207–211

Lammertsma AA, Hume SP (1996) Simplified reference tissue model for PET receptor studies. Neuroimage 4:153–158

Lammertsma AA, Frackowiak RSJ, Hoffman JM, Huang SC, Weinberg IN, Dahlbom M, MacDonald NS, Hoffman EJ, Mazziotta JC, Heather JD, Forse GR, Phelps ME, Jones T (1989) The $C^{15}O_2$ build-up technique to measure regional cerebral blood flow and volume of distribution of water. J Cereb Blood Flow Metab 9:461–4707

Lammertsma AA, Bench CJ, Price GW, Cremer JE, Luthra SK, Turton D, Wood ND, Frackowiak RSJ (1991) Measurement of cerebral monoamine oxidase B activity using L-[^{11}C]deprenyl and dynamic positron emission tomography. J Cereb Blood Flow Metab 11:545–556

Lammertsma AA, Bench CJ, Hume SP, Osman S, Gunn K, Brooks DJ, Frackowiak RSJ (1996) Comparison of methods for analysis of clinical [^{11}C]raclopride studies. J Cereb Blood Flow Metab 16:42–52

Logan J, Fowler JS, Volkow ND, Wolf AP, Dewey SL, Schlyer DJ, Macgregor RR, Hitzmann R, Bendriem B, Gatley SJ, Christman DR (1990) Graphical analysis of reversible radio-ligand binding from time-activity measurements applied to [*N*-^{11}C-methyl]-(-)-cocaine PET studies in human subjects. J Cereb Blood Flow Metab 10:740–747

Logan J, Fowler JS, Volkow ND, Wang GJ, Ding YS, Alexoff DL (1996) Distribution volume ratios without blood sampling from graphical analysis of PET data. J Cereb Blood Flow Metab 16:834–840

Mintun MA, Raichle ME, Kilbourn MR, Wooten GF, Welch MJ (1984) A quantitative model for the in vivo assessment of drug binding sites with positron emission tomography. Ann Neurol 15:217–227

Patlak CS, Blasberg RG, Fenstermacher JD (1983) Graphical evaluation of blood-to-brain transfer constants from multiple-time uptake data. J Cereb Blood Flow Metab 3:1–7

Phelps ML (2004) PET: molecular imaging and its biological applications. Springer, New York

Phelps ME, Huang SC, Hoffman EJ, Kuhl DE (1979) Validation of tomographic measurement of cerebral blood volume with C-11 labeled carboxyhaemoglobin. J Nucl Med 20: 328–334

Renkin EM (1959) Transport of potassium-42 from blood to tissue in isolated mammalian skeletal muscles. Am J Physiol 197:1205–1210

Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6:461–464

Wu Y, Carson RE (2002) Noise reduction in the simplified reference tissue model for neuroreceptor functional imaging. J Cereb Blood Flow Metab 22:1440–1452

43 Evaluation and Image Quality in Radiation-Based Medical Imaging

Matthew A. Kupinski
University of Arizona, Tucson, AZ, USA

1 Introduction . . . 1084
1.1 The Imaging Equation . . . 1084

2 Tasks . . . 1085
2.1 Classification Tasks . . . 1085
2.2 Estimation Tasks . . . 1086
2.3 Combined Tasks . . . 1086
2.4 Distributions . . . 1087

3 Observers . . . 1087
3.1 Figures of Merit . . . 1088

4 Ideal Observers . . . 1089

5 Ideal Linear Observers . . . 1090
5.1 Linear Estimation Tasks . . . 1091

6 Notes on Linear Observers . . . 1092

7 Combination Task Observer Models . . . 1092

8 Discussion . . . 1093

References . . . 1093

C. Grupen, I. Buvat (eds.), *Handbook of Particle Detection and Imaging*,
DOI 10.1007/978-3-642-13271-1_43, © Springer-Verlag Berlin Heidelberg 2012

Abstract: In this chapter, we present methods for assessing image quality in radiation-based medical imaging. Image quality is defined by the ability of an observer to perform a relevant task using the images or data generated by an imaging system. The techniques presented in this chapter may be used to assess the utility of imaging hardware or to fine-tune the processing of image data into reconstructed images. Special attention is paid to nuclear-medicine imaging, but X-ray-based imaging systems are also considered.

1 Introduction

In medical imaging, images are acquired for specific purposes. These purposes include the detection of abnormalities such as tumors or heart defects and the estimation of quantities of interest such as cardiac ejection fraction. Because images are acquired for specific purposes, it is reasonable to assess the performance of imaging systems and image-processing algorithms based on how well these tasks can be performed. The concept of task-based assessment of image quality addresses this very topic. Traditional measures such as resolution and contrast are of secondary importance because they do not necessarily correlate with task performance. Instead, we measure the quality of an imaging system or the processing of image data based on how well relevant observers can perform medically relevant tasks on a population of patients that is of interest. In this chapter, we will review the tasks that arise in medical imaging, the various observer models, and practical methods for their implementation.

1.1 The Imaging Equation

The relationship between objects being imaged and images is of fundamental importance in assessing image quality. Imaging can be represented mathematically as

$$\boldsymbol{g} = \mathcal{H} f(\boldsymbol{r}, t) + \boldsymbol{n}. \tag{1}$$

Here, $f(\boldsymbol{r}, t)$ represents the object being imaged as a function of a spatial coordinate $\boldsymbol{r}$ and time t. In nuclear-medicine imaging, this function represents the distribution of the radiotracter throughout the body and how this distribution changes with time. In X-ray imaging, this function would represent the distribution of X-ray attenuation coefficients and how these change with time, e.g., patient motion. The object $f(\boldsymbol{r}, t)$ could even be vector valued if we considered the dependence of attenuation coefficients on the energy spectrum of the incident beam. For the sake of simplicity, we will consider only scalar-valued objects $f(\boldsymbol{r}, t)$. Also note that the temporal dependence is often ignored when the image acquisition time is short compared to the temporal changes of the object. In cardiac imaging studies, the temporal component cannot, in general, be ignored. It is important to note that the object is represented as a continuous function and not a voxelized representation. Objects are continuous whereas image measurements of the object are generally discrete. The operator $\mathcal{H}$ describes the imaging system and how this system maps the object to a noise-free set of measurements. The physics of the imaging process are all contained in this operator $\mathcal{H}$. The operator $\mathcal{H}$ is often considered a linear, continuous-to-discrete operator in nuclear medicine. In X-ray imaging, the process of measurement can be approximately linearized by taking a logarithm of the measured data; thus, most researchers consider a linear relationship between the object and the log of the data in X-ray imaging.

The vector $\boldsymbol{n}$ is the noise just due to the detectors. We have represented this noise as additive just for notational convenience; the noise might not be additive in the conventional sense. The vector $\boldsymbol{g}$ represents the noisy image data acquired by the imaging system. This vector may or may not be an image that is ready to be viewed. In CT imaging, for example, projection images are measured by the system and a reconstruction algorithm must be applied to the image data $\boldsymbol{g}$ to generate images that can be viewed. We consider both $\boldsymbol{g}$ and $\boldsymbol{n}$ to be $M \times 1$ column vectors where M is the total number of measurement and not necessarily the number of detector elements. For example, in SPECT imaging, the cameras are often rotated around the patient; thus, each detector element at each particular angle represents one element of M. If we were to take multiple acquisitions for a temporal sequence, each time point for each detector element for each angle would comprise the elements of $\boldsymbol{g}$. The ordering of the elements of $\boldsymbol{g}$ is somewhat immaterial as long as the ordering is kept consistent.

For tomographic imaging systems, a subsequent reconstruction operator $\mathcal{O}$ is required to produce a reconstructed image for viewing, i.e.,

$$\widehat{\boldsymbol{f}} = \mathcal{O}\boldsymbol{g}. \tag{2}$$

With this notation, it is important to note that the reconstruction $\widehat{\boldsymbol{f}}$ is finite-dimensional whereas the patient being imaged $\boldsymbol{f}$ is infinite-dimensional. Thus, $\boldsymbol{f}$ and $\widehat{\boldsymbol{f}}$ cannot be directly compared using measures such as mean-squared error.

It should be noted that the operators $\mathcal{H}$ and $\mathcal{O}$ may be nonlinear. For example, for maximum-likelihood, expectation maximization (MLEM) image reconstructions, the operator $\mathcal{O}$ is nonlinear. In this chapter, we will address the objective assessment of imaging systems and not post-processed images. That is, we will cover methods for computing task performance using image data $\boldsymbol{g}$ and not reconstructed images. Any processing on the image data $\boldsymbol{g}$ to produce a reconstructed image $\widehat{\boldsymbol{f}}$ will only reduce (or at best be equal to) the amount of task-specific information present in $\boldsymbol{g}$. Thus, we assess the performance of imaging systems using the raw data produced by those systems.

Image quality is a statistical concept, and thus, a characterization of the random components in the imaging chain is necessary. The noise vector $\boldsymbol{n}$ accounts for randomness due to the measurement process. If the same object is imaged multiple times, the image data returned are not exactly the same due to measurement noise. The object $\boldsymbol{f}$ being imaged is also a random quantity because the same patient is not being imaged every time. Thus, the image data $\boldsymbol{g}$ has random contributions from at least two different sources: the object being imaged and the measurement noise. The measurement noise can depend on the object being imaged. For example, in nuclear-medicine imaging, the image data $\boldsymbol{g}$ is conditionally Poisson. That is, the probability density function $pr(\boldsymbol{g}|\boldsymbol{f})$ is Poisson distributed with mean $\mathcal{H}\boldsymbol{f}$.

2 Tasks

2.1 Classification Tasks

Classification tasks include the detection of a signal such as a tumor or the classification of a tumor as malignant or benign. The observer's decision is to classify the image into one on L classes, where L is a finite number. If only two classes are present, i.e., $L = 2$, then the

classification is called a signal-detection task. For signal-detection tasks, the data are given by

$$H_2 \quad : \quad \boldsymbol{g} = \mathcal{H}\boldsymbol{f} + \boldsymbol{n} = \mathcal{H}(\boldsymbol{f}_b + \boldsymbol{f}_s) + \boldsymbol{n}, \tag{3}$$

$$H_1 \quad : \quad \boldsymbol{g} = \mathcal{H}\boldsymbol{f} + \boldsymbol{n} = \mathcal{H}(\boldsymbol{f}_b) + \boldsymbol{n}. \tag{4}$$

Here, we are using $\boldsymbol{f}$ to represent the continuous function $f(\boldsymbol{r}, t)$ described earlier. Under the signal-present hypothesis (H_2), the object being imaged has both a background and a signal component. Under the signal-absent hypothesis (H_1), the object is just comprised of a random background. Because of obscuration, not all signals in gamma-ray imaging can be represented as additive. In fact, much of the analysis we present does not rely on the signal being additive. For either the signal-absent or signal-present hypothesis, the images are sampled from the distribution:

$$pr(\boldsymbol{g}|H_i) = \int \mathrm{d}\boldsymbol{f}\, pr(\boldsymbol{g}|\boldsymbol{f}, H_i) pr(\boldsymbol{f}|H_i). \tag{5}$$

In Eq. 5, the distribution $pr(\boldsymbol{g}|\boldsymbol{f}, H_i)$ describes the randomness in the image data given a known hypothesis and a fixed object $\boldsymbol{f}$. That is, the distribution $pr(\boldsymbol{g}|\boldsymbol{f}, H_i)$ characterizes the noise. The distribution $pr(\boldsymbol{f}|H_i)$ characterizes the randomness in the objects being imaged given the hypothesis H_i.

2.2 Estimation Tasks

Estimation tasks arise when an image or image data is used to estimate a quantity of interest. For example, the cardiac ejection fraction is commonly estimated using nuclear-medicine images. For estimation tasks, the imaging equation becomes

$$\boldsymbol{g} = \mathcal{H}\boldsymbol{f}(\boldsymbol{\theta}) + \boldsymbol{n}, \tag{6}$$

where $f(\boldsymbol{\theta})$ is the object with some particular set of parameter values $\boldsymbol{\theta}$. Both $\boldsymbol{f}$ and $\boldsymbol{\theta}$ are usually random. Note that we can describe the distribution of the image data using

$$pr(\boldsymbol{g}|\boldsymbol{\theta}) = \int \mathrm{d}\boldsymbol{f}\, pr(\boldsymbol{g}|\boldsymbol{f}, \boldsymbol{\theta}) pr(\boldsymbol{f}|\boldsymbol{\theta}), \tag{7}$$

where $pr(\boldsymbol{g}|\boldsymbol{f}, \boldsymbol{\theta})$ is the noise distribution and $pr(\boldsymbol{f}|\boldsymbol{\theta})$ describes the randomness in the objects for a particular value of $\boldsymbol{\theta}$. This equation is very similar to Eq. 5. There are indeed many similarities between the mathematics describing classification tasks and estimation tasks (Barrett and Myers 2004).

2.3 Combined Tasks

Combined tasks are common in medical imaging; they include any tasks where the observer must determine whether a signal is present or not and then, when present, must estimate parameters that characterize the signal. Any task where the observer detects the presence of a signal and also localizes the signal is a combined task. For a combined task, the image data is described by

$$H_2 \quad : \quad \boldsymbol{g} = \mathcal{H}\boldsymbol{f}(\boldsymbol{\theta}) + \boldsymbol{n} = \mathcal{H}(\boldsymbol{f}_b + \boldsymbol{f}_s(\boldsymbol{\theta})) + \boldsymbol{n}, \tag{8}$$

$$H_1 \quad : \quad \boldsymbol{g} = \mathcal{H}\boldsymbol{f} + \boldsymbol{n} = \mathcal{H}(\boldsymbol{f}_b) + \boldsymbol{n}. \tag{9}$$

Much like before, a useful distribution characterizing the randomness in the image data is

$$pr(\boldsymbol{g}|H_i,\boldsymbol{\theta}) = \int \mathrm{d}\boldsymbol{f}\, pr(\boldsymbol{g}|H_i,\boldsymbol{f},\boldsymbol{\theta})pr(\boldsymbol{f}|H_i,\boldsymbol{\theta}). \tag{10}$$

One caveat to the above equation is that the $\boldsymbol{\theta}$ is not present when $H_i = H_1$, the signal-absent case.

2.4 Distributions

In the previous section, we defined, $pr(\boldsymbol{g}|H_i)$, $pr(\boldsymbol{g}|\boldsymbol{\theta})$, and $pr(\boldsymbol{g}|\boldsymbol{\theta},H_i)$. These distributions are usually called the likelihood distributions because they describe the likelihood of the data given the state relating to the task. The overall probability of the image data is obtained by marginalization. For the case of estimation tasks, this results in

$$pr(\boldsymbol{g}) = \int \mathrm{d}\boldsymbol{\theta}\, pr(\boldsymbol{g}|\boldsymbol{\theta})pr(\boldsymbol{\theta}), \tag{11}$$

where $pr(\boldsymbol{\theta})$ describes the randomness in the parameters being estimated. This distribution is referred to as the prior distribution. It should be noted that a Bayesian definition of probability is not required to define this prior distribution. Expressions similar to Eq. 11 exist for the other tasks.

3 Observers

An observer, quite simply, performs the task. Observers can be humans, computer algorithms, or algorithms that are guided by humans. For the goal of system design, the goal is to generate the most amount of task-specific information in the image data as possible. Thus, so-called ideal observers represent a viable strategy for measuring image quality for imaging hardware. For evaluating image-reconstruction algorithms, human observers or models of human observers are more relevant.

For signal-detection tasks, all observers can be thought of as mapping the image data $\boldsymbol{g}$ to a signal number which represents the confidence level that a signal is present. Thus, the observer requires a mapping function $T(\boldsymbol{g})$ and a threshold t_c and performs the following operation:

$$D_2 \quad : T(\boldsymbol{g}) \geq t_c, \tag{12}$$

$$D_1 \quad : T(\boldsymbol{g}) < t_c, \tag{13}$$

where D_i implies that the observer decided class i. Note the key difference between H_i and D_i; H_i relates to the truth of whether a signal is actually present or not, whereas D_i refers to the decision (possibly incorrect decision) that the observer makes.

For estimation tasks, the observer takes the image data and produces an estimate $\widehat{\boldsymbol{\theta}}$ of the parameters $\boldsymbol{\theta}$,

$$\widehat{\boldsymbol{\theta}} = \boldsymbol{\Theta}(\boldsymbol{g}). \tag{14}$$

For combined tasks, the observer does both of the previous operations. That is,

$$D_2 \quad : T(\boldsymbol{g}) \geq t_c; \widehat{\boldsymbol{\theta}} = \boldsymbol{\Theta}(\boldsymbol{g}), \tag{15}$$

$$D_1 \quad : T(\boldsymbol{g}) < t_c. \tag{16}$$

Thus, this observer only estimates the parameters $\boldsymbol{\theta}$ when the signal is present. Otherwise, there are no parameters to estimate.

3.1 Figures of Merit

For signal-detection tasks, there are two probability density functions which characterize the observer's decision variable: $pr(t|H_2)$ and $pr(t|H_1)$. These are the density functions for the test statistics when the signal is present (i.e., H_2) and when the signal is absent (i.e., H_1). In general, the more overlap between these two densities, the poorer the observer performance. The amount of overlap can be characterized either using the test-statistic signal-to-noise ratio (SNR) or receiver operating characteristic (ROC) (Metz 1986, 1989) analysis. The SNR is given by

$$\mathrm{SNR}^2 = \frac{(\bar{t}_2 - \bar{t}_1)^2}{\frac{1}{2}(\sigma_2^2 + \sigma_1^2)}, \tag{17}$$

where $\bar{t}_i = \langle t \rangle_{t|H_i}$ is the mean test statistic under the H_i hypothesis and $\sigma_i^2 = \left\langle (t - \bar{t}_i)^2 \right\rangle_{t|H_i}$ is the variance of the test statistic under the H_i hypothesis. ROC analysis plots the fraction of correctly classified signal-present images (the true-position fraction or TPF) versus the fraction of incorrectly classified signal-absent images (the false-positive fraction or FPF) as a function of the decision threshold t_c. The TPF and FPF are given by

$$\mathrm{TPF}(t_c) = \int_{t_c}^{\infty} pr(t|H_2)\,\mathrm{d}t, \tag{18}$$

$$\mathrm{FPF}(t_c) = \int_{t_c}^{\infty} pr(t|H_1)\,\mathrm{d}t. \tag{19}$$

An example ROC curve is shown in ❯ *Fig. 1*. The area under the receiver operating characteristic (ROC) curve or AUC is another common figure of merit. A perfect (never makes a mistake) observer has an AUC of 1. An observer that guesses the answer each time has an AUC of 0.5.

For estimation tasks, the ensemble mean-square error (EMSE) is a common figure of merit and is defined as

$$\mathrm{EMSE} = \left\langle \left\langle \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2 \right\rangle_{\boldsymbol{g}|\boldsymbol{\theta}} \right\rangle_{\boldsymbol{\theta}}. \tag{20}$$

This quantity is called *ensemble* because the figure of merit is averaged over the ensemble of the parameters being estimated $\boldsymbol{\theta}$. Other figures of merit can be defined for estimation tasks. As we will see, the choice of the figure of merit changes the optimal estimator.

For combined detection/estimation tasks, there is the estimation receiver operating characteristic (EROC) analysis (Clarkson 2007). The x-axis of an EROC curve is the false-positive fraction as in the standard ROC curve. The y-axis is the average value of a utility function on the parameters being estimated for the true-positive cases. The area under the EROC curve (AEROC) is a figure of merit that characterizes the ability of the observer to perform the combined task. Location ROC analysis (LROC) is a specific form of EROC analysis where the estimation task is to locate the signal.

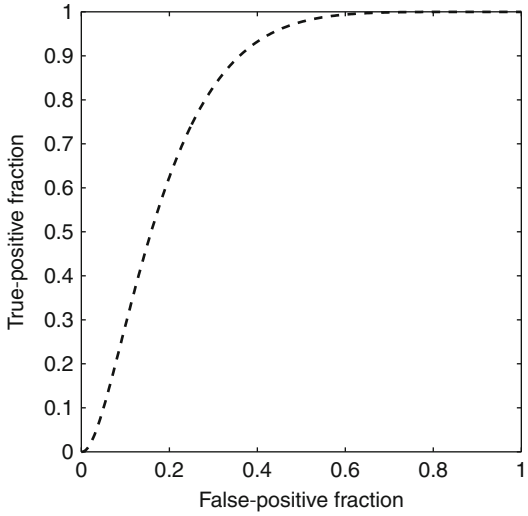


Fig. 1
An example ROC curve

4 Ideal Observers

The Bayesian ideal observer makes decision using the likelihood ratio

$$\Lambda(\boldsymbol{g}) = \frac{pr(\boldsymbol{g}|H_2)}{pr(\boldsymbol{g}|H_1)}, \tag{21}$$

or any monotonic transformation of the likelihood ratio. The Bayesian ideal observer requires complete knowledge of the two likelihoods describing the image data. The ideal observer is the observer that has the best possible ROC curve and the highest area under the ROC curve. The Bayesian ideal observer is not necessarily the observer with the maximum SNR.

For estimation tasks, the best possible observer depends on the costs associated with the estimates. For example, when the cost of an incorrect estimate is the EMSE, then the ideal estimator is known as the posterior mean estimator,

$$\widehat{\boldsymbol{\theta}}_{\mathrm{PM}} = \int \mathrm{d}\boldsymbol{\theta}\, \boldsymbol{\theta} pr(\boldsymbol{g}|\boldsymbol{\theta}) pr(\boldsymbol{\theta}) / pr(\boldsymbol{g}). \tag{22}$$

When all incorrect decisions beyond a tolerance threshold are equally costly, the ideal estimator is the MAP estimator given by

$$\widehat{\boldsymbol{\theta}}_{\mathrm{MAP}} = \arg\max_{\boldsymbol{\theta}} pr(\boldsymbol{g}|\boldsymbol{\theta}) pr(\boldsymbol{\theta}). \tag{23}$$

In either case, complete knowledge of the randomness in the image data and the parameter to be estimated is required to compute the estimate.

The ideal observer for combined tasks depends on a number of factors. A good description of the methods required is given in Khurd and Gindi (2005) as well as in Clarkson (2007).

5 Ideal Linear Observers

A linear observer is one that uses only linear manipulations on the image data to make decisions. For example, a linear observer performing a detection task produces test statistics using

$$t = T(\boldsymbol{g}) = \boldsymbol{w}^{\dagger}\boldsymbol{g}, \tag{24}$$

where $\boldsymbol{w}$ is a vector with the same dimension as $\boldsymbol{g}$ and † denotes the transpose operation. We will assume that the image data $\boldsymbol{g}$ are always real valued (i.e., never complex); so we will not worry about any complex conjugates in this chapter. The linear observer that maximizes SNR (Eq. 17) is known as the Hotelling observer (Barrett and Myers 2004). The Hotelling observer produces test statistics using

$$t = \boldsymbol{w}^{\dagger}\boldsymbol{g} = \boldsymbol{s}^{\dagger}K^{-1}\boldsymbol{g}, \tag{25}$$

where

$$\boldsymbol{s} = \langle \boldsymbol{g} \rangle_{\boldsymbol{g}|H_2} - \langle \boldsymbol{g} \rangle_{\boldsymbol{g}|H_1} = \mathcal{H}\overline{\boldsymbol{f}}_s. \tag{26}$$

The signal $\boldsymbol{s}$ is the difference between the mean data under the two hypotheses, i.e., it is the average signal to be detected. The covariance matrix K is an $M \times M$ matrix that measures the variability in the image data. The ith diagonal element of K is the variance in g_i. The off-diagonal element K_{ij} denotes the covariance between the ith and jth elements of $\boldsymbol{g}$. A large covariance indicates that the ith and jth components of $\boldsymbol{g}$ trend with one another. A near-zero covariance indicates that these components do not vary with one another. The Hotelling observer is not necessarily the linear observer that maximizes the ROC curve.

For most modern medical imaging systems, a direct application of Eq. 25 is difficult because the dimension M of the image data is very large. It is difficult to use conventional methods to estimate and invert this very large matrix. An alternative method for estimating and inverting this large covariance matrix K is to use a covariance matrix decomposition and the Woodbury matrix inversions lemma (Woodbury 1950). We start by decomposing the covariance matrix K into two components using techniques described fully in (Barrett and Myers 2004), to achieve

$$K = \overline{K}_{\boldsymbol{n}} + K_{\overline{\boldsymbol{g}}}. \tag{27}$$

Here $\overline{K}_{\boldsymbol{n}}$ is the noise covariance matrix averaged over all patients and the term $K_{\overline{\boldsymbol{g}}}$ is the covariance of noise-free images. This decomposition is completely general and does not require any assumptions about linear imaging systems or even statistical independence of the objects $\boldsymbol{f}$ and the measurement noise $\boldsymbol{n}$. The key advantage of the decomposition is that the term $\overline{K}_{\boldsymbol{n}}$ is often known from a physical and statistical model or can be easily estimated. For example, in nuclear-medicine imaging, the measurement noise is known to be Poisson and, hence, the matrix $\overline{K}_{\boldsymbol{n}}$ is a diagonal matrix whose diagonal elements are given by the mean of the image data (which is also the variance for Poisson statistics). Because $\overline{K}_{\boldsymbol{n}}$ is known and is, most often, diagonal, then any estimate of K produced using Eq. 27 is necessarily full rank. So once we have determined what $\overline{K}_{\boldsymbol{n}}$ is from the physics of the imaging process, we can use any number of methods to estimate $K_{\overline{\boldsymbol{g}}}$ and the resulting estimate of the full covariance matrix will be full rank and invertible.

The term $K_{\overline{\boldsymbol{g}}}$ is the covariance of noise-free images of objects. This covariance can be determined using simulations of the imaging system and objects or it can be measured for some modalities. For example in nuclear medicine, long exposure images of objects will result in projections that are essentially noise free. These noise-free projections can then be used to produce

a sample estimate of $K_{\overline{g}}$. Thus, the sample covariance matrix to be used in the computation of the Hotelling observer is

$$\widehat{K} = \widehat{\overline{K}}_{\boldsymbol{n}} + \widehat{K}_{\overline{g}}, \tag{28}$$

where $\widehat{\overline{K}}_{\boldsymbol{n}}$ is an estimate of the diagonal $\overline{K}_{\boldsymbol{n}}$ matrix and $\widehat{K}_{\overline{g}}$ is estimated from noise-free sample images.

Another benefit of this matrix decomposition is that the burden of the inversion of a large covariance matrix is mitigated by using the matrix inversion lemma (Barrett et al. 2001). We can rewrite ❯ Eq. 28 as

$$\widehat{K} = \widehat{\overline{K}}_{\boldsymbol{n}} + ZZ^{\dagger}, \tag{29}$$

where Z is a matrix whose dimensions are the number of detector elements M by the number of samples $N_1 + N_2$. The matrix Z is a compact means of representing sample covariance matrices. The matrix inversion lemma states that

$$\left[\widehat{\overline{K}}_{\boldsymbol{n}} + ZZ^{\dagger}\right]^{-1} = \left[\widehat{\overline{K}}_{\boldsymbol{n}}\right]^{-1} - \left[\widehat{\overline{K}}_{\boldsymbol{n}}\right]^{-1} Z\left[I + Z^{\dagger}\left[\widehat{\overline{K}}_{\boldsymbol{n}}\right]^{-1} Z\right]^{-1} Z^{\dagger}\left[\widehat{\overline{K}}_{\boldsymbol{n}}\right]^{-1}. \tag{30}$$

This equation is composed of two separate inverses: $\left[\widehat{\overline{K}}_{\boldsymbol{n}}\right]^{-1}$ which is the inverse of a diagonal matrix and the inverse of a matrix whose dimension is the number of samples by the number of samples. Thus, the matrix inversion lemma greatly simplifies the computational effort required to invert large covariance matrices.

5.1 Linear Estimation Tasks

A linear observer performing an estimation task computes

$$\widehat{\boldsymbol{\theta}} = W^{\dagger}\boldsymbol{g} + \boldsymbol{c}, \tag{31}$$

where W is a matrix whose dimension is the number of detector elements M by the number of parameters to be estimated P. The vector constant $\boldsymbol{c}$ is an $M \times 1$ vector. The linear estimator that minimizes the EMSE is called the Wiener estimator and is given by

$$\widehat{\boldsymbol{\theta}} = K_{\boldsymbol{\theta},\boldsymbol{g}} K^{-1}(\boldsymbol{g} - \overline{\boldsymbol{g}}) + \overline{\boldsymbol{\theta}}, \tag{32}$$

where $K_{\boldsymbol{\theta},\boldsymbol{g}}$ is the cross-covariance between the image data $\boldsymbol{g}$ and the parameters to be estimated $\boldsymbol{\theta}$, and $\overline{\boldsymbol{\theta}}$ is the mean of the parameters to be estimated. This equation is similar to the Hotelling observer for detection tasks. Just like the case of the Hotelling observer, the covariance matrix K can be estimated directly from samples. A matrix decomposition can also be employed to ease the burden of estimating and inverting K. As before,

$$K = \overline{K}_{\boldsymbol{n}} + K_{\overline{g}}. \tag{33}$$

However, $\overline{K}_{\boldsymbol{n}}$ is the noise covariance averaged over object variability and parameter $\boldsymbol{\theta}$ variability. The second term $K_{\overline{g}}$ is the covariance matrix of the noise-free image data but including parameter variations. As before, the first term in ❯ Eq. 33 is often diagonal and can be estimated using image data. The matrix inversion lemma can again be employed to invert estimates of the covariance matrix computed using the above expansion.

6 Notes on Linear Observers

The Hotelling and Wiener observers use only linear manipulations of the image data. Because of this, they are limited in their performance in certain situations. A paper by Whitaker et al. (2008) has shown that the Wiener estimator performs poorly when there is signal location and/or shape variability. The authors showed that the Wiener estimator was unable to use the image data to simultaneously estimate signal size, location, and amplitude. It is also expected that the Hotelling observer will perform very poorly if the location of the signal is random. While the computational efforts required for linear estimation are not great, their application is generally limited to specific tasks such as the detection of a known signal at a known location or the estimation of signal amplitude when the signal location and size are known.

7 Combination Task Observer Models

The observer models we have discussed thus far have been for detection tasks or estimation tasks only. Combined observer models do, however, exist and these models have led to a new class of observer model that can be applied to any type of detection task, estimation task, or combination detection/estimation task. The pioneering work of Khurd and Gindi (2005) developed an observer model that maximized that area under the LROC curve. This observer model required a search over potential signal locations and the application of an observer that computes a test statistic as a function of signal location. That is, the observer's test statistic is given by

$$t = \max_{\boldsymbol{r}'} \left\{ \left\langle \Lambda(\boldsymbol{g}|\boldsymbol{r}) u(\boldsymbol{r}, \boldsymbol{r}') \right\rangle_{\boldsymbol{r}} \right\}, \tag{34}$$

where $\Lambda(\boldsymbol{g}|\boldsymbol{r})$ is the ideal observer when the signal is located at position $\boldsymbol{r}$, and $u(\boldsymbol{r}, \boldsymbol{r}')$ is the utility function of locating the signal at position $\boldsymbol{r}'$ when the true location is $\boldsymbol{r}$ (Khurd and Gindi 2005). If t is above a detection threshold, then the estimate of the signal location is given by

$$\boldsymbol{r} = \arg\max_{\boldsymbol{r}'} \left\{ \left\langle \Lambda(\boldsymbol{g}|\boldsymbol{r}) u(\boldsymbol{r}, \boldsymbol{r}') \right\rangle_{\boldsymbol{r}} \right\}. \tag{35}$$

Clarkson generalized this analysis to allow for the estimation of other parameters instead of just signal location (Clarkson 2007) and he generated a new form of a ROC analysis called Estimation ROC or EROC analysis.

The work of Whitaker (Whitaker et al. 2008) went a step further and generated an estimation observer model that uses only the second step shown in ❯ Eq. 35 – the scanning, linear estimator. In addition, Whitaker et al. also used a simplifying assumption to replace the expectation in ❯ Eq. 35 with a Gaussian function. Using Gaussian assumptions, scanning Hotelling observers have also been developed that implement ❯ Eq. 34. Both the scanning linear estimator and scanning Hotelling observers are nonlinear because of the maximization step. However, the computation of the objective function depends on only first- and second-order statistics of the image data $\boldsymbol{g}$. While the objective function may be a straightforward computation, the maximization of this objective function can be difficult.

8 Discussion

We have discussed Bayesian ideal observer models and found that, in general, they require complete knowledge of the statistics of the image data; this is usually not possible for realistic situations. Ideal linear observers on the other hand can be computed efficiently and practically. However, the utility of these observer models is limited to simple tasks such as the detection of a signal at a known location. More modern observer models such as the scanning linear estimator can be related to an ideal observer; so they have a strong foundation in signal-detection theory. However, they are also practical to compute. These newer observer models have found a great deal of use in medical imaging, homeland security, and astronomical imaging.

Objective assessment of image quality requires a knowledge or measurement of the statistics governing the imaging problem. Object variability, measurement noise, and task variability (i.e., the randomness in the truth state H_i) all play a role in the development of observer models for assessing image quality. Clearly the computational burden of computing these observers is more difficult than conventional quality metrics such as resolution or contrast. However, these metrics address the reason the images were acquired in the first place – the task. Thus, the effort to optimize imaging systems based on these measures of image quality should result in systems that enhance overall task performance.

References

Barrett HH, Myers KJ (2004) Foundations of Image Science. Wiley, Hoboken

Barrett HH, Myers KJ, Gallas BD, Clarkson E, Zhang H (2001) Megalopinakophobia: its symptoms and cures. In: Antonuk LE, Yaffe MJ (eds) Medical imaging 2001: physics of medical imaging. SPIE, Bellingham, WA, pp 299–307

Clarkson E (2007) The estimation receiver operating characteristic curve and ideal observers for combined detection/estimation tasks. JOSA A 24(12):B91–B98

Khurd P, Gindi G (2005) Decision strategies maximizing the area under the LROC curve. In: Eckstein MP, Jiang Y (eds) Medical imaging 2005: image perception, observer performance, and technology assessment, vol 5749. SPIE, Bellingham, WA, pp 150—161

Metz CE (1986) ROC methodology in radiologic imaging. Invest Radiol 21:720–733

Metz CE (1989) Some practical issues of experimental design and data analysis in radiological ROC studies. Invest Radiol 24:234–245

Whitaker MK, Clarkson E, Barrett HH (2008) Estimating random signal parameters from noisy images with nuisance parameters: linear and scanning-linear methods. Optics Exp 16(11):8150–8173

Woodbury MA (1950) Inverting modified matrices. Statistical Research Group, Princeton University, Princeton

44 Simulation of Medical Imaging Systems: Emission and Transmission Tomography

Robert L. Harrison
University of Washington, Seattle, WA, USA

1 Introduction ***1097***

2 History of Simulation ***1097***

3 Statistical Methods for Simulations ***1099***
3.1 Random Number Generators 1099
3.2 Sampling from Probability Density Functions (PDFs) 1100
3.2.1 Sampling from the Exponential Function Using the Inversion Method 1100
3.2.2 The Acceptance–Rejection Sampling Method 1102
3.2.3 Table Lookup 1103
3.2.4 Library Functions and Sampling from the Normal and Poisson Distributions 1103

4 Basic Principles and Physics of Medical Imaging Simulation ***1105***
4.1 Sources of Photons 1106
4.1.1 Nuclear Decay: Source of Photons for Emission Tomography 1106
4.1.2 Electromagnetic Radiation: Source of Photons for X-ray CT 1108
4.1.3 Secondary Sources of Photons 1109
4.2 Tracking Photons Through Matter 1109
4.2.1 Description of Attenuation 1110
4.2.2 Where Will a Photon Interact? 1112
4.2.3 What Type of Interaction? 1112
4.2.4 Simulating Photoelectric Absorption 1113
4.2.5 Simulating Compton Scatter 1113
4.2.6 Simulating Coherent Scatter 1114
4.2.7 Simulating Pair Production 1115
4.3 Simulating Detection 1115
4.3.1 Tracking Photons in the Detector Crystal 1115
4.3.2 Converting Deposited Energy to a Signal 1116

C. Grupen, I. Buvat (eds.), *Handbook of Particle Detection and Imaging*,
DOI 10.1007/978-3-642-13271-1_44, © Springer-Verlag Berlin Heidelberg 2012

4.3.3 Histogramming Simulated Events 1117
4.4 Acceleration of Photon-Tracking Simulations 1119

5 Available Simulation Software 1120

6 Choosing a Simulation Tool 1121

7 Online Resources 1122

8 Conclusion 1122

9 Cross-References 1122

References 1123

Abstract: Simulation is an important tool in medical imaging research. In patient scans the true underlying anatomy and physiology is unknown. We have no way of knowing in a given scan how various factors are confounding the data: statistical noise; biological variability; patient motion; scattered radiation, dead time, and other data contaminants. Simulation allows us to isolate a single factor of interest, for instance when researchers perform multiple simulations of the same imaging situation to determine the effect of statistical noise or biological variability. Simulations are also increasingly used as a design optimization tool for tomographic scanners. This article gives an overview of the mechanics of emission and transmission tomography simulation, reviews some of the publicly available simulation tools, and discusses trade-offs between the accuracy and efficiency of simulations.

1 Introduction

A simulation is a computational model of a physical system. The model is a simplified version of reality, but one where the exact form of the simulated system is known and where details that are hidden in the physical system can be observed.

Simulation has a long history. For example, as far back as ancient Greece, people have built geocentric orreries, mechanical simulations of the solar system used to predict the position of the sun, moon, and planets relative to earth. At some point – certainly by the eighteenth century – some added stochastic elements to their simulations, early forerunners of what we now call Monte Carlo simulation. Monte Carlo simulation is now a much-used scientific tool for problems that are analytically intractable and for which experimentation is too time-consuming, costly, or impractical.

Tomographic patient scans are good examples of such problems. The true anatomy/physiology that is being probed will never be known; the physical processes and detection systems are very complex. We have no way to know in a given scan how various factors are confounding the data: stochastic noise, biological variability, scattered radiation, patient motion, and the detection system. Simulation provides an unparalleled window for examining these factors: the same experiment can be run repeatedly to examine stochastic noise, or a series of experiments can be run to examine the effect of a single parameter. As a result, simulation is used extensively in the design of tomographs, in the investigation of physical effects that cannot be quantified using experiments, in the development of data correction and image reconstruction algorithms, and in the development of imaging protocols.

2 History of Simulation

Early simulations, like the orreries, used mechanisms to help understand or predict the workings of larger/more complex objects. We still make use of this type of simulator, for instance, flight simulators for the training of pilots. Such simulations are usually deterministic: if the planets started in this orientation, then in a month Saturn will rise at midnight; if we change the wing flaps so, the plane will turn to the right. However, much of the modern use of simulation is stochastic, or Monte Carlo simulation.

"Do random events ever lead to concrete results? Seems unlikely – after all, they're random" (Burger and Starbird 2005). It is somewhat counterintuitive to think that flipping a coin millions, billions, or trillions of times could give us insight into what is happening in a nuclear bomb, but it was precisely this problem that lead to the development of the fundamental methods of Monte Carlo simulation. Two scientists at Los Alamos Scientific Laboratory, John von Neumann and Stanislaw Ulam, used it to investigate the distance that neutrons traveled through radiation shielding. Despite the fact that the basic properties of neutrons were, by that point, well understood, the problem could not be solved with analytical calculations. Monte Carlo simulation became a key tool in the Manhattan Project, the American Second World War effort to develop an atomic bomb.

Their methods were based on the observation that integrals can be solved stochastically. For instance, if we want to integrate a function $y = f(x)$ on the interval $[0,1]$ with range $[0,1]$, we can generate N random pairs

$$(x_i, y_i) \in \{(x,y) : x \in [0,1], y \in [0,1]\}, \tag{1}$$

then estimate the integral as

$$\int_{x=0}^{1} f(x)\,\mathrm{d}x = \frac{m}{N}, \tag{2}$$

where m is the number of i such that

$$y_i < f(x_i). \tag{3}$$

In general, the more random pairs we generate, the better our estimate of the integral will be. Doing a quick experiment to estimate the integral of $y = x^2$ on the interval $[0,1]$, the estimate with $N = 100$ is 35/100 (*Fig. 1*); an estimate with $N = 10{,}000$ is 3,378/10,000, closer to the expected result of 1/3.

The first recorded use of this kind of integration was by Georges Louis LeClerc, Comte de Buffon (1707–1788), an influential French scientist who used random methods in a number of studies. He developed a method of estimating π by dropping a needle on a lined background that became a Victorian-era parlor game. He reportedly tested it by throwing baguettes over his shoulder onto a tile floor. In the nineteenth and early twentieth centuries, simulation was increasingly used as an experimental means of confirming theory, analyzing data, or supplementing intuition in science and mathematics, especially in statistics.

The early studies were seminal: the idea that randomness is an experimental tool rather than an impediment to science was revolutionary. However, there is a significant difference between them and typical modern Monte Carlo simulations studying problems that are otherwise intractable. The early work dealt with problems that could be (often had been) solved with existing theory: for instance Archimedes (287–212 BC) developed a rigorous method to establish π's value well before LeClerc's time. Modern simulations invert this process, using stochastic methods as the primary means of determining a problem's solution.

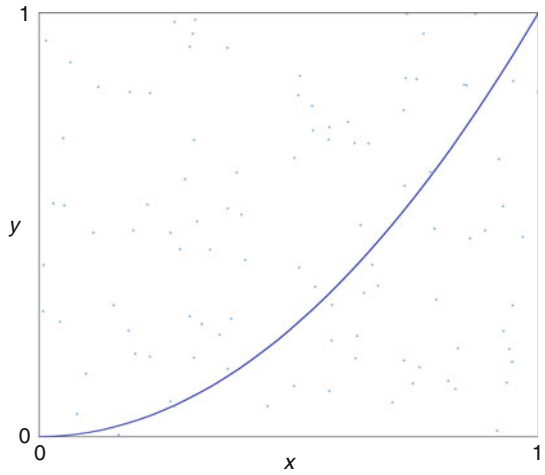


Fig. 1
We can estimate an integral of a function by randomly generating points in a region containing the function

3 Statistical Methods for Simulations

3.1 Random Number Generators

How can a computer generate a random number?

In general they do not. There are sources of truly random numbers, e.g., time between (detected) decays from a radioactive source – there are even lists of random numbers derived from such sources, though these lists are far too short for complex simulations, which often require billions or even trillions of random numbers. In general Monte Carlo simulations depend on pseudorandom number generators (often called an RNG, with the pseudo left out): algorithms that produce sequences of numbers that are random-like, but that are quick, reliable, and produce sequences that can be reproduced for program testing or debugging.

An RNG is a deterministic computer algorithm that produces a series of numbers that shares many of the characteristics of truly random samples. A simple RNG that used to be very common is the linear congruential generator (LCG). It is easy to understand and implement. Given the nth sample, I_n, the next sample is

$$I_{n+1} = (aI_n + c) \bmod m, \tag{4}$$

where a, c, and m must be carefully chosen for best performance – one should not make these numbers oneself, tables of appropriate values can be found online. LCGs have the advantages that they are easy to understand, implement, and relatively quick. However, it is not really appropriate for most modern uses – maybe for one's homework or video game software. Even with the best choice of constants LCGs present problems. One is that they repeat themselves after m samples; even more serious, if one repeatedly fills n-tuples using LCGs, $(I_j, I_{j+1}, \ldots, I_{j+n})$, the results fall on hyperplanes in n-space (Knuth 1997).

However, there are many very good pseudo-RNGs available, often in open source or public domain versions. Suites of tests have been developed to check how "random" they are (e.g., TestU01, L'Ecuyer and Simard 2007). In particular, many publicly available Monte Carlo simulations now use the Mersenne Twister (Matsumoto and Nishimura 1998). It is complicated but fast, passes most tests except for some that are for RNGs used in encryption, and has a freely available source code.

Most RNGs uniformly sample positive integer values between 0 and N. To sample the uniform distribution on $[0,1]$, $\mathbf{U}[0,1]$, a sample, u, from the RNG is divided by N – the range of the distribution can also be open or half open as needed, e.g., $[0,1)$. Most RNGs allow N to be large enough that the double precision numbers between 0 and 1 are reasonably uniformly sampled using this method. The remainder of this section shows how the uniform distribution can be transformed into other distributions of interest.

3.2 Sampling from Probability Density Functions (PDFs)

A PDF is a nonnegative, real-valued function with an integral of 1 over its range. It gives the relative likelihood of the possible outcomes for a random sample from a given distribution. For example, for the $\mathbf{U}[0,1]$ the PDF is

$$p(x) = \begin{cases} 1 & 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}. \tag{5}$$

For sampling purposes we often use the cumulative distribution function (CDF), which is defined as the integral of the PDF. For $\mathbf{U}[0,1]$ the CDF is:

$$P(x) = \int_{-\infty}^{x} p(x)\,\mathrm{d}x = \begin{cases} 0 & x < 0 \\ x & 0 \le x \le 1 \\ 1 & 1 \le x \end{cases}. \tag{6}$$

3.2.1 Sampling from the Exponential Function Using the Inversion Method

In general, whatever distribution we want to sample from, we start by sampling from $\mathbf{U}[0,1]$ and then convert the sample to the desired distribution. The inversion method is a commonly used conversion method. It uses the uniform sample to "look up" a number on the other distribution's CDF. We will demonstrate this using the exponential distribution.

The exponential distribution is very useful; it is, for instance, the distribution for the time between decays in a radioactive sample or the distance an X-ray or gamma ray travels before an interaction in a uniform medium. The PDF of the exponential function is

$$f(x;\lambda) = \begin{cases} \lambda \mathrm{e}^{-\lambda x} & 0 \le x \\ 0 & x < 0 \end{cases}, \tag{7}$$

where λ is the rate of exponential decay (or attenuation coefficient) (Fig. 2a). The mean and standard deviation of the distribution are both $\frac{1}{\lambda}$. The CDF of the exponential function (Fig. 2b) is

$$F(x;\lambda) = \begin{cases} 1 - \mathrm{e}^{-\lambda x} & 0 \le x \\ 0 & x < 0 \end{cases}. \tag{8}$$

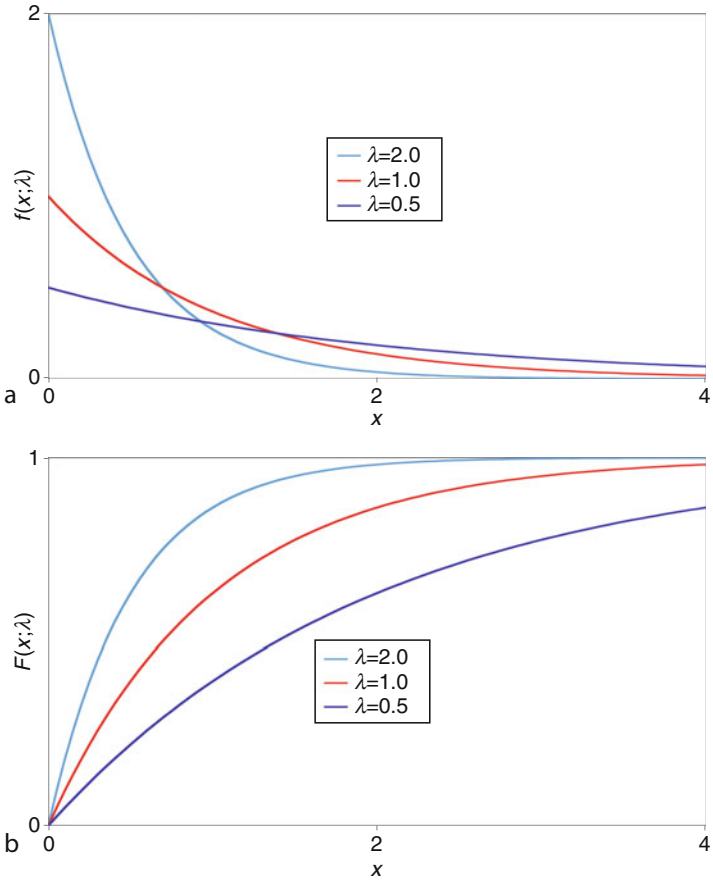


Fig. 2
Exponential Distribution. (a) PDF (b) CDF

To use the inversion method to sample from the exponential distribution, we use the following algorithm:

– Sample u from the uniform distribution on $(0,1)$.
– Locate u on the y-axis of the CDF of the target function, in this case the exponential CDF.
– Find the x-value that corresponds to u (*Fig. 3*):

$$x = F^{-1}(u) \Rightarrow x = -\frac{\ln(1-u)}{\lambda} \tag{9a}$$

as $1-u$ has the same distribution as u, we simplify this to

$$x = -\frac{\ln(u)}{\lambda}. \tag{9b}$$

This sampling method can be used whenever the target CDF is invertible.

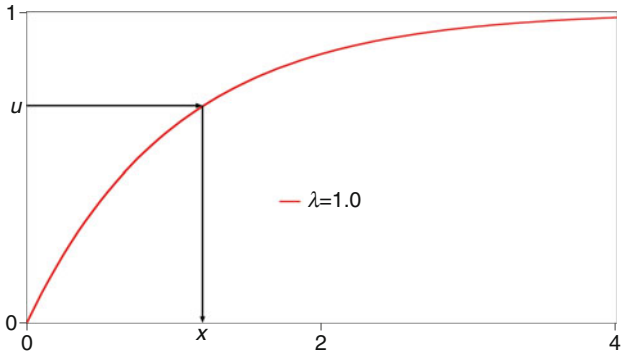


Fig. 3
Inversion Method. To sample from a CDF using the inversion method, take a sample from U[0, 1), *u*, and find *x* such that $F(x) = u$

3.2.2 The Acceptance–Rejection Sampling Method

The acceptance–rejection sampling method, also called rejection sampling, replaces sampling from a difficult-to-sample PDF with sampling from an easy-to-sample PDF. It is often used when the CDF is not analytically invertible or the inversion is computationally burdensome. It is an exact method, not an approximation.

To sample from a PDF $f(x)$:

1. Choose an easy-to-sample PDF $g(x)$ (e.g., with an invertible CDF) and a constant c such that
$$c * g(x) \geq f(x) \tag{10}$$
for every x.
2. Generate a random number v from $g(x)$ (e.g., using the inversion method).
3. Generate a random number u from $\mathbf{U}[0, 1]$.
4. If
$$c * u < \frac{f(v)}{g(v)}, \tag{11}$$
then accept v as the random sample.
5. Otherwise reject v and start again at step 2.

What happens is that sampling v from $g(x)$ gives the wrong distribution for $f(x)$ (of course), but in step 4 we reject v in proportion to how far off the sampling is at v (Fig. 4). On average the method requires c iterations of steps 2–5 to produce a sample. As each iteration requires two random numbers, the acceptance–rejection method requires $2 * c$ RNG calls to produce a sample. In some simulations a significant portion of the computational cost is in the random sampling, so the choice of c and $g(x)$ can make a significant difference to the computational efficiency.

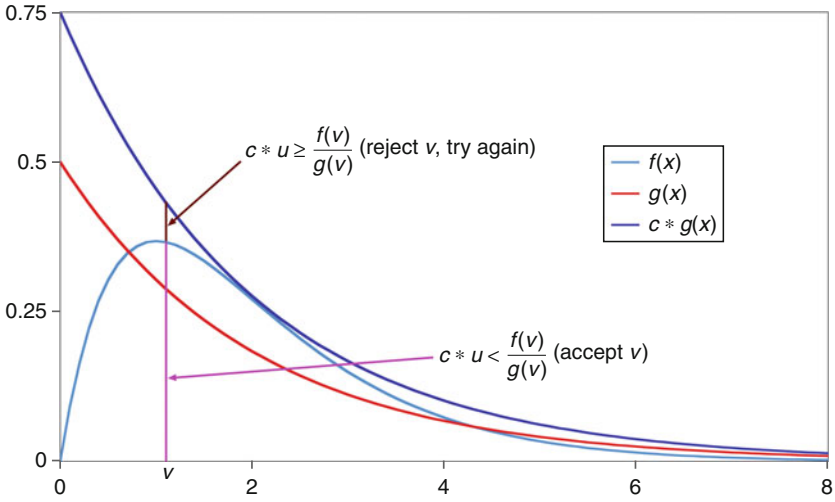


Fig. 4

Acceptance–Rejection Method: a random sample from $g(x)$, v, is accepted as a random sample from $f(x)$ if $c * u < \frac{f(v)}{g(v)}$, rejected otherwise

3.2.3 Table Lookup

A PDF that closely approximates the target PDF, for instance a piecewise linear approximation, can also be used for sampling. Such approximations often require a table lookup and a linear interpolation. They can be very computationally efficient, but anytime one samples from the wrong distribution it results in a biased simulation. For table lookups, increasing the size of the table can often reduce the bias to an acceptable level. The bias can be completely eliminated by using the acceptance–rejection method, but the extra random sample may eliminate the efficiency improvement.

Sometimes bias is unavoidable, for instance if the PDF is based on experimental data, which is inevitably only available at certain discrete points and which includes some experimental uncertainty.

3.2.4 Library Functions and Sampling from the Normal and Poisson Distributions

Random sampling functions for many common PDFs can be found in many open source mathematical/scientific/statistical code libraries. It is usually best to use these for difficult-to-sample distributions. I recommend their use for two other important distributions: the normal and Poisson distributions.

The normal distribution (also known as the Gaussian distribution and bell curve) is often used to approximate unknown distributions, for instance the errors in experimental measurements. It is often reasonable to do so, particularly when the unknown distribution is actually

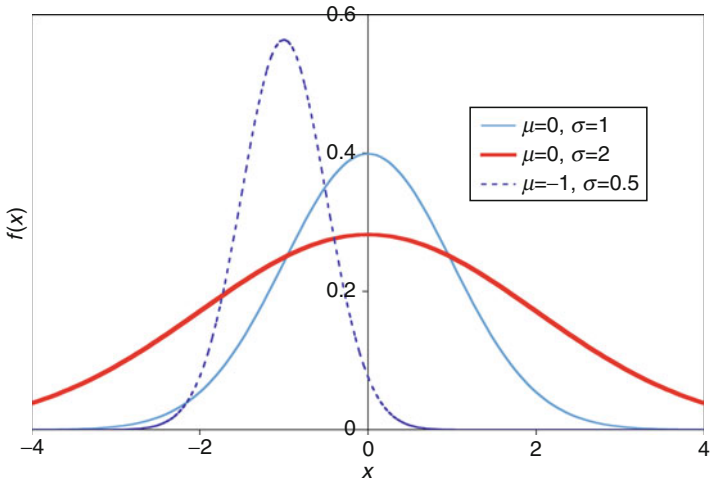


Fig. 5
A family of PDFs for the normal distribution

a combination of many smaller distributions. By the central limit theorem, the sum of a large number of independent random variables will be approximately normally distributed.

The PDF for the normal distribution is

$$\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \tag{12}$$

where μ is the mean of the distribution and σ is the standard deviation (Fig. 5). The CDF of the normal distribution is a complicated function, but invertible, so the inversion method could be used for sampling. However, more computationally efficient algorithms have been developed, in particular the ziggurat, Box–Muller, and Marsaglia polar algorithms.

The Poisson distribution is a discrete distribution that takes on nonnegative integer values. When one counts events that happen at a constant rate, the number of events that occur in a given time period will be a Poisson random variable. Thus, for instance, the number of events detected on a given line of response in a PET or SPECT scan is a Poisson random variable, or the number of decays from a radioactive sample in a minute. Indeed, the Poisson distribution is closely related to the exponential distribution: when the distribution of time elapsed between events or decays is an exponential random variable, the number of events or decays in a time interval is a Poisson random variable.

The probability mass function (PMF, the discrete equivalent of a PDF) for the Poisson distribution (Fig. 6) is

$$f(k) = \frac{\lambda^k e^{-\lambda}}{k!} \tag{13}$$

for k a nonnegative integer, where λ is both the mean and variance ($\sqrt{\lambda}$ is the standard deviation).

Given the relationship between the Poisson and exponential distributions, we could use the exponential distribution to sample from the Poisson: sample repeatedly from the exponential distribution, until the sum of the samples is greater than the time period for the Poisson.

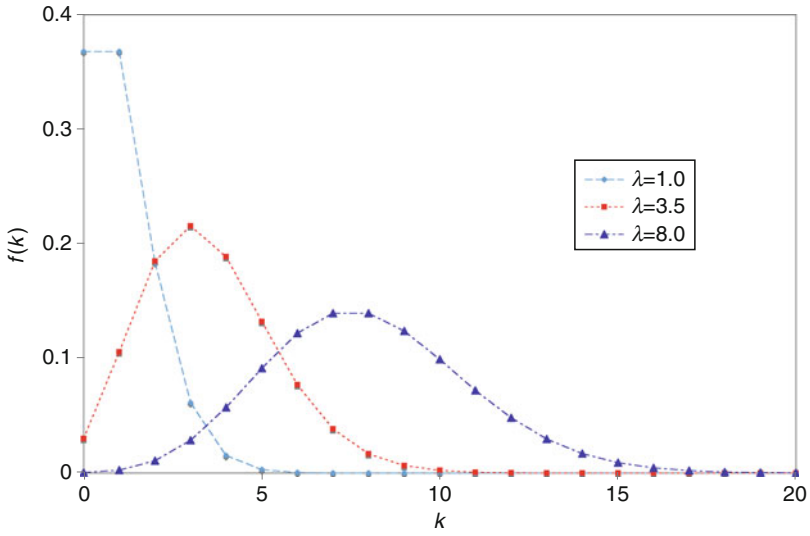


Fig. 6
Poisson PMF (the function only takes values at the integers – the lines are only a visual aid)

The value for the Poisson random variable is then the number of samples minus one (as the exponential sample that ended the sequence occurred outside the time interval). However, this algorithm is very slow unless λ is very small – the average number of exponential distribution samples needed is $\lambda + 1$. Once again, I recommend that you use a library function.

4 Basic Principles and Physics of Medical Imaging Simulation

Emission (PET and SPECT) and transmission (X-ray CT) tomography rely on detection of photons to populate the data arrays that are then reconstructed to create images. In the most general terms, simulation of these modalities must include (1) the source or sources of photons, (2) the interactions of these photons with matter, and (3) the conversion of photon–matter interactions into detected events. For each physical process at each step, the simulation needs a model of the process. This section will give an overview of the models used at each of these stages. The section ends with a discussion of acceleration methods.

In general, two kinds of models are used: deterministic models and stochastic models. Simulations that mainly make use of the former are often called analytic simulations; those that use more of the latter types of models are called particle- or photon-tracking simulations, or Monte Carlo simulations. In truth, almost all simulations include both deterministic and stochastic elements. Even users of analytic simulations want to see the effect of noise on their data; so, at a minimum, these simulations include the option of adding noise to the simulated measurements. At the other extreme, to include accurate stochastic models for every possibility in a simulation would be prohibitively expensive computationally: one must always reduce some aspects to deterministic computations (or ignore them completely). Realistic simulations of patient scans using photon-tracking simulations take considerably longer that the scans themselves

as it is! When designing a simulation – or choosing one to use – one must balance accuracy and efficiency. Where the balance is struck will depend on the problem or problems to be simulated.

A second goal of this section is to give the user a sense for this balance for the three main stages of simulation.

4.1 Sources of Photons

4.1.1 Nuclear Decay: Source of Photons for Emission Tomography

The primary source of photons in emission tomography is nuclear decay in the patient. The patient is injected with or ingests a tracer, a chemical labeled with a radioisotope. The chemical distributes through the body, possibly being metabolized along the way. As emission tomography scans typically last for 5–20 min, there is patient motion as well. The first question in a simulation of emission tomography is how much of this distribution process and patient motion should be simulated. The simulation can model stationary radioisotope distributions, or time-varying distributions. As a time-varying distribution can be viewed as a series of stationary distributions (there will always be some minimum time frame over which we are integrating), the balance to be stuck here is between ease of use for a particular application and software complexity.

There are two methods generally used to define the spatial domain of nuclear medicine simulations: decomposition into geometric primitives (e.g., cylinders, ellipsoids, boxes with a rule for which primitive takes precedence in case of intersections) or a pixelated volume. For a given time frame, the activity is uniform within a primitive or pixel. Generally the primitives are more efficient for relatively simple objects, pixels for very complex objects (◉ *Fig. 7*). This is true both in terms of computational efficiency and for the simulation user. When navigating through a pixelated volume, the simulation must update the object properties at each pixel change; it is much quicker to navigate through a few geometric primitives. Similarly, it is much easier for a user to specify a small number of cylinders, ellipsoids, etc., than to define a pixelated volume. However, for complicated objects like the human body it becomes much harder both to define the object as a collection of primitives and to navigate through it – the user often has a pixelated map of a patient, either from an imaging study or from something like the Visible Human Project® (National Library of Medicine 2003–2010); for the computer it is always clear what the next pixel will be in any direction, while to navigate through a collection of geometric

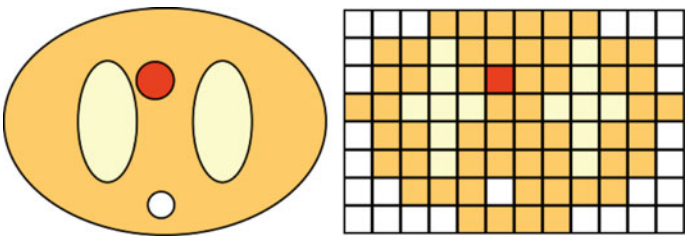


Fig. 7
Description of activity in the patient volume. The contents of a volume can be described as a collection of geometric primitives or can be pixelated

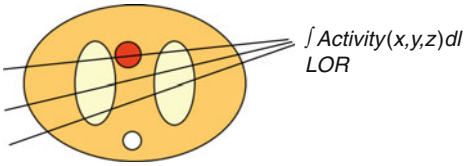


Fig. 8
Integral of Activity. An analytic simulation takes the integral of the activity for each LOR in the output data

primitives requires complicated algorithms that become less and less efficient as the number of primitives grows larger. Some simulations allow or require both types of description, for instance using pixelation for the patient volume and geometric primitives for the tomograph.

Once we have a distribution of radioisotope, we need to decide the simulated time and position of each decay. It is at this point that typical analytic and photon-tracking simulations diverge: the analytic simulations ignore individual decays, and instead integrate the activity distribution along each detection line of response (LOR) to get an approximate average response (Fig. 8):

$$\int_{\text{LOR}} A(x, y, z)\, \mathrm{d}l, \tag{14}$$

where A is the activity and the integral is along the LOR.

Photon-tracking simulations, on the other hand, need to generate one decay at a time. There are several ways to do this, but one typical way is to step through the pixels or primitives in the patient volume and

- Calculate the mean number of decays for the current time frame (the product of the activity in becquerels and the time in seconds).
- Sample a Poisson random variable with this mean to get the number of decays in the pixel/primitive.
- For each decay:
  - Randomly pick a point within the pixel or primitive for the decay.
  - Randomly pick a time within the time frame from a truncated exponential distribution with λ the half-life of the radioisotope (if the half-life is long compared to the time frame, the uniform distribution can be used).

There are many possible decay modes, but the three occurring in common medical radioisotopes are alpha decay (emission of a helium nuclei), negative or positive beta decay (emission of an electron or positron plus a neutrino), and gamma decay (emission of a photon) (Fig. 9). Alpha and negative beta decays have been of little interest in emission tomography simulations, as the particles do not travel very far from the decay and no detectable photons are produced. This may be changing as there is increasing interest in the dosimetry (the study of the radioactive dose absorbed by the patient) of tomographic imaging. In isotopes with alpha or beta decays, most of the dose comes from the kinetic energy of these particles.

Gamma decay is the most straightforward to simulate: the decay emits a photon, the energy of which is determined by the radioisotope. These are the photons we detect in SPECT. Some isotopes have a single decay mode, always producing the same energy photons. Others have

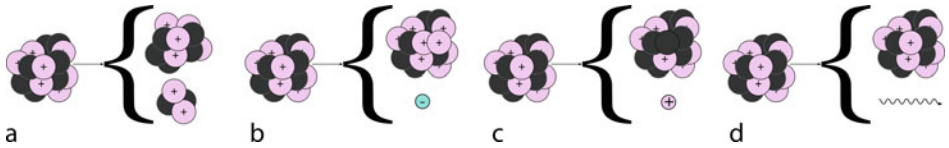


Fig. 9
Decay Modes. (a) An alpha decay ejects a helium nucleus. (b) A beta-negative decay ejects an electron. (c) A beta-positive decay ejects a positron. (d) A gamma decay ejects a photon

multiple decay modes, producing photons with different energies and different probabilities, and still others go through a chain of decays producing several photons of different energies (or alpha or beta decays). In the latter case there is a time correlation between the various decay products; often the half-lives of the intermediate decay products are so short that the decays are, for the purposes of a tomographic simulation, simultaneous. Not all simulations support isotopes with multiple decay products.

Once a photon has been created at a given location, its direction of travel must be randomly selected. To do this, we pick a random unit direction vector $\vec{v}$:

– Sample θ from the uniform distribution on $[0, 2\pi)$.
– Sample μ from the uniform distribution on $[-1, 1]$.
– Let $\phi = \cos^{-1}(\mu)$.
– Set $\vec{v} = (\cos(\theta)\sin(\phi), \sin(\theta)\sin(\phi), \cos(\phi))$.

Positron decay is the process of interest for PET (see Chap. 38, "PET Imaging: Basics and New Trends"). The positron travels a short distance before annihilating with an electron and producing two antiparallel photons (traveling in the opposite direction – also called anti-collinear). The positron's travel can be simulated, but to do so is very computationally expensive. Most simulations either ignore the travel distance and assume the decay and annihilation locations are identical, or they use a simplified stochastic model to select the annihilation location (e.g., Palmer and Brownell 1992). Two 511 keV photons are created by the annihilation. One photon's direction should be randomly chosen as above, the other should be started in the opposite direction. Annihilation photons are not exactly antiparallel – the angle between the two photon directions is $180^\circ + \eta$, where η is a normal random variable with mean 0° and standard deviation approximately 0.5° (Evans 1955). Many simulations include this.

4.1.2 Electromagnetic Radiation: Source of Photons for X-ray CT

In general, medical X-rays are generated by accelerating electrons to a high velocity and having them collide with a metal target. Ideally this results in a point source of X-rays that are then collimated to eliminate those that will not hit the opposing detector array. While a wide range of photon energies are used, a typical X-ray CT uses photon energies around 80–120 keV. The beam is not monoenergetic, but a continuous distribution of energies over the entire range.

It is possible, in some packages, to include the X-ray generation and source collimation, but to do so is extremely computationally expensive. CT simulations do not generally include it, but instead assume beams of photons from a point source with an idealized distribution of energies,

either monoenergetic or an experimentally derived or simulated distribution. Simulations often give the user a choice of the energies to simulate, as different imaging situations sometimes call for different beam energies – for instance a lower energy might be used for a child. The energy is usually expressed as the voltage on the X-ray tube in kV. This will set the maximum possible energy for the resulting photons: if the voltage is set to 100 kV the maximum photon energy will be 100 keV.

4.1.3 Secondary Sources of Photons

There are several secondary sources of radiation that are generally considered contaminants: K- and L-shell fluorescence, also known as lead X-rays, and bremsstrahlung (braking radiation) (see Chap. 1, "Interactions of Particles and Radiation with Matter"); some detector materials are radioactive, most notably the lutetium used in LSO, a common detector material for PET, and also, when imaging radioisotopes with complex decay schemes, photons from decay products other than that being imaged.

K- and L-shell fluorescence occurs when an electron from a higher energy shell falls into an unoccupied position in a lower energy shell. A photon is emitted with the energy difference. As X-rays and gamma rays knock electrons out of orbit when they Compton scatter or are photoelectrically absorbed, they can cause such fluorescence. The probability of fluorescence increases with photon energy and with the atomic number (Z) of the material the photon is traversing. For the photon energies seen in medical imaging, fluorescence occurs almost exclusively in high-Z materials, mainly occurring in the collimator and shielding of the tomograph. For X-ray CT and thallium SPECT scans, fluorescence X-rays can be mistakenly identified as events. For other scans they can add to dead-time problems.

Bremsstrahlung occurs when a charged particle decelerates in an electric field. The particle's lost momentum is emitted as a photon. In medical imaging this typically occurs after a beta decay. However, the probability of bremsstrahlung is very low in low-Z materials: as almost all beta decays are absorbed in the patient, bremsstrahlung does not have much effect on scans.

Decay in LSO detectors has been shown (using simulation!) to have a small but noticeable effect on random coincidence rates and dead time in PET.

There are many isotopes used in medical imaging that have decay products other than those used for imaging. The biggest problems occur when higher-energy photons are created, as these can often penetrate through the collimation and deposit energy in the photopeak window. Lower-energy photons can cause dead-time problems. Both these issues have been explored with simulation.

4.2 Tracking Photons Through Matter

X-rays and gamma rays are used for medical imaging because the human body is, to an extent, transparent to them. However, there is considerable attenuation: even a 511 keV photon has a mean free path of only 10.4 cm in water. In SPECT, PET, and X-ray CT, many more photons will interact in the patient than will reach the detector without interactions. Simulations must account for this attenuation: it often makes the biggest contribution to the structure of noise in emission tomography, and in transmission tomography it is the measurement of interest. It is often of interest to model the interactions as well – this is the purpose of photon tracking in

simulations. The main photon–matter interactions that occur in medical imaging are photoelectric absorption, Compton scatter, and coherent (or Rayleigh) scatter. There are a few isotopes that emit photons with energies > 1,022 keV as decay products; for these, pair production can also occur.

Descriptions of these processes can be found in Chap. 1, "Interactions of Particles and Radiation with Matter." This section will focus on the mechanics of simulating them.

4.2.1 Description of Attenuation

A simulation needs a description of the attenuating media that photons (and other particles, if tracked) will pass through. The discussion of how emission tomography simulations represent the patient (Sect. 4.1.1) could be repeated here – for the representation of the attenuation in the patient the issues are similar. However, if the tomograph is going to be simulated in any detail it cannot be reasonably represented as a pixelated object. Some of the materials used in the tomographs, such as lead and detector materials, are too attenuating to approximate with pixels. Small differences in shape can make rather large differences in photon transmission/absorption. An extremely large number of pixels would be required to do an adequate job of defining the tomograph elements; tracking through so many pixels would be computationally expensive. Fortunately most tomograph elements can be well approximated using geometric primitives. For this reason, simulations often use a pixelated representation of the patient and geometric primitives to represent the tomograph (*Fig. 10*).

Attenuation characteristics depend on the elemental composition and density of a material and vary with photon energy. A good source for attenuation coefficients is Cullen et al. (2010). Many simulations provide the user with a set of materials sufficient for most simulations, including, e.g., a selection of tissue types for the body, lead and tungsten for the collimator, and the most common detector materials. The simulation then uses pre-computed tables giving the properties for the materials at a range of energies. In analytic simulations, where contaminants

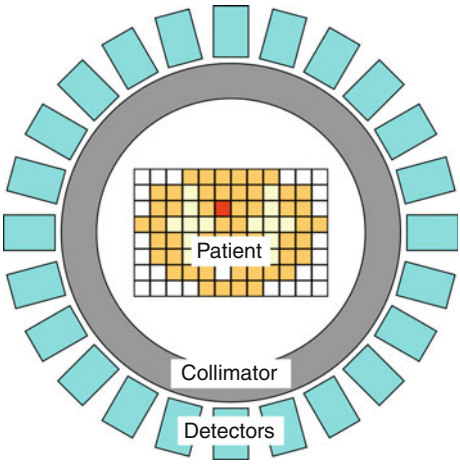


Fig. 10
PET Simulation setup with pixelated object and geometric primitives for the tomograph

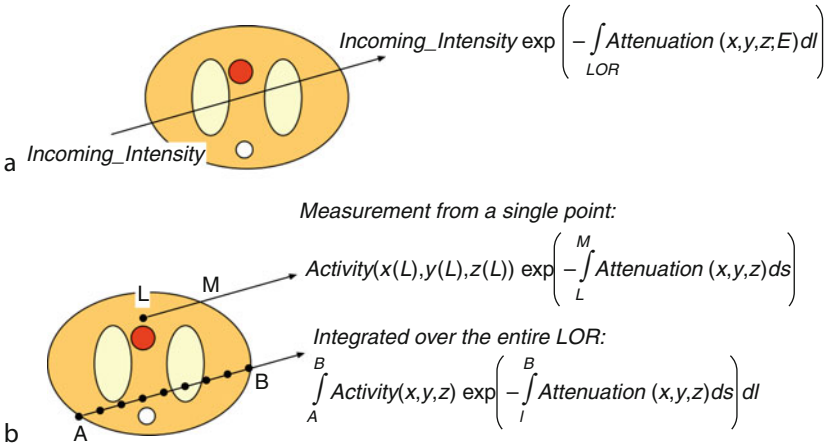


Fig. 11
Analytic Simulation of X-ray CT and SPECT

like scatter are ignored or modeled simply, the variation with energy is often ignored. For X-ray CT, a measurement will be

$$I_0 \exp\left(-\int_{\mathrm{LOR}} \mu(x,y,z;E)\,\mathrm{d}l\right), \tag{15}$$

where I_0 gives the incoming intensity of the beam leaving the X-ray tube, μ gives the attenuation coefficient at the position (x,y,z) for the energy E, and the integral is over the LOR that is being measured (Fig. 11a). As μ will change with energy, this value must either be weighted and integrated over the range of energies (with I_0 a function of energy), or taken at an average energy.

Attenuation also affects measurements in emission tomography. Equation 14 must be adjusted for attenuation. In PET, between the two antiparallel photons the attenuation of the entire LOR is encountered. The equation for the mean measurement on a LOR becomes a combination between Eq. 14 and Eq. 15, with Eq. 14 taking the place of I_0 and the energy for the attenuation being fixed at the attenuation of the annihilation photons,

$$\int_{\mathrm{LOR}} A(x,y,z)\,\mathrm{d}l \exp\left(-\int_{\mathrm{LOR}} \mu(x,y,z;511\,\mathrm{keV})\,\mathrm{d}l\right). \tag{16}$$

In SPECT the photons do not travel the whole line of response, but rather the portion of the line of response from the point of decay to the detector. Thus we must pay more attention to the beginning and end points of the attenuation integral; they come inside the integral for the attenuation:

$$\int_{-\infty}^{\infty} A(x,y,z) \exp\left(-\int_{l}^{\infty} \mu(x,y,z;E)\,\mathrm{d}s\right)\mathrm{d}l, \tag{17}$$

where A is the map of activity, μ is the map of attenuation, E is the energy of the photons, and $+/-\infty$ is used to indicate that we are integrating over the entire LOR (Fig. 11b).

The integrals above (Eqs. 15–17) are what the imaging modalities want to measure: they neglect the factors contaminating the data. One of the chief sources of contamination is the photons that are attenuated. These photons undergo an interaction somewhere along the LOR. Most of them scatter, and some of those end up being detected, for the most part on another LOR.

4.2.2 Where Will a Photon Interact?

If we have a photon at position (x, y, z) traveling in direction (u, v, w) with energy E, where will it interact? For photon-tracking simulations this is often the most computationally expensive calculation – not because it takes so long to calculate one time, but because it needs to be calculated so many times.

As mentioned in Sect. 3.2, the exponential distribution (Eq. 7) gives the distribution of distances a photon travels before an interaction in a uniform medium. We can sample from this distribution using Eq. 19, but we are not tracking through a uniform medium. We circumvent this problem by sampling the number of free paths to travel, p, from a medium-independent distribution,

$$p = -\ln(u), \tag{18}$$

where u is sampled from the uniform distribution on $(0, 1)$. Note that this is the same as Eq. 9b except that we have not divided it by the attenuation coefficient: if we were traveling through a uniform medium, we would divide by the attenuation coefficient to get the distance to travel. To adjust this to nonuniform media, we travel until

$$\sum_{i=1}^{D} \mu_i d_i = p, \tag{19}$$

where d_i is the length of the photon path across the ith material segment and μ_i is the attenuation for $i < D$ at the photon's E. We pick D and d_i for $i = D$ so that the equality in Eq. 19 holds. If the photon leaves the space defined for our simulation before the sum reaches p, we say it has escaped; otherwise we simulate an interaction at the position

$$(x, y, z) + \left(\sum_{i=1}^{D} d_i\right)(u, v, w). \tag{20}$$

4.2.3 What Type of Interaction?

The attenuation coefficient, μ, can be broken down into four components (*Tables 1* and 2): the photoelectric component, μ_{pe}; the Compton scatter component, μ_c; the coherent (or Rayleigh) scatter component, μ_r; and, if the photon energy is over 1,022 keV, the pair production component, μ_{pp}:

$$\mu = \mu_{pe} + \mu_c + \mu_r + \mu_{pp}. \tag{21}$$

To sample which type of interaction to simulate we

– Sample u from the uniform distribution on $[0, 1)$.
– Simulate photoelectric absorption if $0 \le u < \frac{\mu_{pe}}{\mu}$.
– Simulate Compton scatter if $\frac{\mu_{pe}}{\mu} \le u < \frac{\mu_{pe} + \mu_c}{\mu}$.

Table 1

Approximate likelihood of photon interaction types in water and lead as a function of energy

	Water			Lead		
Photon energy (keV)	50	500	5000	50	500	5000
Photoelectric absorption	12%	<0.1%	<0.1%	91%	51%	3%
Compton scatter	79%	99.7%	92%	1%	42%	46%
Coherent scatter	9%	0.2%	<0.1%	8%	7%	<0.1%
Pair production	0%	0%	8%	0%	0%	51%

Table 2

Variation of attenuation components with photon energy (*E*) and attenuating material's atomic number (*Z*)

Interaction type	Variation with *E*	Variation with *Z*
Photoelectric absorption	$\propto \frac{1}{E^3}$	$\propto Z^3$
Compton scatter	~ constant $E \leq 100$ keV $\propto \frac{1}{E}$ $E > 100$ keV	$\propto Z$
Coherent scatter	$\propto \frac{1}{E}$	$\propto Z^2$
Pair production	Rapid increase above 1022 keV	$\propto Z^2$

– Simulate coherent scatter if $\frac{\mu_{pe} + \mu_c}{\mu} \leq u < \frac{\mu_{pe} + \mu_c + \mu_r}{\mu}$.

– Otherwise simulate pair production.

4.2.4 Simulating Photoelectric Absorption

When a photon is photoelectrically absorbed it disappears: we do not track it any further. However, the energy of the photon cannot disappear. In biologic tissue the energy will dissipate over a very small range and is of interest only when including dosimetry in a simulation. In collimators the energy may be re-emitted as a lead X-ray. In scintillating detectors the energy will be re-emitted as many scintillation photons (or as a "lead" X-ray, though in most crystals this will be reabsorbed quickly and create scintillation photons).

4.2.5 Simulating Compton Scatter

Compton scatter is an interaction between the photon and an atomic electron. The photon changes direction and loses energy in the collision. Compton scatter is the dominant interaction type for medical X-rays and gamma rays in biologic tissues. At typical emission tomography photon energies, the scatter angles are noticeably forward-peaked (i.e., the likeliest scatter angles are around 0°), but there is a sizeable probability of scatter into any angle. The outgoing energy and the change in angle are related by the equation

$$\frac{E_{\text{out}}}{511} = \frac{\frac{E_{\text{in}}}{511}}{1 + \frac{E_{\text{in}}}{511}(1 - \cos\alpha)}, \tag{22}$$

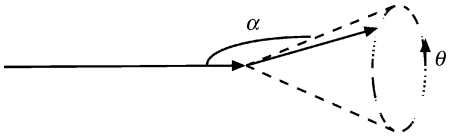


Fig. 12
The scatter angle defines a cone of possible outgoing directions. A second angle is sampled to choose from amongst them

where E_{in} and E_{out} are the incoming and outgoing energies of the photon in keV and α is the angle between the incoming and outgoing photon directions.

α can be sampled from the Klein–Nishina distribution, a free electron approximation of the Compton scatter distribution (Zerby 1963). In the Klein–Nishina distribution, the likelihood of scattering at α is proportional to

$$\left(\frac{E_{\text{out}}}{E_{\text{in}}}\right)^2 \left(\frac{E_{\text{in}}}{E_{\text{out}}} + \frac{E_{\text{out}}}{E_{\text{in}}} - 1 + \cos\alpha\right). \tag{23}$$

Kahn (1954) developed an oft-used acceptance–rejection method for sampling from the Klein–Nishina distribution.

The Klein–Nishina distribution does not take into account the binding energy of the electron. This becomes a significant factor for low-energy photons in high-Z materials and requires an adjustment as described in Ljungberg (1998). Alternatively, the distribution can use lookup tables based on one of the high-energy physics databases (e.g., Cullen et al. 2010). These reflect the latest experimental data.

α is just the angle between the photon's incoming and outgoing direction vectors. This defines a cone of possible outgoing directions: a second angle should be chosen from the uniform distribution on $[0, 2\pi)$ to sample from this cone (*Fig. 12*).

The energy the photon loses by Compton scattering is transferred to the electron. This energy manifests itself in the same ways as discussed in Sect. 4.2.4.

4.2.6 Simulating Coherent Scatter

Coherent scatter is an interaction between a photon and an entire atom. The photon changes direction, but does not lose any energy. At typical emission tomography photon energies, the scatter angles are extremely forward-peaked, with an extremely small probability of scatter of more than 10°. Coherent scatter occurs predominantly at low energies in high-Z materials.

At low photon energies in an element with atomic number Z, the differential cross section per atom is given by the incoherent sum of the Thomson cross section over the Z electrons in the atom (Strutt 1871) and the likelihood of scattering at α is proportional to

$$Z(1 + \cos^2\alpha). \tag{24}$$

However, as photon energy and Z increase, the photon's wavelength approaches the atomic radius resulting in interference in the scattering amplitudes across the atom, which is accounted for using atomic form factors. Photoelectric absorption also alters Eq. 24, with analytic approximations of the effect failing to account for experimental measurements at high energies in high-Z materials (Kaplan et al. 1998).

4.2.7 Simulating Pair Production

Pair production occurs when a photon with energy greater than 1,022 keV "becomes" a positron–electron pair. (The minimum energy is due to $E = mc^2$. The creation of other antimatter–matter pairs is also possible, but the energies required are well outside those seen in medical imaging.) At energies seen in medical imaging, the interaction occurs only in the presence of an atomic nucleus (which is needed to balance the momentum), and the likelihood increases approximately with Z^2 and quite quickly with photon energy. Pair production is not important in biologic tissues (the tissues are nearly transparent at these energies and Compton scatter is still dominant – see ❯ *Table 1*), but in high-Z materials like lead and detector materials its frequency approaches that of Compton scatter for some of the highest-energy gamma rays (2–4 MeV).

When pair production occurs, the excess photon energy above 1,022 keV is shared (not necessarily equally) between the electron and positron as kinetic energy. The positron, after losing most of its kinetic energy, will annihilate with an electron producing two 511 keV photons; this process can be simulated as it is after a positron decay. The electron will also lose its kinetic energy, with the same effects as those discussed in the section of photoelectric absorption.

4.3 Simulating Detection

Many different methods are used to detect photons in medical tomography. However, most of the systems use scintillating crystals that convert an X-ray or gamma ray into a flood of visible light photons, and then a solid-state device or photomultiplier tube (PMT) to convert the scintillation photons into electric signals. These signals are then converted to a measured response on a LOR.

These processes can be simulated, but are usually not simulated as part of an analytic or photon-tracking simulation. The added computational expense of, e.g., tracking hundreds of scintillation photons is too great. Instead, some parts of the detector response are modeled based on experimental or simulated measurements. In many analytic simulations it is ignored completely; the detectors are treated as if they are perfect absorbers with perfect energy resolution.

4.3.1 Tracking Photons in the Detector Crystal

The physics of tracking photons through detectors is no different than that described above, but our goal is different: We keep track of where energy is deposited in the detector, as it is this energy that gets converted to scintillation photons.

As in tracking through the patient, some simulations ignore the kinetic energy imparted to electrons by absorption and Compton scatter, assuming that this energy is deposited locally. This is a relatively good approximation: electrons (and K- and L-shell fluorescence photons) have a short range in detector crystals, so most will deposit their energy in the detector element they were created in.

4.3.2 Converting Deposited Energy to a Signal

Emission and transmission tomography handle detection differently. In transmission tomography, the signals from many photons are summed together and the summed signal used as a measurement of the transmitted intensity of the incident beam (see Eq. 15); emission tomography attempts to process each photon separately and count single events or coincidences.

Many transmission tomography simulations do little simulation of the detectors. They take the integral in Eq. 15, sometimes at several different energies to account for the energy distribution of the incident beam. More detailed photon-tracking simulations are usually done with in-house software or general-purpose Monte Carlo packages (Zaidi and Ay 2007). Some packages allow full simulation from decay, through scintillation photons, to the electronics, but it is not generally practical to include so much detail. Instead, a detailed simulation or experiment might be done to quantify the response of the detector and electronics to photons of various energies or to different fluxes and the results tabulated. Equation 15 can then be modified to use this data.

In emission tomography the main decisions involve the handling of deposited energy, spatial location, and time.

The detected energy of photons is used as a basis to accept or reject them: events with energy too low are assumed to be scattered photons, events with energy too high are assumed to be the piled-up energy from two or more photons. In SPECT, events with too low energy are sometimes binned into secondary data arrays used to estimate scatter contamination. There is some error in the detected energy: some simulations ignore this, but many add a normally distributed error term to the total deposited energy $E_{\text{deposited}}$,

$$E_{\text{detected}} = E_{\text{deposited}} + n, \tag{25a}$$

where E_{detected} is the detected energy, and n is a random sample from a normal distribution with mean 0, standard deviation $\frac{F\sqrt{E_{\text{decay}}E_{\text{deposited}}}}{2\sqrt{2\ln 2}}$ for F the full width half maximum (FWHM, the measure usually given by camera manufacturers) of the detector system's energy response, and E_{decay} the energy for which the FWHM was measured. Alternatively a model making the error term a function of the photon's interaction positions and energy deposited at each can be developed using experimental or simulated measurements, i.e.,

$$E_{\text{detected}} = \sum_i (E_{i,\text{deposited}} + n_i), \tag{25b}$$

where the summation is over the interactions in the detector crystal, $E_{i,\text{deposited}}$ is the energy deposited at interaction i, and n_i is a noise term possibly dependent on both the energy and position of the ith interaction.

Determining the detected position of a photon is often done in conjunction with determining the energy, as the interaction positions and energy deposited at each are needed for both calculations. In many SPECT simulations the photon is positioned at the energy-weighted centroid, $\vec{x}_c$, of the interactions in the detector crystal

$$\vec{x}_c = \frac{\sum_i E_i \vec{x}_i}{\sum_i E_i}, \tag{26}$$

where the summation is over the interactions in the detector crystal, the $\vec{x}_i$ are the two-dimensional position vectors of the interactions (the depth of the interaction in the crystal

is ignored), and the E_i are the energy deposited by the interactions. This approximation misses the edge effects in the crystal, the effect of the PMT positioning on resolution, and the effect of pileup: such effects can be included in a model based on experimental or simulated data.

There are also many time-related effects in emission tomography: pileup and dead time in the detectors, and in PET coincidence timing, random coincidences, and time-of-flight measurements. Dead time and random coincidences can be modeled by keeping track of time during the simulation and modeling the loss of events (for dead time) and coincidence window (for random coincidences) directly, or by keeping track of the singles rates in the detectors. Typical dead-time models are paralyzable,

$$r_m = re^{-\tau r}, \tag{27}$$

or non-paralyzable

$$r_m = \frac{r}{1+\tau r}, \tag{28}$$

where r and r_m are the actual and measured count rates, respectively, and τ is the dead-time coefficient.

Random coincidences are typically modeled as using the singles rates at each detector element, S_i, to determine the random coincidence rate along the LOR between two detectors, R_{ij},

$$R_{ij} = 2\tau S_i S_j, \tag{30}$$

where τ is the length of the coincidence window. This model does not include the effect of triples (three photons arriving in a time window) and it includes photons involved in coincidences as singles as well, but these are typically small effects. More importantly, this formula does not account for isotopes with decay chains that produce multiple photons over a short (compared to the coincidence window) time period. For such isotopes, simulations must model randoms by keeping track of time on a photon-by-photon basis.

Pileup is more complicated to model, as it is caused by two or more photons interacting in a detector unit at almost the same time, but affects the spatial positioning of the photons. Again, correct modeling of pileup requires keeping track of time on a photon-by-photon basis, but also an understanding of how the tomograph being modeled handles the signals electronically: in modern cameras several methods are used to minimize pileup (Wernick and Aarsvold 2004).

Simulating time-of-flight measurements requires keeping track of time only for the photons from a single annihilation/decay. Photons from different decays can be assumed to be uncorrelated in time, and thus randomly distributed along the time-of-flight axis. For instance, random events can be modeled with ❯ Eq. 30, then assigned a random location in the time-of-flight dimension.

4.3.3 Histogramming Simulated Events

One advantage of simulations is that they give the user the ability to see the history of an event to a much greater extent than one can experimentally. In general, users want to be able to categorize events by the various aspects of their history. For example, instead of getting sinograms that mix unscattered and scattered (and random) events together, simulations give the user the

Table 3

Potential fields for histogramming (some fields may be needed once for each photon in PET)

Data	Comment	Typical number of bins when used
Sinogram or projection data	Output data type for the scanner simulated	10^7–10^9
Time-of-flight position	For time-of-flight PET; really part of above	16–256
Starting location	Starting location of photon/decay (may want separate decay/annihilation location for PET)	Approximately equal to sinogram/projection data
Bed position	When patient bed moves during scan	1–20
Starting energy	Energy of photon at creation	5–50 when needed
Energy entering detector	Energy of photon entering detector	2–50
Detected energy	Energy detected by detection system	2–50
Number of scatters	How many times photon scattered (may be split into patient/collimator/detector components)	2–10 (often desire just scattered/unscattered; detected photons rarely have scattered more than 10 times)
Scatter location(s)	Where the photon scattered in patient or collimator	Approximately equal to starting location data
Event type	E.g., alpha/beta/gamma decay, lead X-ray, bremsstrahlung, pileup, random coincidence, pair production	Number of event types desired
Location of interaction in detector	The locations used in Eq. 25b and 26	100–10^7 per interaction
Deposited energy of interaction	The energies used in Eq. 25b and 26	50–1,000 per interaction
Time of decay	May need to be kept to sub-nanosecond accuracy while processing, but typically only for, e.g., kinetic modeling/patient motion in output array	10–200
Time of detection	Same as above	Same as above

opportunity to histogram the data into separate unscattered and scattered (and random) sinograms. This histogramming process is also called scoring, sorting, or binning in some software. Some of the fields of interest are listed in *Table 3* along with a typical number of bins needed. Ideally the output can be created with any combination of fields and arbitrarily many dimensions and number of histogram bins. For example, one might want to create a histogram for coincidences from PET simulation with the following dimensions:

$$(x_r, y_r, \phi, \theta, E_1, E_2, s_1, s_2, x, y, z), \tag{31}$$

where (x_r, y_r, ϕ, θ) give the projection data format used for reconstruction of 3D PET data, (E_1, E_2) give the detected energies of the two photons, (s_1, s_2) give the number of scatters each of the photons underwent, and (x, y, z) give the decay position for the event.

Note, a simulation with this output histogram would be completely impractical as the number of bins in the histogram is product of the number of bins used for each of the variables: typical projection data for a modern PET scanner is already on the order of 100 million bins, multiplying this by any reasonable number of bins for the other fields would result in an array too large to fit even on modern disk drives; furthermore, to generate enough events to get a good estimate of the mean for every bin using Monte Carlo simulation, which requires many thousands of operations to produce a single event, would take far too long even using a supercomputer. Note that this histogram does not even include many fields of interest; however, careful experimental design can help to disentangle factors that cannot be directly simulated.

I know of no program that directly supports all the fields in ❯ *Table 3*, though it may be possible to modify some to handle any of them or to post-process the data to derive the desired data.

4.4 Acceleration of Photon-Tracking Simulations

Photon- and particle-tracking simulations can take weeks, months, or even years to simulate clinical scans on a single CPU. Thus there is much interest in accelerating them. Many simulations include variance reduction techniques, fictitious interaction tracking, or convolution-forced detection.

Variance reduction techniques, also known as importance sampling, are techniques that reduce the variance of estimates without adding any bias (Haynor et al. 1991; Haynor 1998). The key to variance reduction is recognizing that many of the photons that we track do not contribute to the measurements we are interested in, for instance a photon that scatters and leaves the tomograph. Unfortunately we do not know in advance which photons these will be – a photon that appears to be leaving the tomograph can scatter again and be detected. We can, however, estimate how likely it is a given decay or photon will be detected. In variance reduction, we simulate more decays or photons that we think are likely to be detected. To avoid bias, we give each decay or photon a weight and increment our histogram by this weight rather than just counting detections. When we over-/under-sample a given set of decay or photons, we reduce/increase the weight by the same factor that we over-/under-sampled by. The two main variance reduction techniques are called stratified sampling and forced detection. In stratified sampling we choose decay positions and initial photon directions in proportion to the likelihood that they will be detected, i.e., in emission tomography more decays are sampled inside the tomograph field of view than outside and we will oversample initial photon directions that are headed toward a detector. In forced detection, when a photon interacts in the patient we force the photon to scatter rather than be absorbed; force the scatter to be in a detectable direction; and force the photon to pass through the remaining attenuation to the detector.

Fictitious interaction tracking, also known as delta tracking or scattering and Woodcock tracking, is another unbiased acceleration technique (Rehfeld et al. 2009). It is a method for accelerating tracking through objects with many changes in attenuation, for instance pixelated phantoms. With fictitious interaction tracking it is no longer necessary to step through every pixel on a photon's path through the object.

Convolution-forced detection is used for estimating scatter distributions quickly (de Jong et al. 2001). It is similar to forced detection, but replaces the scatter with an estimate of the tomograph's spatial response. It does bias the output histograms, but properly implemented this bias can be small.

5 Available Simulation Software

There are many simulation software packages appropriate for emission and transmission tomography that are available in the public domain, as open source software, or semi-publicly (e.g., with the author's permission). Buvat and Castiglioni (2002) give a good overview of those available for PET and SPECT, and Zaidi and Ay (2007) a good overview for X-ray CT (including in-house software not available to the public). While some newer/overlooked packages are available, these articles remain excellent starting points for those seeking information about such software.

In general, the software can be split into three categories: general-purpose particle-tracking software, in general produced by one of the large high-energy physics laboratories; photon-tracking software specifically designed for emission or transmission tomography; or analytic simulation software specifically designed for emission or transmission tomography. The general-purpose packages are capable of the most accurate simulations and are the most flexible, but tend to be the slowest and most difficult to set up and use. The photon-tracking software designed for emission and transmission tomography was, for the most part, developed in response to these shortcomings. By restricting the simulations to tomographs and restricting the tracking to those physical effects that have a measurable impact on medical imaging, the simulations can be faster (less to simulate and knowledge of the geometry allows for some optimization) and easier to use (many decisions are already made by the software designer). Analytic simulation software is the least accurate and flexible, but extremely fast and easy to use. These characteristics are summarized in *Fig. 13*.

General-purpose particle-tracking software includes Geant (Allison et al. 2006), EGS (Nelson et al. 1985; Kawrakow 2000), MCNP (Forster et al. 2004), and Penelope (Sempau et al. 2003). These packages can provide extremely accurate simulations of photon, electron, and positron transport through arbitrarily complex structures. They allow the user to tailor the complexity of the tracking to some extent – e.g., by leaving out bremsstrahlung radiation or by ignoring the tracking of optical photons. Realistic simulations of a clinical imaging scan may take months or years on a single CPU, and the software can be difficult to adapt to emission and transmission tomography simulation. There are several extensions that significantly reduce the ease of use problems: GATE (Geant4 Application for Emission Tomography; Jan et al. 2004), PET-EGS (Castiglioni et al. 1999), and PenelоPET (España et al. 2009). GATE is the most flexible of these, able to simulate PET, SPECT, and X-ray CT. The other two are PET-specific.

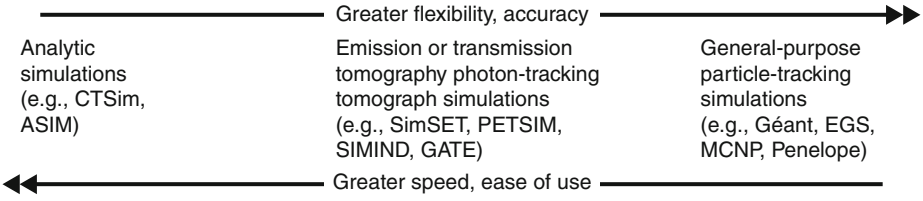


Fig. 13
In simulations, greater flexibility and accuracy comes at the expense of speed and ease of use

There are a number of photon-tracking packages that have been developed specifically to provide greater speed and/or ease of use than the general-purpose packages. Two commonly used packages are aimed at emission tomography: SimSET (Simulation System for Emission Tomography; Lewellen et al. 1998) simulates both PET and SPECT, and SIMIND (Dewaraja et al. 2002) is a fully featured SPECT-only simulation. Buvat and Castiglioni (2002) give details of several other publicly available emission tomography simulations. In general, these simulations are a factor of 10–100 faster than the general-purpose software.

Zaidi (2007) lists no publicly available photon-tracking software specifically tailored to transmission tomography. However, it appears that a package called SINDBAD may soon be available (Tabary et al. 2007).

There is an open source analytic simulation for transmission tomography, CTSim (Rosenberg 2002). I know of no open source analytic simulations for emission tomography, though our group is hoping to release an open source version of ASIM (Comtat et al. 1999) soon. Analytic simulations can be extremely fast, particularly in situations where many realizations of the same, or close to the same, situation are needed.

6 Choosing a Simulation Tool

Where possible, one should use an already existing simulation rather than writing one's own. In the following section we list some readily available packages; often other packages found in the literature can be acquired by contacting the author.

In general, choosing the fastest/easiest-to-use simulation appropriate for one's application is a good idea. Analytic simulations are often used for testing image reconstruction software and evaluating clinical protocols. Often noise and reconstructed image resolution, which analytic simulation reproduces reasonably accurately, have the most impact. And it should be noted that clinically there is a lot of noise that a physics simulation cannot account for in any case. In PET, for instance, the test-retest variability for patient standard uptake value measurements is ~10% with much of the variation being biological.

When more quantitative accuracy is needed, using a photon-tracking or general-purpose simulation may be required. For instance, when developing corrections for effects like scatter or beam hardening, or when examining the effects of small differences in tomograph design, these simulations would be called for. One should consider what level of detail is appropriate for a given simulation: a complete description of a particular tomograph might not be appropriate when a general question is being asked, a photon-tracking package may suffice. On the other hand, when investigating a new tomograph geometry, one may be forced to use a general-purpose package: the emission and transmission tomography simulations only simulate certain defined geometries. The general-purpose packages are also required when examining the fine details of detector/electronics performance, e.g., when designing new detectors/detector arrays. Here the problem can often be reduced to looking at a single detector or detector block, but tracking of scintillation photons may be necessary (Van der Laan et al. 2010) along with simulation of the electronics. Buvat et al. (2005) give a good overview of the features one might use to decide between photon- and particle-tracking simulations.

Finally, an alternative for some PET researchers might be a database of data previously simulated using the PET-SORTEO simulator (Reilhac et al. 2004; Reilhac et al. 2005). It provides simulated data for FDG, Raclopride, and Fdopa brain studies.

Table 4
Web addresses for simulation tools

Tool name	Address
Simulation packages	
Geant	http://geant4.web.cern.ch/geant4
EGSnrc	http://irs.inms.nrc.ca/software/egsnrc
EGS4	http://www.irs.inms.nrc.ca/EGS4/get_egs4.html
MCNP	http://mcnp-green.lanl.gov/index.html
GATE	http://www.opengatecollaboration.org
SimSET	http://depts.washington.edu/simset/html/simset_main.html
SIMIND	http://www.radfys.lu.se/simind
CTSim	http://ctsim.org
Simulated PET data	
PET-SORTEO	http://sorteo.cermep.fr
Digital anatomic phantoms	
NCAT phantom	http://www.bme.unc.edu/~wsegars/
Zubal phantom	http://noodle.med.yale.edu/zubal/data.htm
Visible Human Project®	http://www.nlm.nih.gov/research/visible/visible_human.html

7 Online Resources

Table 4 gives current Web addresses of some simulation tools.

8 Conclusion

Simulation of emission and transmission tomography is a rich and well-developed field. There are many approaches to simulation, but the underlying physics is uniform. To choose an approach one must first decide what level of accuracy/detail is needed in the simulated data. In general, greater accuracy requires more time setting up and configuring a simulation and results in slower simulations. There is a wealth of publicly available simulation tools that will serve most needs.

9 Cross-References

Chapter 1, "Interactions of Particles and Radiation with Matter"
Chapter 5, "Statistics"
Chapter 10, "Radiation Protection"

Acknowledgments

This work was supported in part by PHS grants CA42593 and CA126593.

References

Allison J et al (2006) Geant4 developments and applications. IEEE Trans Nucl Sci 53: 270–278

Burger EB, Starbird MP (2005) The heart of mathematics: an invitation to effective thinking. Springer, New York, p 546

Buvat I, Castiglioni I (2002) Monte Carlo simulations in SPET and PET. Q J Nucl Med 46:48–61

Buvat I, Castiglioni I, Feuardent J, Gilardi MC (2005) Unified description and validation of Monte Carlo simulators in PET. Phys Med Biol 50:329–346

Castiglioni I, Cremonesi O, Gilardi MC, Bettinardi V, Rizzo G, Savi A, Bellotti E, Fazio F (1999) Scatter correction techniques in 3D PET: a Monte Carlo evaluation. IEEE Trans Nucl Sci 46:2053–2058

Comtat C et al (1999) Simulating Whole-body PET scanning with rapid analytical methods. In: Proceedings of IEEE Nuclear Science Symposium and Medical Imaging Conference, vol 3, Seattle, WA, October 24–30, pp 1260–1264

Cullen DE, Hubbell JH, Kissel L (2010) Photon and electron interaction data http://www-nds.iaea.org/epdl97/. University of California Lawrence Livermore National Laboratory, Livermore, CA

de Jong HWAM, Slijpen ETP, Beekman FJ (2001) Acceleration of Monte Carlo SPECT simulation using convolution-based forced detection. IEEE Trans Nucl Sci 48:58–64

Dewaraja YK et al (2002) A parallel Monte Carlo code for planar and SPECT imaging: implementation, verification and applications in I-131 SPECT. Comp Meth Prog Biomed 67:115–124

España S et al (2009) PeneloPET, a Monte Carlo PET simulation tool based on PENELOPE: features and validation. Phys Med Biol 54:1723–1742

Evans RD (1955) The atomic nucleus. McGraw-Hill, New York

Forster RA et al (2004) MCNP (TM) Version 5. Nucl Instr Meth Phys Res B 213:82–86

Haynor DR, Harrison RL, Lewellen TK (1991) The use of importance sampling techniques to improve the efficiency of photon tracking in emission tomography simulations. Med Phys 18:990–1001

Haynor DR (1998) Variance reduction techniques. In: Ljungberg M, Strand S-E, King MA (eds) Monte Carlo calculations in nuclear medicine. Institute of Physics, Bristol, pp 13–24

Jan S et al (2004) GATE: a simulation toolkit for PET and SPECT. Phys Med Biol 49:4543–4561

Kahn H (1954) Applications of Monte Carlo, USAEC Report AECU-3259. Rand Corporation, Los Angeles

Kaplan MS, Harrison RL, Vannoy SD (1998) Coherent Scatter Implementation for SimSET. IEEE Trans Nucl Sci 45:3064–3068

Kawrakow I (2000) Accurate condensed history Monte Carlo simulation of electron transport. I. EGSnrc, the new EGS4 version. Med Phys 27:485–498

Knuth DE (1997) Art of computer programming, vol 2, 3rd edn, Seminumerical Algorithms. Addison-Wesley, Reading

L'Ecuyer P, Simard R (2007) TestU01: A C library for empirical testing of random number generators. ACM Trans Math Softw 33(4): Article Number 22

Lewellen TK, Harrison RL, Vannoy S (1998) The SimSET program. In: Ljungberg M, Strand S-E, King MA (eds) Monte Carlo calculations in nuclear medicine. Institute of Physics, Bristol, pp 77–92

Ljungberg M (1998) Introduction to the Monte Carlo method. In: Ljungberg M, Strand S-E, King MA (eds) Monte Carlo calculations in nuclear medicine. Institute of Physics, Bristol, pp 1–12

Matsumoto M, Nishimura T (1998) Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator. ACM Trans Model Comput Simul 8:3–30

National Library of Medicine (2003–2010) The National Library of Medicine's Visible Human Project. http://www.nlm.nih.gov/research/visible/visible_human.html, National Institutes of Health, Washington, DC

Nelson WR, Hirayama H, Rogers DWO (1985) The EGS4 Code System. SLAC-265, Stanford Linear Accelerator Center, Menlo Park, CA

Palmer MR, Brownell GL (1992) Annihilation density distribution calculations for medically important positron emitters. IEEE Trans Med Imag 11(3):373–378

Rehfeld NS et al (2009) Introducing improved voxel navigation and fictitious interaction tracking in GATE for enhanced efficiency. Phys Med Biol 54:2163–2178

Reilhac A et al (2004) PET-SORTEO: validation and development of database of simulated PET volumes. IEEE Trans Nucl Sci 52:1321–1328

Reilhac A et al (2005) PET-SORTEO: a Monte Carlo-based simulator with high count rate capabilities. IEEE Trans Nucl Sci 51:46–52

Rosenberg KM (2002) CTSim 3.5 User Manual. Heart Hospital of New Mexico, Albuquerque

Sempau J, Fernandez-Varea JM, Acosta E, Salvat F (2003) Experimental benchmarks of the Monte Carlo code PENELOPE. Nucl Instr Meth Phys Res B 207:107–123

Strutt JW (1871) On the light from the sky, its polarization and colour. Philos Mag 41(107–120): 274–279

Tabary J, Hugonnard P, Mathy F (2007) SINDBAD: a realistic multi-purpose and scalable X-ray simulation tool for NDT applications. In: Proceedings of International Symposium on Digital Industrial Radiology and Computed Tomography, Lyon, France, June 25–27, 2007

Van der Laan DJ, Schaart DR, Maas MC, Beekman FJ, Bruyndonckx P, van Eijk CWE (2010) Optical simulation of monolithic scintillator detectors using GATE/GEANT4. Phys Med Biol 55(6):1659–1675

Wernick MN, Aarsvold JN (2004) Emission tomography: the fundamental of PET and SPECT. Elsevier Academic, London, p 175

Zaidi H, Ay MR (2007) Current status and new horizons in Monte Carlo simulation of X-ray CT scanners. Med Bio Eng Comput 45:809–817

Zerby CD (1963) A Monte Carlo calculation of the response of gamma-ray scintillation counters. In: Alder B (ed) Methods in computational physics, vol 1. Academic, New York, p 110

45 High-Resolution and Animal Imaging Instrumentation and Techniques

Nicola Belcari · Alberto Del Guerra
University of Pisa, Pisa, Italy

1 Introduction 1126

2 Small-Animal Imaging 1126

3 Key Technologies 1127
3.1 Present Technology for Small-Animal PET Systems 1127
3.2 Spatial Resolution Considerations in PET 1131

4 Present Technology for Small-Animal SPECT Systems 1133
4.1 Spatial Resolution Considerations in SPECT 1135
4.2 PET and SPECT Comparison 1137

5 Future Improvements in Small-Animal Instrumentation 1137
5.1 New Photodetectors 1138
5.2 New Detector Materials 1138

6 Small-Animal CT Imaging 1140

7 Multimodality Approach 1141
7.1 PET/CT and SPECT/CT 1142
7.2 PET/SPECT 1143

8 Other High-Resolution Applications of Radiation-Imaging Instrumentation: Breast Cancer Investigation 1146

9 Summary 1147

10 Cross-References 1147

References 1148

C. Grupen, I. Buvat (eds.), *Handbook of Particle Detection and Imaging*,
DOI 10.1007/978-3-642-13271-1_45, © Springer-Verlag Berlin Heidelberg 2012

Abstract: During the last decade we have observed a growing interest in "in vivo" imaging techniques for small animals. This is due to the necessity of studying biochemical processes at a molecular level for pharmacology, genetic, and pathology investigations. This field of research is usually called "molecular imaging."

Advances in biological understanding have been accompanied by technological advances in instrumentation and techniques and image-reconstruction software, resulting in improved image quality, visibility, and interpretation. The main technological challenge is then the design of systems with high spatial resolution and high sensitivity.

This chapter gives a short overview of the state-of-the-art technologies for high-resolution and high-sensitivity molecular imaging techniques, namely, positron emission tomography (PET) and single photon emission computed tomography (SPECT) as well as the basics of small-animal x-ray computed tomography (CT). Multimodality techniques merging molecular information with anatomical details are also introduced. Finally, the new trends in detector technology for other high-resolution applications like breast cancer investigation are presented.

1 Introduction

Molecular imaging can be defined as the representation, characterization, and quantification of biological processes occurring in a living subject at the cellular and subcellular levels (Massoud and Gambhir 2003).

The emergence of molecular imaging is due to the recent advances in molecular and cell biology techniques and in the development of small-animal imaging instruments.

Over the past few decades, many new technologies have been introduced into the field of "in vivo" imaging. These have included anatomical imaging techniques such as x-ray CT (x-ray computed tomography), US (ultrasound) and MRI (magnetic resonance imaging), and molecular or functional imaging techniques such as SPECT (Single Photon Emission Computed Tomography), PET (Positron Emission Tomography) and optical imaging. Some of these imaging technologies have seen clinical use for decades and, in recent years, they have found one of the most fruitful applications in the preclinical investigation using small laboratory animals. In this field, high-end technologies are continuously introduced and tested, pushing the imaging technology beyond the state of the art. In this chapter, high-resolution instrumentation applied in the field of imaging using ionizing radiation sources (i.e., positron, γ, or x-ray sources) are reviewed with a special attention to the small-animal application.

2 Small-Animal Imaging

Molecular imaging investigations are translated from man to small animals, such as mice and rats, down to cellular level, in order to obtain results on simplified human models before the direct study on patients.

Due to recent biogenetic innovations, transgenic animals showing particular anomalies obtained by genetic modification are now available. Rat and mouse have long been used by molecular biologists to study fundamental cellular events in vivo. Although the theoretical

capabilities of present imaging instrumentation are ideal for translational imaging purposes, the practical use of clinical scanners for imaging very small animals such as mice presents several challenges. This happens especially because mice are much smaller than humans (about 1,500 times in volume). Hence, small-animal imaging requires the development of instruments able to achieve a finer spatial resolution with respect to the available clinical scanners. For example, the imaging of the rat brain requires a spatial resolution less than 2 mm full width at half maximum (FWHM), while for mice it would be ideal to use instruments with a resolution less than 1 mm (FWHM) so that images can be obtained with the same visual acuity as in humans.

Another requirement for small-animal imaging instrumentation is the ability of measuring very small signals because, in many biological processes, the effective concentration of the involved molecules, such as hormones, transmitters, messengers, and carriers, is often smaller than 10^{-12}–10^{-15}mol ml^{-1}. For their visualization, dedicated molecular "probes" are used as a source of contrast.

Although new specific imaging probes and drugs are now available, highly sensitive imaging techniques are required for molecular imaging. High-sensitivity instrumentation is especially required when fast dynamic processes with characteristic time of the same order of the scanning time (usually a time period of about 30–45 min during which small animals can be anesthetized safely) are studied in small animals or when low-activity regions (as for low-uptake processes in gene or stem cells research) have to be visualized and quantified. In fact, biological constraints limit the total activity to be injected. For instance, in order to avoid the perturbation of the stationary conditions or the saturation of the receptors, the typical injected activity in a mouse for brain receptor investigation is not greater than 5–10 MBq. In addition, there are limitations on the maximum volume of injected solution (~10% of the total blood volume).

3 Key Technologies

Among the various molecular imaging techniques, PET and SPECT have demonstrated to be very valuable investigation methods, and they are characterized by an extensive clinical use. However, the relatively small size of the objects under study in small-animal imaging (small organs or brain regions of rats and mice) makes it difficult to use imaging instruments developed for human subjects, i.e., the sensitivity and spatial resolution of the available clinical SPECT or PET scanners are not satisfactory for the quantitative and qualitative in vivo investigations, e.g., the assessment of gene expression (Massoud and Gambhir 2003). For example, the best spatial resolution of presently available clinical PET scanner is not better than 4–6 mm FWHM.

The success of PET and SPECT systems for small animals is largely due to the continuous development of high-resolution, high-sensitivity instrumentation for γ-ray detection that has strongly improved the performance of small-animal scanners with respect to the clinical ones. This is especially true for the spatial resolution figure of merit so as to fulfill the requirements for small-size animal study.

3.1 Present Technology for Small-Animal PET Systems

Current clinical PET systems are designed in a ring geometry. This design enables the patient to be completely surrounded by detectors, which are still based upon the block detector concept

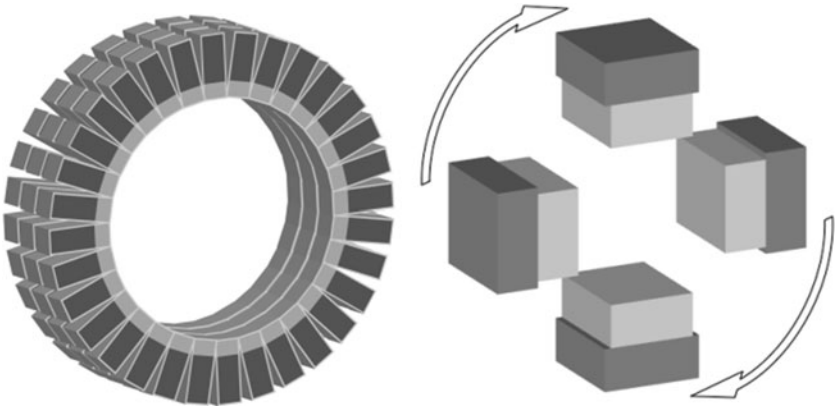


Fig. 1

Two different configurations for the construction of a small-animal PET scanner. *Left*: ring geometry, where the detectors are arranged on a ring geometry and surround the animal (adapted from Volterrani et al. 2010, Chap. 12, Fig. 12.2, p. 258). *Right*: Example of a rotating detectors configuration with four heads, where each one is in time coincidence with the opposite one

that relies on quadrant light sharing over single-channel PMT (Zanzonico 2004). This design ensures both high detection efficiency and a suitable resolution for human applications.

In the last 20 years, the advent of novel high-resolution position-sensitive photodetectors has dramatically expanded the possibilities for the construction of more simple and flexible systems with improved performances.

The main design of small-animal PET instruments is usually based on a miniaturized structure of a clinical scanner with small detector elements surrounding the animal in a small bore ring (Chatziioannou et al. 1999). However, some alternative designs using rotating planar detector pairs have been realized (Del Guerra et al. 1998) (*Fig. 1*).

To achieve an improved spatial resolution, the detectors take the form of position-sensitive photomultiplier tubes (PS-PMT) coupled to pixellated matrices of scintillators. This represents the simplest approach for the gamma interaction position encoding. In this case, the position of the gamma interaction is coded by the crystal position where the interaction occurs. The crystal position is, in turn, identified via "light-sharing" technique, i.e., by calculating the centroid of the light spot emerging from the crystal with a high-granularity position-sensitive PMT. The PS-PMT is now at the third generation. Early PS-PMTs were based on a typical round shape, with size up to 10 cm diameter, with a crossed wire anode structure. The second generation has a square-shaped metal-channel dynode structure with a very fine anode structure made by crossed metal plates or a matrix of independent square anodes. This type of PS-PMT provides a very good intrinsic spatial resolution and uniformity but was limited in size (up to about 2 cm side). With the third generation there has been a great improvement in the active area dimensions (up to 5 cm in side) and active-to-total area ratio (up to about 90%), still maintaining the performance as close as possible to the second-generation tubes. These 5 cm tubes are based on the metal-channel dynode structure with an anode matrix of 16 × 16 elements on a 3 mm pitch (*Fig. 2*).



Fig. 2
Example of first (*left*, Hamamatsu R2486 with crossed-wire anode structure), second (*center*, Hamamatsu R8520 with crossed-plate anode structure), and third (*right*, Hamamatsu H8500 with multi-anode structure) (images from the Hamamatsu web site: www.hamamatsu.com)

To fully exploit the performance of PS-PMT, each anode should be independently acquired (multi-anode readout). However, in order to simplify the complexity of the readout system, the most widely used approach for the PS-PMT readout is the utilization of resistive chains (Popov et al. 2001; Olcott et al. 2005) so as to strongly reduce the number of output channels. Many different combinations of scintillator pixellization and PS-PMT are presently used depending on the specific application and on the required field-of-view (FOV) size. However, this solution results in a twofold contribution to the degradation of the spatial resolution of the detector (see Sect. 3.2) for the finite crystal size (crystal pitch error) and the possible error in the identification of the hit crystal (coding errors).

As an alternative method for the readout of matrices of scintillators, solutions based on semiconductor photodetectors are also used with the intent of overcoming the coding limitation. For example, matrices of small-area Avalanche Photodiodes (APDs) are used for the parallel readout of the pixellated matrices of a scintillator (*Fig. 3*). In this case, each crystal is coupled to a single APD (Pichler et al. 2001). This method, similarly to those used in the early PET of the 1970s but now with a pixel pitch down to 1 mm, allows a very simple, almost "perfect" one-to-one coding at a very high count-rate. Consequently, there is no resolution loss due to light sharing or electronic coding, and the true geometric crystal resolution can be achieved. An additional advantage is the negligible dead time and pile-up probability.

On the other hand, detectors based on a scintillator block coupled to high-granularity position-sensitive photodetectors (Anger camera principle) have been proposed (Joung et al. 2001; Balcerzyk et al. 2009). In this way, the limitation introduced by the finite size of the crystal elements could be overcome by measuring, with high precision, the centroid of the light spot.

An additional issue for the development of dedicated PET instrumentation is to provide an adequate efficiency for the detection of 511 keV photons. In particular, high-sensitivity instrumentation and design are required. In order to maximize the efficiency of the PET system, the PET heads should be positioned close to the object, and the thickness of the photon absorber should be at least one attenuation length at 511 keV. In this case, some lines of response (especially those having a high geometrical inclination) are subjected to a potentially significant depth-of-interaction (DOI) error (*Fig. 4*). A number of techniques for designing detectors with DOI capability have been proposed, based either on the direct measurement of the DOI within the crystal (Moses and Derenzo 1994; Balcerzyk et al. 2009) or by segmenting the crystal

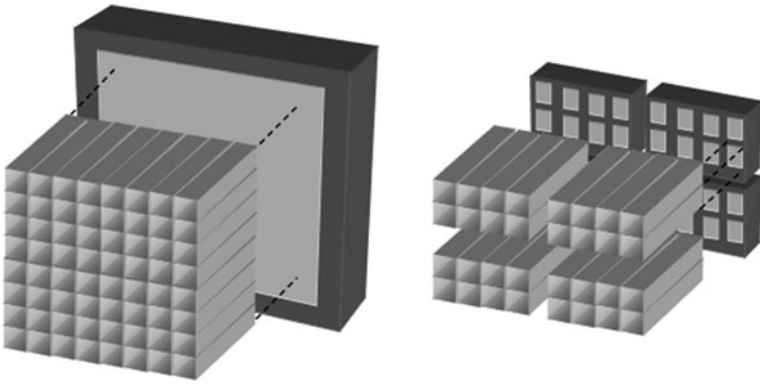


Fig. 3

Example of scintillating matrix to photodetector coupling. *Left*: a matrix is read out by a large-area PS-PMT (electronic coding). *Right*: each element of the matrix is coupled to a single photodetector (e.g., APD) (one-to-one coupling)

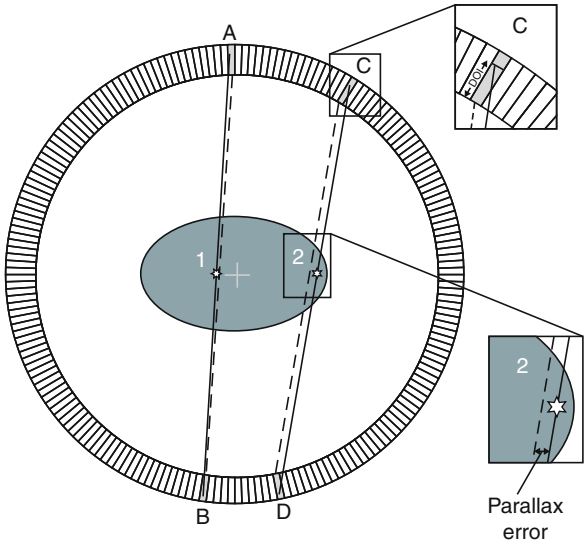


Fig. 4

Representation of the parallax error close (case 1) or far (case 2) from the tomograph center (indicated by a cross sign in the figure). In case 1, the two γ rays reach the crystals with a trajectory that is parallel to their axis. In this case, the parallax error is negligible and is independent from the depth of interaction. In case 2, the γ rays reach the detector with an inclined trajectory. In this case, a large parallax error can be observed, especially at deep interactions in the crystal (adapted from Volterrani et al. 2010, Chap. 12, Fig. 12.7, p. 268)

into two layers so that the photodetection system is able to discriminate events occurring in one layer from the other (Saoudi et al. 1999; Seidel et al. 1999).

The simultaneous improvement of spatial resolution and sensitivity is one of the challenges of PET imaging since the two figures are often in contrast, i.e., increasing one could cause the reduction of the other. As an example, increasing the scintillator thickness or reducing the detector ring diameter would cause an increase of sensitivity but also a reduction in spatial resolution and vice versa. This fact will become clearer when the factors affecting spatial resolution in PET are described in next section.

Every year, new small-animal PET prototypes are produced or proposed by many research groups offering or promising even better performance (Miyaoka et al. 2002; Rouze and Hutchins 2003; Tai et al. 2003). At the same time, some fully engineered scanners are released as commercial products. Nowadays, several products are present on the market.

3.2 Spatial Resolution Considerations in PET

The best achievable spatial resolution in PET is limited due to both the physics of the β^+ decay and the available technology for the position detection of two γ rays in coincidence.

In fact, the positron is emitted with a non zero kinetic energy, and it is slowing down in tissue via Coulomb interactions. The energy loss continues until the positron reaches the thermal equilibrium with the surrounding tissue and annihilates with an electron. The positron range depends on tissue (water equivalent) electron density. In water, the range of the positron emitted by a typical PET radioisotope is of the order of 1–2 mm. This range effect degrades the spatial resolution of a PET tomograph. In addition, the annihilation occurs not quite at rest. As a consequence, the two photons are not emitted, in the laboratory frame, at exactly 180°, but they have an angular deviation from non-collinearity of ±0.25 degrees. The annihilation of the positron with an electron can occur via different intermediate states contributing in a different way to the angular deviation from collinearity of the annihilation γ pair. In fact, in water, the positron has a probability of 62% to annihilate with an electron in a non-bound state. This effect generates a narrow ($|\Delta\theta| < 4$ mrad) component of angular deviation. A probability of 38% is for the formation of a positron–electron bound state, called positronium, that can be bound, depending on the relative electron and positron spin orientation, in ortho-positronium ($S = 1$, $\uparrow\uparrow$) or para-positronium ($S = 0$, $\uparrow\downarrow$) states. The para-positronium (25% of the bound states) usually decays by self-annihilation generating again a narrow contribution to the noncollinearity. On the other hand, the ortho-positronium (75% of the bound states) annihilates via a pick-off process with a free electron. The latter effect largely contributes ($|\Delta\theta| > 4$ mrad) to the noncollinearity. Range effect and noncollinearity are the two fundamental processes that intrinsically limit the spatial resolution in PET (*Fig. 5*, *left*). The best spatial resolution achievable is also limited by other factors, i.e., crystal size, crystal position readout coding, and image-reconstruction algorithm. The best spatial resolution of a PET scanner can be expressed in terms of the FWHM of the point-spread function (PSF) after a filtered back projection (FBP) reconstruction with the following formula (Derenzo and Moses 1993):

$$\mathrm{FWHM} = 1.2\sqrt{\left(\frac{d}{2}\right)^2 + b^2 + (0.0022\,D)^2 + r^2 + p^2},$$

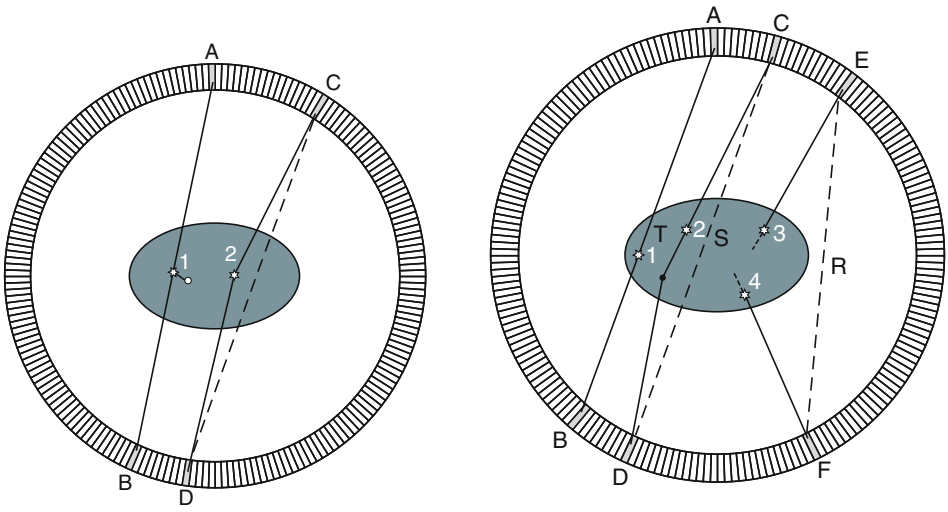


Fig. 5

***Left*: Representation of the *range effect* (case 1) and *non-collinearity* (case 2). In both cases the reconstructed line of response does not pass through the emission point generating a degradation of the spatial resolution (from Volterrani et al. 2010, Chap. 12, Fig. 12.6, p. 267). *Right*: Representation of *true* (T), *scattered* (S) and *random* (R) events (from Volterrani et al. 2010, Chap. 12, Fig. 12.9, p. 272)**

where 1.2 is the degradation factor due to tomographic reconstruction, d is the crystal pitch, b is the coding error, D is the tomograph ring diameter, r is the effective source diameter including positron range, p is the parallax error, all values are expressed in mm. The last contribution for a ring tomograph can be approximated by the expression:

$$p = \alpha \frac{r}{\sqrt{r^2 + R^2}},$$

where r is the distance from center, R is the tomograph radius, and α depends on thickness and type of the scintillator (for example, $\alpha = 12.5$ for 30 mm thick LSO or BGO). The parallax error is related to the crystal thickness. In ring geometry, the lines that do not pass through the tomograph center are affected by an uncertainty that increases as the distance from the center increases. Considering all the contributions included in the formula for the FWHM, it can be found that the spatial resolution in PET is intrinsically limited and cannot be better than 0.6–0.8 mm FWHM which represents the theoretical limit value.

Due to the available technology, present small-animal systems have a typical spatial resolution of about 1.3–1.6 mm FWHM at the center of the field of view (FOV). This is good enough for most cases (especially for rats). Although in PET modality the maximum achievable spatial resolution is limited by the physics of the β^+ decay, latest systems, incorporating the high-granularity detectors, and in some cases, DOI capability offer a volumetric spatial resolution of about 1 mm^3 that is very close to the theoretical one. On the other hand, the system sensitivity was continuously improved up to a present maximum of 8–10% at the center of the FOV (Laforest et al. 2004).

Another factor reducing image quality in addition to the spatial resolution degradation is noise. The major sources of noise in PET are scattered and random counts (*Fig. 5*, *right*). Several techniques are implemented in high-resolution PET to reduce noise, including the use of new materials as well as acquisition techniques such as time-of-flight PET as described in Sect. 4.

Thanks to its exquisite sensitivity, radioisotope imaging using high-performance PET represents the gold-standard technique for molecular imaging investigation (Del Guerra and Belcari 2007).

4 Present Technology for Small-Animal SPECT Systems

Small-animal SPECT are based on two main design approaches: systems that are obtained by incorporating a novel collimator on clinical systems (Schramm et al. 2003; Beekman et al. 2005) and dedicated systems based on compact high-resolution detectors (McElroy et al. 2002; Weisenberger et al. 2003; Furenlid et al. 2004; Kastis et al. 2004). The key feature of both types of scanners is the collimator, which is specially designed to obtain high-resolution in vivo images of small organs or a whole small animal. In human SPECT applications, parallel-hole collimators are used. These are based on a regular array of round, square, or hexagonal holes in a high-density medium (lead or tungsten). However, the spatial resolution is relatively poor for animal application, even if some examples exist (Del Guerra et al. 1998).

For small-animal imaging, an appropriate collimator is required for high (submillimeter) spatial resolution and high sensitivity. The most widely applied collimator solution is the pinhole (or multi-pinhole) collimator, which increases the spatial resolution of the imaging system by projecting a magnified, nonoverlapping view of the animal onto the detector. In this case, a large bulky gamma camera can be used also for small animals. In fact, the large area of these cameras can be used to compensate for the low detector resolution. For these systems, the detectors are based on a large-area single scintillation crystal coupled to an array of photomultiplier tubes (PMT). The scintillator is usually thallium-doped sodium iodide (NaI:Tl), which has ideal properties for detecting medium-low-energy γ rays (e.g., ^{99m}Tc, ^{201}Tl, ^{123}I, and ^{111}In) in terms of light yield, energy resolution, and efficiency. A limitation of this design is the low number of pinholes that can be used since the large magnification needed to bypass the low resolution of the detector produces large projections that may overlap as the number of pinholes increases. The main approach to address the problem of overlapping projections is the use of iterative estimation methods (Meikle et al. 2001) which are based on the earlier concept of coded aperture (Barrett 1972), which is, in turn, a still valid solution for small-animal imaging (Meikle et al. 2002).

To increase the number of pinholes the solution based on dedicated high-resolution detectors is the preferred one. In this way, the animal can be surrounded by tens (Liu et al. 2002) or even up to hundreds of pinholes so as to compensate for the relatively low photon count, making it possible to use very low pinhole diameters. The use of multiple overlapping pinhole projections increases the absolute detection efficiency, but this effect is not reflected into an increased signal-to-noise ratio due to the reduced information conveyed by each detected photon in a multiplexed system with respect to a non-multiplexed system caused by the ambiguity on the defined line of flight (Meikle et al. 2002). Such kind of detectors can be arranged around the animal like multi-head SPECT, up to four heads, or set up in a polygon-like manner,

effectively forming a large number of pinhole cameras to image the small animal (Beekman and Vastenhouw 2002).

Solutions for the construction of high-resolution detectors can be subdivided into two design categories: those using one small-size multi-crystal array coupled to one or more PS-PMTs and those using solid-state detectors. Scintillating detectors typically employ arrays of small (1–2 mm) scintillation crystals, such as NaI:Tl, CsI:Tl, and YAP:Ce, that are optically coupled to PS-PMTs. The maximum active area achievable in this way is smaller than those offered by standard gamma cameras, but they require less magnification to achieve similar performance due to the better intrinsic spatial resolution. A drawback of pixellated detectors is a poorer energy resolution due to a lower light output from the small crystals with respect to monolithic scintillators.

On the other hand, solid-state detectors provide a promising alternative technology as compact high-resolution gamma cameras. Semiconductor detector technology is the new horizon in dedicated instruments for high-resolution nuclear imaging and such solid-state detectors with direct γ-ray conversion have been proposed. The requirements for a good detector for SPECT, i.e., high spatial resolution, high energy resolution, and good efficiency for the detection of medium energy γ rays, are only partially fulfilled by solutions based on scintillators and photomultipliers. In particular, in a gamma camera, the spatial resolution is usually limited by the low granularity of the photomultiplier tube array, while the energy resolution is limited by the scintillation phenomena and by the relatively low quantum efficiency (about 20%) of the photomultiplier. On the other hand, semiconductor detectors have similar detection efficiency, but they have better performance, by about a factor 2, both in spatial and energy resolution. Semiconductor detectors are now available in relatively small sizes, but they are well suited for the construction of small-animal imaging and dedicated instruments. The working principle of semiconductor detectors for the detection of γ rays with an energy below 1 MeV relies on the direct conversion of the photon energy in a number of electron–hole pairs, which, when transported to the corresponding electrode, represent a measurable charge signal, whose intensity is proportional to the energy of the incident photon. The described interaction principle is a direct conversion mechanism. In fact, the incident photon energy is directly converted in an electrical signal. The advantage of this approach stays in the reduction of the spatial resolution degradation that occurs in scintillators due to the degradation of the spatial information for the spreading of the light through the scintillator. Conversely, in semiconductors, the lateral spreading of the electron cloud is limited and nearly negligible with respect to the pixel size, which is the major resolution limit.

CdTe and CdZnTe are typical semiconductor materials for the construction of nuclear-imaging detectors. Thanks to their characteristics (see *Table 1*), they can be used for the construction of compact gamma cameras.

Table 1

Physical properties of CdTe and $Cd_{0.9}Zn_{0.1}Te$, in terms of density (ρ), effective Z (Z_{eff}), attenuation coefficient (μ) at 140.5 keV, energy to produce an electron–hole pair (E_{e-h}) and photofraction (PhF)

Material	ρ (g/cm^3)	Z_{eff}	μ (140.5 keV) (cm^{-1})	E_{e-h} (eV)	PhF (%)
CdTe	5.85	50	4.15	4.43	84
$Cd_{0.9}Zn_{0.1}Te$	5.78	49.1	3.96	4.64	83

Present CdTe or CdZnTe detectors are available in 1–5 mm thickness, and 5 mm is the most commonly used value for SPECT. Pixel pitches are also available in the submillimeter range. The typical energy resolution for 140.5 keV γ rays is about 5%.

In particular, CdZnTe-based detectors (Funk et al. 2003), thanks to the superior energy resolution with respect to scintillators together with their good photofraction, make it possible to reject Compton events so as to increase the signal-to-noise ratio in the final image. In addition, the excellent energy resolution facilitates multi-radionuclide studies, thus enabling dual- or triple-radionuclide imaging.

4.1 Spatial Resolution Considerations in SPECT

SPECT does not suffer from intrinsic spatial resolution limitations. On the other hand, the main goal is to find the optimal compromise between spatial resolution and sensitivity.

For parallel-hole collimators (Fig. 6), the spatial resolution, $R_{C,\,\text{parallel}}$, in terms of FWMH, is given by Zavattini and Del Guerra (2004):

$$R_{C,\,\text{parallel}} \approx D + d(D/L),$$

where D is the hole diameter, d is the collimator–object distance, and L is the hole length.

$$R_{\text{TOT}} = \sqrt{{R_{C,\,\text{parallel}}}^2 + {R_D}^2},$$

where R_D is the intrinsic spatial resolution of the detector.

Hence, the spatial resolution improves when the object gets closer to the collimator and when the holes get smaller. Conversely, the sensitivity g does not depend on the distance, and it increases when the holes get larger, i.e.,

$$g \propto \left(\frac{D}{L}\right)^2 \cdot \left(\frac{D^2}{(D+h)^2}\right),$$

where h is the thickness of septum between holes (Zavattini and Del Guerra 2004).

The use of pinhole collimators is a solution for ultrahigh-resolution SPECT. A pinhole collimator consists in a single hole shaped like a double cone (Fig. 7).

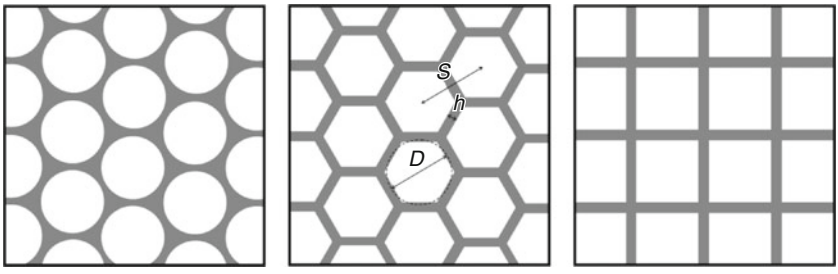


Fig. 6

Typical hole patterns in parallel-hole collimators having round holes in a hexagonal pattern (*left*), hexagonal holes in a hexagonal pattern (*center*), and square holes in a square pattern. *D*, *S*, and *h* represent the hole diameter, the hole-to-hole distance, and the septum thickness, respectively (from Volterrani et al. 2010, Chap. 9, Fig. 9.6, p. 224)

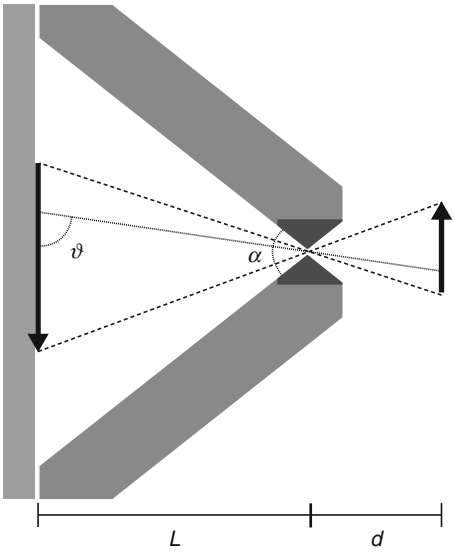


Fig. 7
Schematic representation of a pinhole collimator (from Volterrani et al. 2010, Chap. 9, Fig. 9.7, p. 226)

These collimators are usually made of high-Z materials (such as gold) in order to reduce the radiation penetration at the edge of the hole. For this type of collimators, the FWHM of the spatial resolution is given by (Zavattini and Del Guerra 2004):

$$R_{C,\,\text{pinhole}} = D_e \cdot (d + L)/L$$

where d is the distance between the object and the pinhole and L is the distance between the pinhole and the scintillator. D_e represents the effective diameter of the pinhole, depending also on the attenuation coefficient of the collimator material and on the hole aperture. Differently from parallel-holes collimator, the sensitivity of a pinhole is proportional to $1/d^2$. Thus, the closer the object to the hole, the higher the sensitivity. However, the sensitivity rapidly decreases as the distance increases. By using the intrinsic geometrical magnification of a pinhole collimator, it is then possible to obtain high-resolution images with a FWHM of the same order of $D_e (L > d)$ and a good sensitivity (only using small d). The magnification allows the use of instruments with an intrinsic resolution larger than the system resolution without significantly affecting it. On the other hand, by reducing the pinhole to object distance, the size of the FOV is proportionally reduced. Due to these reasons, pinhole collimators are often used in combination with large detectors, such as those used for humans with a high magnification factor, or with high-resolution detectors. In this way, a value of the order of few hundred microns is obtainable. By using more holes (multi-pinholes), it is possible to enlarge the FOV to obtain "whole-body" high-resolution images of small animals.

$$R_{\text{TOT}} = \sqrt{{R_{C,\,\text{pinhole}}}^2 + \left(R_D \frac{d}{L}\right)^2}.$$

4.2 PET and SPECT Comparison

PET and SPECT offer overall performances that fulfil the requirements of most of molecular imaging investigations, and each one possesses relative strengths and weaknesses. In PET, the spatial resolution is limited by the physics of the β^+ decay, while, in SPECT, there is no such limitation. In the last years, SPECT images of a mouse with a spatial resolution of few hundreds of microns have been reported.

On the other hand, PET still offers a higher sensitivity than SPECT thanks to the electronic collimation obtained by the coincidence γ-pair detection instead of the physical collimation used in SPECT. This gives a substantial advantage in sensitivity of PET over SPECT, typically more than one order of magnitude. Nanomolar concentrations (10^{-9} mol l^{-1}) of radiopharmaceutical are typically measured with PET and picomolar concentrations (10^{-12} mol l^{-1}) are achievable. However, the sensitivity gap between PET and SPECT is reducing over the years thanks to the use of multi-pinhole collimators. The SPECT sensitivity is now about 10–40 times less than the best PET scanners with a similar spatial resolution as for PET.

PET isotope and tracer availability is more limited than for SPECT, but light PET isotopes can be used as a substitute of the analog stable elements present in the parent molecules. In addition, the ^{18}F can be used as a substitute for hydrogen. For this reason, PET is essential for imaging small-molecule drugs. SPECT, on the other hand, offers easier access to longer-lived isotopes, which are well suited for labeling larger molecules such as peptides and antibodies.

In this respect, a second important consideration is the kinetics of the drugs under study. For drug candidates with slow kinetics, the best approach is to use SPECT tracers because of the longer half-life of these radioisotopes. These drugs fall usually in the category of biopharmaceuticals, i.e., large molecules for which the addition of a SPECT radioisotope do not impair the action of the drug or biomarker. For small molecules on the other hand, the PET isotope ^{11}C is the ideal radioisotope since it can be synthesized in the candidate drug. Unfortunately, the very short half-life and the low specific activity of this tracer restrict the use of ^{11}C tracers to molecules that can be relatively quickly synthesized and for imaging studies of short duration. For other small-molecule applications, radio halogens for PET and SPECT may have to be used instead. In this respect, one of the advantages of using radio halogens for SPECT is the ability to utilize ^{125}I and ^{131}I tracers used commonly for in vitro and clinical in vivo studies, respectively.

5 Future Improvements in Small-Animal Instrumentation

A major concern for the development of the next generation of PET systems for small-animal imaging is the improvement of sensitivity, still pushing the spatial resolution close to its intrinsic limit. To overcome the limitation of present systems novel scintillating crystals for γ detection are required in combination with new photodetectors able to exploit the crystal features. On the other side, small-animal SPECT has not reached its resolution limit yet, but the best present systems have a spatial resolution that is perfectly adequate for the study of small-animal organs. In this case, the main challenge is to increase the sensitivity and the field of view to obtain ultrahigh-resolution systems able to visualize the entire animal with a high sensitivity.

5.1 New Photodetectors

The development of novel photodetectors has a fundamental role for the future of PET imaging. The possible solutions proposed by various research groups have a twofold approach. On the one hand, systems overcoming the coding limitation have been introduced through APD-based solutions (one-to-one coupling) (Pichler et al. 2004), while, on the other hand, solutions based on a further refinement of the gamma camera concept using a monolithic scintillating crystal have been proposed (Moehrs et al. 2004).

An example of a novel photodetector is the Silicon Photomultiplier (SiPM). The SiPM (Golovin and Saveliev 2004) is a silicon-diode detector that shows great promise as a photodetector for scintillators for both one-to-one coupling and Anger camera approaches. In short, the SiPM is a densely packed matrix of small Geiger-mode avalanche photodiode (GAPD) cells (typically ~$40 \times 40\ \mu m^2$), with individual quenching resistances for each cell. The Geiger-mode operation of each cell produces a large gain (of the order of 5×10^5) at low bias voltage (~50 V). All the cell outputs are connected in parallel to produce the summed signal. This microcell structure of the SiPM gives a proportional output for moderate photon flux. The performance is in many ways comparable or superior to that of a conventional PMT, but with the compactness and other benefits of a semiconductor detector. Such a compact silicon detector is well disposed for being developed into a close-packed array so as to have a position-sensitive detection surface, and this justifies the wide interest and the great efforts dedicated to the SiPM development (Herbert et al. 2006). Examples of monolithic matrices of SiPM are now available, including up to 64 elements in an 8 × 8 array (12 mm × 12 mm) with 1.5 mm pitch. These devices are well suited for PET (Llosa et al. 2009) as well as for SPECT applications.

5.2 New Detector Materials

On the other hand, due to the availability of higher-granularity photodetectors, it becomes possible to use slabs of scintillators instead of pixel matrices. In this case, the position encoding is performed by calculating the centroid of the light spot reaching the photocathode (Anger camera principle). Hence, in this second solution, the limitation introduced by the finite size of the crystal elements could be overcome. An additional advantage of this approach is the reduced cost of a scintillator slab with respect to matrices. The main drawback is the complexity of the readout system since each element (e.g., anode) of the photodetector has to be independently acquired (Pichler et al. 2001).

A factor affecting sensitivity in a scintillating crystal is the probability of photoelectric emission, which is proportional to ρZ_{eff}^4. Among all the available scintillators, the highest value is for BGO (*Table 2*). New material research is focused on keeping the value of ρZ_{eff}^4 as close as possible to the BGO value (so as to control parallax error) while improving those parameters where BGO is not optimal, e.g., decay time (improvement in count-rate properties) and light yield (better energy and spatial resolution).

Nowadays, a detector based on a matrix of high-Z, fast, scintillating crystals (e.g., LSO:Ce) coupled to a position-sensitive photodetector is a well-established technology and represents the state of the art. However, all high-Z materials have a photofraction below 50% at 511 keV.

This implies that events under the full-energy peak, usually chosen as the "best" events, have a large contribution from multiple Compton interactions in the crystal. Hence, the spatial resolution is degraded by multiple site charge deposit. An alternative approach is the use

Table 2
Physical features of typical scintillating materials used in SPECT and PET

Material	Density (g/cm^3)	Z_{eff}	Light yield (Ph./MeV)	Decay time (ns)	Wavelength (peak) (nm)	Attenuation length (at 511 keV) (mm)	Refractive index	Photo-fraction (%)
NaI:Tl[a]	3.76	51	41,000	230	410	29.1	1.85	17
BGO[b]	7.13	75	9,000	300	480	10.4	2.15	40
LSO:Ce[b]	7.4	66	30,000	40	480	11.4	1.82	32
GSO:Ce[b]	6.71	59	8,000	600	430	14.1	1.85	25
CsI:Tl[a]	4.51	52	66,000	1,000	565	22.9	1.80	21
LuAP:Ce[b]	8.3	64.9	12,000	18	365	10.5	1.94	30
LaBr:Ce[a,b]	5.29	46.9	63,000	16	380	21	1.9	15
YAP:Ce[a,b]	5.37	33.5	23,000	27	370	21.3	1.95	4.2

[a]SPECT
[b]PET

of medium-low-Z materials, such as YAP:Ce scintillators (Del Guerra et al. 1996) or silicon crystals (Auricchio et al. 2005). In such materials, most of the events will be Compton events, e.g., in YAP:Ce, the photofraction is only 4% at 511 keV. However, by selecting the entire energy spectrum, one achieves high sensitivity, and a high fraction of these events will be single interactions, therefore providing the correct interaction position (Zavattini et al. 2005). Of course, this technique is only possible when Compton scattering in the object can be neglected or easily modeled.

A further improvement, of a great potential impact in PET imaging performance, would be the reduction of the time resolution of about one order of magnitude. In addition to the obvious reduction of random events, it would become possible to utilize the time-of-flight (TOF) information for image reconstruction. This is the so-called TOF-PET concept (*Fig. 8*), where the time difference of the detection of two annihilation γ rays is measured in the two coincident detectors. The timing information can be used to remove the coupling between voxels that are separated by more than the TOF measurement distance. For example, using $LaBr_3$:Ce, a time resolution of 300–500 ps might be possible. Such a figure corresponds to a net reduction of the background variance, which is equivalent to an improvement in sensitivity.

Thallium-doped sodium iodide (NaI:Tl) is the most common material for SPECT both in the form of a monolithic slab or in pixellated matrices. A possible alternative is a thallium-doped cesium iodide (CsI:Tl) scintillator which has high-Z components, high density, and excellent scintillation light yield, which makes it a good candidate for SPECT. In addition, CsI:Tl crystals can be grown in tiny columnar structures (Nagarkar et al. 1996) to form a continuous slab of material, where the light spread is strongly reduced yielding a high spatial resolution. To maintain a high resolution, these crystals might be read out with high-spatial-resolution photodetectors.

In addition to CdZnTe detectors described in Sect. 4, other direct conversion solutions based on single-photon-counting devices have been proposed. For example, fine-pitch silicon hybrid pixel detectors read out by the Medipix2 chip (Accorsi et al. 2004) and electron-multiplying CCD cameras (de Vree et al. 2005) have been employed for the construction of a high-resolution gamma camera specially tailored for low-energy radionuclides imaging such as ^{125}I.

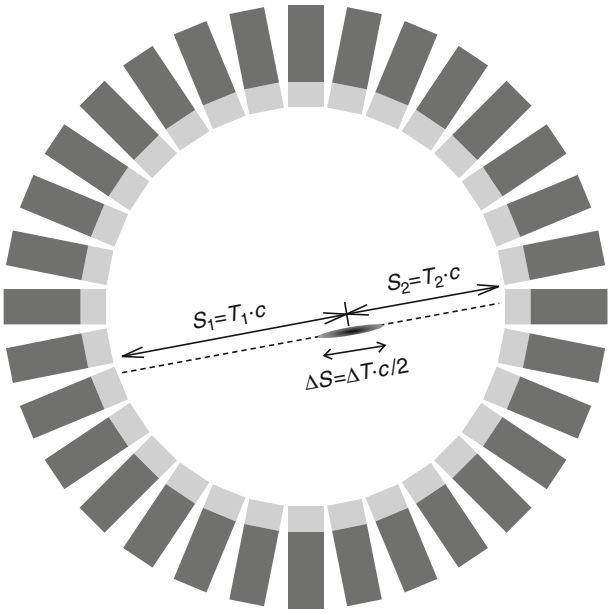


Fig. 8

Time-of-flight PET principle. The γ-pair emission point is determined by the length of S_1 and S_2 segments calculated from times of flight T_1 and T_2. The quantity that is measurable in practice is the time difference $T_1 - T_2$ that can be used to calculate the distance of the emission point from the line-of-flight center. Due to the error in measuring the time difference, the error on the position of the emission point is given by $\Delta S = \Delta T \cdot c/2$ (from Volterrani et al. 2010, Chap. 12, Fig. 12.5, p. 262)

6 Small-Animal CT Imaging

Computed Tomography (CT) is a very important technique of noninvasive diagnosis, which provides a 3D map of the local X-ray attenuation properties of the scanned object or patient. Dedicated CT scanners for the imaging of small objects and animals at very high-resolution (micro-CT) have been developed (Stock 2008). Those scanners find application in biomedical research (Paulus et al. 2000) and in nondestructive testing (NDT) of materials and components (Levine et al. 1999). All micro-CT scanners are based on the same design: an x-ray source, an imaging detector, and a system that either rotates the animal or the specimen within a stationary scanner or rotates the scanner around the object. Thanks to the availability of relatively large-area flat-panel x-ray detectors, the x-ray transmission is usually made with cone-beam geometry. For the CT with a rotating object, this is generally rotated around the vertical axis within a fixed x-ray source/imaging detector system. This is unsuitable when working with an anesthetized animal since it would modify its physiological behavior if held in this vertical orientation for an extended period of time. Rotating gantry systems are hence preferred for in vivo imaging of small animals. In this case, the x-ray source and the imaging detector rotate around the animal.

A tungsten-anode x-ray tube with a relatively small focal spot (~10 μm) emitting low-energy x-rays (30–50 kVp) is commonly used. Small-animal CT systems can be optimized for spatial resolution, to obtain an image that approaches histological microscopy as closely as possible, by using a micro focus x-ray tube, combined with high geometric magnification. A small focal spot is mandatory in systems implementing a high geometric magnification, to minimize penumbral blurring. Transmission images are collected with flat panel detectors made up of magnified CCD- or CMOS-based detectors featuring high spatial resolution (~50 μm pixel spacing). The system spatial resolution is limited by several factors, including the pixel spacing of the x-ray detector, the geometric magnification, the focal spot size, and the stability of the rotation mechanism.

The demand for high-precision methods for the calibration of high-resolution micro-CT scanners is motivated by the very high dependence of the scanner performances on the mechanical positioning of the various components. An uncorrected displacement of the axis of rotation (AoR) in the order of magnitude of the detector pitch may result in a significant image blurring, and consequently in a loss of image spatial resolution. Displacements are continuously added in the system by many factors, including mechanical instability, vibrations, thermal drift of bearings and components, etc. Many methods have been proposed for the measurement of the misalignment parameters from the analysis of the acquisition data both using a reference phantom for the 2D fan-beam geometry (Gullberg et al. 1987) and 3D cone-beam geometry (Boone 2006), as well as a semiautomatic optimization-based method without the needs of dedicated phantom measurements (Panetta et al. 2008).

X-ray projection views are acquired at hundreds of equally spaced angular positions around the object of interest either with a step-and-shoot (where the rotation is stopped while the detector acquires data) or with a continuous rotation. Views are then used to reconstruct a CT image, typically using a convolution back-projection approach (implemented in 3D) (Feldkamp et al. 1984). A spatial resolution of 15–50 μm, over a field of view up to 100 mm, is usually obtained.

In vivo imaging requirements include additional design constraints on the system, the first of which is the dose limitation and second the speed of the scan that should be reduced to a few minutes. The dose reduction is sometimes obtained by reducing the spatial resolution with the use of 50–100 μm voxel size. This is because image noise is proportional to $(\text{voxel size})^{-2}$ at constant animal exposure (Zhu et al. 2009). Hence, to obtain a high-resolution image of a living animal, it would be necessary to deliver an unacceptably high whole-body x-ray dose. At the same time, the lower resolution required allows one to use a reduced magnification factor and, hence, a larger x-ray tube focal spot size, facilitating higher tube current and shorter exposures.

In its present use, imaging of small animals obtained with CT is used for complementing the functional information obtained by other modalities discussed above with anatomical information. However, the conjunction of two different and complementary imaging techniques is more than the sum of functional and anatomical information and constitutes a new imaging paradigm: the multimodality imaging.

7 Multimodality Approach

PET and SPECT are intrinsically non-morphological imaging techniques. However, in most cases, the shape of the body and of the organs of the animal can be visualized due to nonnegligible nonspecific uptake of the radiotracers. Nevertheless, in some cases, a rough anatomical

visualization is not sufficient, especially in bio-distribution studies, where it is necessary to exactly know the position of the radiotracer target. In addition, when quantitative information on small target sites is needed, the image suffers from an appreciable partial volume error that cannot be corrected without the knowledge of the target morphology. In this sense, it is obvious that the information from a morphological imaging technique, such as CT, is of great help for the PET or SPECT image analysis.

As a bonus for next-generation systems, there will be the possibility of easy integration with other modality imaging systems such as PET/CT (Fontaine et al. 2005), PET/SPECT (US Patent 6303935; Bartoli et al. 2007) or PET/MR (Mackewn et al. 2005). In fact, the combination of different imaging modalities offers the possibility to perform experiments more effectively than with a single modality alone, especially when the two modalities can be obtained within a short time with respect to each other (or simultaneously) and without moving the animal, possibly on the same gantry. In this sense PET/MR seems to be the perfect choice, combining the exquisite sensitivity of PET with the morphological/functional/spectroscopic high-resolution imaging of MR. The present major limitation on the development of simultaneous PET/MR scanners is the sensitivity of PMTs to magnetic fields. A reliable solution could be the development of magnetic-field-insensitive position-sensitive photodetectors such as the newly developed silicon photomultipliers (SiPM) (Golovin and Saveliev 2004).

7.1 PET/CT and SPECT/CT

On the shadow of the successful application of combined PET/CT scanners in the clinical environment, this technique has been recently transferred to small-animal scanners. In fact, the morphological information from CT can be used to get a finer spatial localization of the radiotracer distribution within the body as well as to obtain the attenuation coefficient map of the object under study for attenuation and scatter correction of the PET images. In clinical practice, PET/CT is mainly used in oncology studies, and it offers noticeable advantages for treatment planning. In small-animal imaging, the main advantages of combined PET/CT are the possibility to visualize the animal anatomy, especially in bio-distribution and oncology studies, and to perform partial volume correction for tracer quantification.

In addition, CT images can be used to improve the emission images. In fact, the emission images are affected by a quantitative error due to the attenuation of radiation by the object under study. Even if much smaller than for humans, the magnitude of this correction in small animals is non-negligible. For example, in PET, the attenuation correction factor is 45 for a 40 cm diameter man, 1.6 for a 5 cm diameter rat, and 1.3 for a 3 cm diameter mouse. This effect is usually corrected in small-animal PET scanners by obtaining a direct measure of the attenuation map of the object by means of transmission-based methods (Chow et al. 2005). The transmission images can be either obtained with a positron-emitting source (usually ^{68}Ge) rotating around the object by using the PET detectors (de Kemp and Nahmias 1994) or with a CT scanner (Kinahan et al. 1998). Although the PET-based method gives a direct measurement of the attenuation coefficient at 511 keV, the CT-based method has some advantages. In the CT case, the attenuation coefficients are measured with a continuous x-ray spectrum, ranging from 10 to 70 keV. Hence, for any tissue X, the CT-energy linear attenuation coefficient ($\mu_{CT,\,X}$) has to be scaled to the 511 keV value ($\mu_{PET,\,X}$). This is usually done with the formula (Mackewn et al. 2005):

$$\mu_{PET,X} = \frac{\mu_{CT,X} \times \mu_{PET,H_2O}}{\mu_{CT,H_2O}},$$

where μ_{PET,H_2O} and μ_{CT,H_2O} are the known linear attenuation coefficients of water at 511 keV and CT energy, respectively. This relation is valid for all the materials, such as soft tissues. In fact, the mass attenuation coefficient (μ/ρ) is remarkably similar for all non-bone materials since Compton scatter dominates for these materials. On the other hand, bone has a higher photoelectric absorption cross section due to the presence of calcium. Therefore, a different coefficient of correction is usually adopted for bone. The advantages of the CT-based attenuation correction method over the PET-based method include: lower statistical noise in transmission images acquired on CT versus PET, reduction of cross talk from PET annihilation photons and transmission photons during postinjection transmission studies, and reduction of the overall study time. Moreover, despite the difficulties in scaling the attenuation to 511 keV and the lack of a water correction for beam hardening, the PET emission images corrected with the CT-based method are as accurate as those corrected with the PET-based method (Chow et al. 2005).

7.2 PET/SPECT

Present multimodality imaging strategies are mostly based on the combination of complementary imaging techniques such as PET/CT and SPECT/CT, where functional and morphological information on the same subject can be derived and combined. On the shadow of the successful application of combined PET/CT scanners in the clinical environment, multimodality techniques have been recently transferred to small-animal scanners.

On the other hand, there is a growing interest for the preclinical application of nuclear imaging techniques able to perform dual-tracer imaging. It is defined as the application of two radionuclide-labelled tracers to image two different biological or molecular targets at the same time. In this way, a direct comparison between the two targets can be derived so as to provide additional diagnostic value that is difficult for a single-tracer imaging to provide. Dual-tracer imaging is now a well-established technique. It has been in use for over a decade for parathyroid, liver, and cardiac disease (e.g., in metabolic and perfusion studies for the assessment of myocardial viability), and studies in brain receptors and perfusion (for example, in the imaging of dopamine neurotransmission and brain perfusion in differential diagnosis of Parkinson's disease). Dual-isotope imaging is performed both in clinical and preclinical environments with SPECT by using two different isotopes with well-separated γ-ray emission energies. When injected together, the two isotopes can produce two different images that are obtained by applying different energy windows between the two primary photons. However, the main limitation of the dual-radionuclide imaging is the cross talk between the two radionuclides (Chang et al. 2006). Such cross talk is mainly due to the scattered (in the object, in the detector, or in the collimator) radiation from the higher-energy radionuclide that is detected in the lower-energy window or to the higher-energy secondary emission peaks from the lower-energy radioisotope that can be detected in the higher-energy window.

The following step would be the combination of the PET and SPECT for dual-tracer imaging that will potentially open new possibilities for the design of clinical and preclinical protocols. In this case, one single γ-emitting isotope is used in combination with a β^+-emitting isotope. For years, PET and SPECT were considered as antagonist techniques. In practice, each of the two modalities has peculiar advantages: PET can usually provide superior imaging performance in sensitivity and quantification compared to SPECT, and the PET tracers can offer several advantages with respect to SPECT tracers from the biological compatibility point of view; on the other hand, SPECT is not intrinsically limited in spatial resolution. In this view,

the combination of the two modalities is a real plus and can be technically advantageous in certain aspects as compared to conventional SPECT dual-tracer. A significant improvement in dual-tracer PET/SPECT imaging would be the possibility to also perform the two modalities simultaneously (simultaneous PET/SPECT).

In principle, a dual-tracer study with PET and SPECT can be also done with two separate scanners that can share the same type of animal bed holder. In fact, the animal can be subjected to a series of scans with the two modalities by transferring the animal bed from one scanner to the other once the animal is well fixed onto the bed. The image fusion for dual-tracer study can be done a posteriori by knowing the relative position of the animal with respect to the PET and SPECT fields of view.

On the other hand, a scanner, where PET and SPECT detectors are physically integrated with a common bed, will potentially ease the imaging procedure and speed up post-processing.

Dual-modality SPECT/PET systems are now available and a number of prototypes have been proposed in the past years. Present systems are very different in integration level and detector technology.

The simplest way to combine two imaging modalities is to physically integrate two separate imaging systems in a single gantry. For example, separate systems can be installed on the same gantry with nonoverlapping fields of view. For this reason, with this kind of systems PET and SPECT modalities cannot be performed at the same time. Another way to combine different imaging modalities is to juxtapose back to back two standalone systems in a way that they have the same axis and can share the same animal bed. Once combined, the two systems can be controlled by the same user console. The advantage of this solution, called "dockable," is a greater flexibility in exchange for a higher cost.

A further step toward a full combination of the SPECT and PET modalities is to have a solution where the two fields of view are overlapping. In this way, in addition to the possibility of performing both emission techniques on the same animal, it allows, in principle, the simultaneous execution of PET and SPECT investigation. To make it feasible, a possible solution is to combine two rotating dual-head parallel-plane detectors, where a pair is dedicated to PET and the other is for SPECT (Bartoli et al. 2007). An example of a solution using rotating detectors is shown in ❯ *Fig. 9*. In this case, two opposing heads (e.g., heads 1 and 2 in ❯ *Fig. 9*) are used for PET (coincidence mode), while the other pair (detectors 3 and 4) is equipped with the collimators and independently acquires single events (SPECT mode).

A more complicated solution is based on a dual-layer scintillator read out by a common photodetector (e.g., photomultiplier tubes or avalanche photodiodes). In this case, a first layer is optimized for the detection of medium energy γ rays commonly used in SPECT, while a second layer is optimized for the high-energy annihilation radiation used in PET. Such a detector module can be used as the basic detection unit in a multimodality scanner capable of performing PET and SPECT imaging simultaneously. Pulse-shape discrimination is used to separate the signals originating from the different detector materials. Good candidates for the first layer are high-light-output scintillators, while the second layer should be made of fast, high-density crystals. Examples of possible dual-layer scintillators for SPECT/PET are (first layer/second layer) YSO/LSO (Dahlbom et al. 1997; Saoudi and Lecomte 1999), LSO/GSO (Saoudi and Lecomte 1999), NaI/LSO (Schmand et al. 1998; Pichler et al. 2003), and YAP/LSO (Guerra et al. 2007). Another example is the modification of a PET system based on a ring geometry to perform SPECT and also simultaneous PET/SPECT using a slit-slat collimator insert (Shao et al. 2007).

PET and SPECT dual-tracer imaging, particularly the simultaneous acquisition, may open the door for many potentially new clinical and preclinical applications.

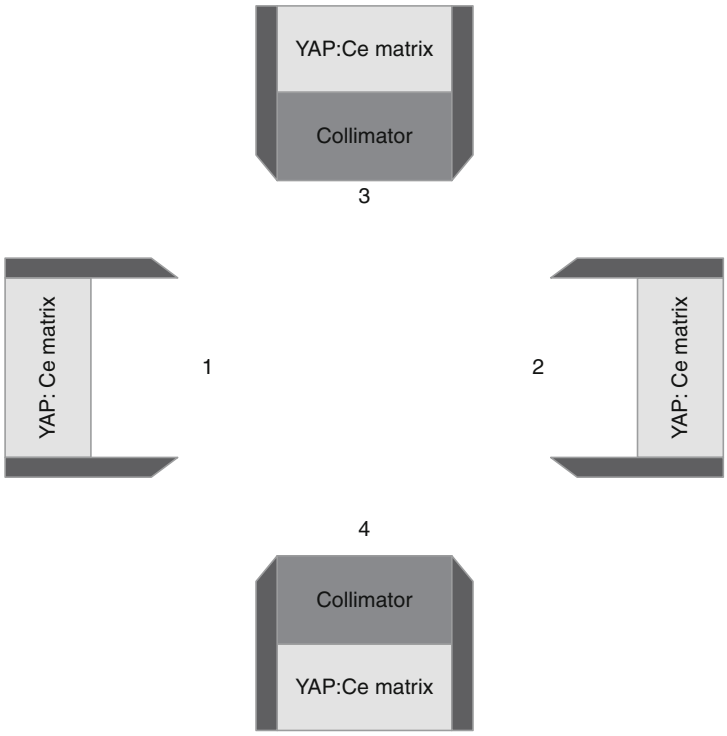


Fig. 9

Example of a simultaneous PET/SPECT solution using rotating detectors. Heads 1 and 2 are in time coincidence and are used for PET. Heads 3 and 4 are equipped with parallel hole collimators and are used for SPECT

Even if designed to integrate multimodality PET and SPECT imaging techniques, none of the above-mentioned systems are specifically designed for true simultaneous PET/SPECT acquisitions.

The requirements for simultaneous PET and SPECT go beyond the ability to simply detect single low-energy (for SPECT) and coincidence 511 keV (for PET) γ rays. The technical challenges for this imaging method are multifold and are related to the type of system in use. In fact, a possible solution for simultaneous PET/SPECT can use either a common detector working simultaneously in PET and SPECT (*joint modality*) or PET and SPECT can be performed using separate detectors either of the same type (*split modality*) (Bartoli et al. 2007) or using different technologies (*separate modality*).

A common challenge for any solution is related to the fact that the simultaneous PET/SPECT acquisition implies the simultaneous presence of single-photon- and positron-emitting isotopes. Hence, each modality should be able to work in presence on a secondary γ source. In particular, it is necessary to perform a correction procedure for the subtraction of the contribution of the down-scattered 511 keV γ rays in the energy range of the SPECT energy window (Bartoli et al. 2007).

8 Other High-Resolution Applications of Radiation-Imaging Instrumentation: Breast Cancer Investigation

In recent years, there has been an increasing focus on PET and SPECT systems designed for specific applications. One of such applications is the investigation of breast cancer with nuclear imaging techniques. The application of PET and SPECT techniques to breast cancer study is usually called Positron Emission Mammography (PEM) and Single Photon Emission Mammography (SPEM). Due to the different geometry and superior performance required with respect to whole-body scanners, these applications need the development of dedicated, higher-performance instrumentation. For example, for the PET case, thanks to the technological advances in high-resolution detectors for the small-animal imaging devices, many research groups are now working on the development of dedicated devices for PEM with high sensitivity and high spatial resolution (Freifelder and Karp 1997; Robar et al. 1997; Doshi et al. 2000; Lecoq and Varela 2002; Belcari et al. 2003) based on similar technologies. Although these two applications differ in optimal geometry and absolute performance requirements, both can gain from the advent of novel instrumentation for gamma-ray detection.

Some prototypes of PET systems for breast examination (PEM) are now available. They are typically accomplished with two parallel planar detectors positioned above and below the compressed breast (Robar et al. 1997; Doshi et al. 2000; Lecoq and Varela 2002). The main advantage of this configuration is the possibility to compress the breast. The higher solid-angle coverage and the reduced scattering improve the sensitivity and spatial resolution. Clinical trials have been performed with the dual-planar configuration. This configuration is well suited for the integration with conventional x-ray mammography units for emission–transmission image fusion or guided biopsy (Murthy et al. 2000).

Despite the good performance shown by the planar configuration with compression, the evolution of such a technique suffers from some intrinsic drawbacks. First of all, the compression, even if mild, limits the time duration of the examination, ultimately reducing the data statistics, hence the signal-to-noise ratio. In addition, the planar geometry does not allow a complete sampling of the FOV. In such a way, the image-reconstruction artifacts limit the spatial resolution and the signal-to-noise ratio in the final image, especially for smaller tumors. Other configurations are also being investigated (Freifelder and Karp 1997; Qi et al. 2002) to overcome such limitations. A possible solution proposed for obtaining pseudo-tomography acquisition is the translation of rectangular detectors in a fixed position (Weinberg et al. 2002). In the last years, some research groups have proposed solutions based on a pair of planar detectors rotating around the breast to fully exploit the superior spatial resolution of the tomography approach. This configuration could also allow the integration of PET with x-ray CT (Lamare et al. 2005).

On the other hand, single-photon emission technologies have been largely used in breast cancer investigation. The standard technique, introduced in early 1990s, is called scintimammography and utilizes a high-resolution gamma camera to obtain a single planar projection of a compressed breast (Khalkhali et al. 1994). In this case, the detector is usually made up of a scintillation camera comprising a pixelated matrix of scintillators (usually NaI:Tl) coupled to PS-PMTs with a typical active area of 15 cm × 20 cm. Examples of prototypes based on monolithic slabs (e.g., $LaBr_3$:Ce) have been also proposed (Pani et al. 2007). In addition to scintimammography, specialized tomographic scanners (SPECT scanners) have also been proposed

for breast cancer applications. High-resolution SPECT instruments dedicated to breast imaging have been designed in various designs each trying to best fit the breast size and geometry. Planar detectors can be arranged around the breast while rotating around the vertical axis with the patient in a prone position. In this case, the systems are called vertical axis of rotation (VAOR) SPECT. Dedicated VAOR systems should be advantageous since they permit a more accurate diagnosis of structures and patterns of scintimammography, while eliminating the hard compression of the breast (Tornai et al. 2004; Belcari et al. 2007). The small radius of rotation around the breast improves resolution-sensitivity tradeoff, and there is less attenuation and scatter between the breast and the camera as compared to more conventional horizontal axis of rotation (HAOR) SPECT scans (Metzler et al. 2002). Some prototypes of dedicated systems have been developed in the past years confirming the expected performance improvement with respect to present standard techniques (Loudos et al. 2004). An additional advantage of VAOR geometry is the possibility to integrate the SPECT system with a dedicated x-ray tomography scanner (Del Guerra et al. 2003). Other examples utilize arrays of collimated CdZnTe detectors tiled on either a cylindrical surface or a hemispherical surface surrounding the breast coupled to a stationary multiple pinhole collimated system (Singh and Mumcuoglu 1997; Tenney et al. 2009).

9 Summary

Dedicated high-resolution γ-ray detectors have found a successful field of application in small-animal scanners that are now well-established instruments. They are commercially available, and their development is still going on. The major trend in the next years will be a deep integration of different and complementary imaging techniques to have a more complete set of information from the anatomical, functional, and molecular point of view. Thanks to the recent advances, the γ-ray-detector technology is mature enough to make it possible to build dedicated instrument for breast cancer or other dedicated applications with adequate performance for the clinical use. The development of new scintillation materials and new photodetectors like SiPM could favor a further improvement of such dedicated instruments.

10 Cross-References

- Chapter 13, "Photon Detectors"
- Chapter 15, "Scintillation Counters"
- Chapter 17, "Gamma-Ray Detectors"
- Chapter 35, "Radiation-Based Medical Imaging Techniques: An Overview"
- Chapter 36, "CT Imaging: Basics and New Trends"
- Chapter 37, "SPECT Imaging: Basics and New Trends"
- Chapter 38, "PET Imaging: Basics and New Trends"
- Chapter 39, "Image Reconstruction"
- Chapter 44, "Simulation of Medical Imaging Systems: Emission and Transmission Tomography"

References

Accorsi R, Autiero M, Celentano L, Laccetti P, Lanza RC, Marotta M, Mettivier G, Montesi MC, Riccio P, Roberti G, Russo P (2004) Toward a Medipix2 coded aperture gamma microscope. IEEE Nucl Sci Symp Conf Rec 4:2461–2464

Auricchio N, Cesca N, Di Domenico G, Moretti E, Sabba N, Gambaccini M, Zavattini G, Andritschke R, Kanbach G, Schopper F (2005) SiliPET: design of an ultrahigh resolution small animal PET scanner based on stacks of semiconductor detectors. IEEE Nuc Sci Symp Conf Rec 5:3010–3013

Balcerzyk M, Kontaxakis G, Delgado M, Garcia-Garcia L, Correcher C, Gonzalez AJ, Gonzalez A, Rubio JL, Benlloch JM, Pozo MA (2009) Initial performance evaluation of a high resolution Albira small animal positron emission tomography scanner with monolithic crystals and depth-of-interaction encoding from a user's perspective. Meas Sci Technol 20:104011

Barrett HH (1972) Fresnel zone plate imaging in nuclear medicine. J Nucl Med 13:382–385

Bartoli A, Belcari N, Del Guerra A, Fabbri S (2007) Simultaneous PET/SPECT imaging with the small animal scanner YAP-(S)PET, 2007 IEEE nuclear science symposium conference record, Honolulu, HI, 26 Oct – 3 Nov. CD-ROM: M18-126. Paper copy: vol 5, pp 3408–3413. ISBN 1-4244-0923-3

Beekman FJ, Vastenhouw B (2002) Design and simulation of U-SPECT an ultra-citahigh resolution molecular imaging system. In: Conference proceedings of 2002 IEEE nuclear science symposium and medical imaging conference, Norfolk

Beekman FJ, van der Have F, Vastenhouw B, van der Linden AJA, van Rijk PP, Burbach JPH, Smidt MP (2005) U-SPECT-I: a novel system for submillimeter-resolution tomography with radiolabelled molecules in mice. J Nucl Med 46:1194–1200

Belcari N, Camarda M, Del Guerra A, Herbert D, Motta A, Vaiano A, Di Domenico G, Zavattini G (2003) Development of a planar head PEM system based on an array of PSPMT and YAP crystals. In: IEEE nuclear science symposium conference records, Portland, OR, vol 5, pp 2179–2182

Belcari N, Del Guerra A, Camarda M, Spontoni L, Vecchio S, Bennati P, Cinti MN, Pani R, Campanini R, Iampieri E, Lanconelli N (2007) Tomographic approach to single-photon breast cancer imaging with a dedicated dual-head camera with VAOR (SPEMT): detector characterization, 2007 IEEE nuclear science symposium conference record, Honolulu, HI, 26 Oct – 3 Nov. CD-ROM: M13-61. Paper copy: vol 4, pp 2901–2905. ISBN 1-4244-0923-3

Boone JM (2006) A geometric calibration method for cone beam CT systems. Med Phys 33:1695–1706

Chang CJ, Huang WS, Su KH, Chen JC (2006) Separation of two radionuclides in simultaneous dual-isotope imaging with independent component analysis. Biomed Eng Appl Basis Comm 18:264–269

Chatziioannou AF, Cherry SR, Shao Y, Silverman RW, Meadors K, Farquhar TH, Pedarsani M, Phelps ME (1999) Performance evaluation of microPET: a high-resolution lutetium oxyorthosilicate PET scanner for animal imaging. J Nucl Med 40: 1164–1175

Chow PL, Rannou FR, Chatziioannou AF (2005) Attenuation correction for small animal PET tomographs. Phys Med Biol 50(8):1837

Dahlbom M, MacDonald LR, Schmand M, Eriksson L, Andreaco M, Williams C (1997) A YSO/LSO phoswich array detector for single and coincidence photon imaging. IEEE Trans Nucl Sci 45(3):1128–1132

de Kemp RA, Nahmias C (1994) Attenuation correction in PET using single photon transmission measurements. Med Phys 21:771–7780

de Vree GA, Westra AH, Moody I, van der Have F, Ligtvoet KM, Beekman FJ (2005) Photon-counting gammacamera based on an electron-multiplying CCD. IEEE Trans Nucl Sci 52:580–588

Del Guerra A, Belcari N (2007) State-of-the-art of PET, SPECT and CT for small animal imaging. Nucl Instrum Meth A 583:119–124

Del Guerra A, de Notaristefani F, Di Domenico G, Gigante M, Pani R, Piffanelli A, Turra A, Zavattini G (1996) Use of a YAPCe matrix coupled to a position sensitive photomultiplier for high resolution positron emission tomography. IEEE Trans Nucl Sci NS-43:1958

Del Guerra A, Di Domenico G, Scandola M, Zavattini G (1998) High spatial resolution small animal YAP-PET. Nucl Instrum Meth A 409(1–3): 537–541

Del Guerra A, Di Domenico G, Fantini A, Gambaccini M, Milano L, Sabba N, Taibi A, Tartari A, Tuffanelli A, Zavattini G, Pani R, Pellegrini R, Soluri A, Cinti MN, Bevilacqua A, Bollini D, Gombia M, Lanconelli N, Arfelli F, Longo R, Olivo A, Pani S, Poropat P, Rigon L (2003) A dedicated system for breast cancer study with combined SPECT-CT modalities. Nucl Instrum Meth A 497:129–134

Derenzo SE, Moses WW (1993) Quantification of brain function. In: Uemura K et al (eds) Tracer kinetics and image analysis in brain PET. Elsevier, Amsterdam, pp 25–40

Doshi NK, Shao Y, Silverman RW, Cherry SR (2000) Design and evaluation of an LSO PET detector for breast cancer imaging. Med Phys 27:1535

Feldkamp LA, Davis LC, Kress JW (1984) Practical cone-beam algorithm. J Opt Soc Am A 6:612–619

Fontaine R, Belanger F, Cadorette J, Leroux J-D, Martin J-P, Michaud J-B, Pratte J-F, Robert S, Lecomte R (2005) Architecture of a dual-modality, high-resolution, fully digital positron emission tomography/computed tomography (PET/CT) scanner for small animal imaging. IEEE Trans Nucl Sci NS-52:691–696

Freifelder R, Karp JS (1997) Dedicated PET scanners for breast imaging. Phys Med Biol 42:2463–2480

Funk T, Parnham KB, Patt BE, Li J, Iwanczyk JS, Iwata K, Hwang AB, Hasegawa BH (2003) A new CdZnTe-based gamma camera for high resolution pinhole SPECT. In: IEEE NSS-MIC proceedings, vol 4, pp 2320–2324

Furenlid LR, Wilson DW, Chen Y-C, Kim H, Pietraski PJ, Crawford MJ, Barrett HH (2004) FastSPECT II: a second-generation high-resolution dynamic SPECT imager. IEEE Trans Nucl Sci 51: 631–635

Golovin V, Saveliev V (2004) Novel type of avalanche photodetector with Geiger mode operation. Nucl Instrum Meth A 518:560–564

Guerra P, Rubio JL, Ortuño JE, Kontaxakis G, Ledesma MJ, Santos A (2007) Performance analysis of a low-cost small animal PET/SPECT scanner. Nucl Instrum Meth Phys Res A 57:98–101

Gullberg GT, Tsui BMV, Crawford CR, Edgertorn ER (1987) Estimation of geometrical parameters for fan beam tomography. Phys Med Biol 32:1581–1594

Herbert DJ, Saveliev V, Belcari N, DAscenzo N, Del Guerra A, Golovin A (2006) First results of scintillator readout with silicon photomultiplier. IEEE Trans Nucl Sci NS-53(1):389

Joung J, Miyaoka RS, Kohlmyer SG, Lewellen TK (2001) Investigation of bias-free positioning estimators for the scintillation cameras. IEEE Trans Nucl Sci NS-48:715–719

Kastis GA, Furenlid LR, Wilson DW, Peterson TE, Barber HB, Barrett HH (2004) Compact CT/SPECT small-animal imaging system. IEEE Trans Nucl Sci 51:63–67

Khalkhali I, Mena I, Diggles L (1994) Review of imaging techniques for the diagnosis of breast cancer: a new role of prone scintimammography using technetium-99 m sestamibi. Eur J Nucl Med 21:357–363

Kinahan PE, Townsend DW, Beyer T, Sashin D (1998) Attenuation correction for a combined 3D PET/CT scanner. Med Phys 25:2046–2053

Laforest R, Longford D, Siegel S, Newport DF, Yap J (2004) Performance evaluation of the microPET-Focus – F120. IEEE Nucl Sci Symp Conf Rec 5:2965–2969

Lamare F, Bowen SL, Visvikis D, Cortes P, Wu Y, Tran V-H, Boone JM, Cherry SR, Badawi RD (2005) Design simulation of a rotating dual-headed PET/CT scanner for breast imaging. IEEE Nucl Sci Symp Conf Rec 3:1524–1529

Lecoq P, Varela J (2002) Clear-PEM, a dedicated PET camera for mammography. Nucl Instrum Meth A 486:1–6

Levine ZH, Kalukin AR, Frigo SP, McNulty I, Kuhn M (1999) Tomographic reconstruction of an integrated circuit interconnect. Appl Phys Lett 74:150

Liu ZL, Kastis GA, Stevenson GD, Barrett HH, Furenlid LR, Kupinski MA, Patton DD, Wilson DW (2002) Quantitative analysis of acute myocardial infarct in rat hearts with ischemia-reperfusion using a high-resolution stationary SPECT system. J Nucl Med 43:933–939

Llosa G, Belcari N, Bisogni MG, Marcatili S, Collazuol G, Piemonte C, Barrillon P, Bondil-Blin S, De La Taille C, Del Guerra A, Lacasta C (2009) Monolithic 64-channel silicon photomultiplier matrices for small animal PET. 2009 IEEE nuclear science symposium conference record, Orlando, FL, 25–31 Oct 2009. CD ROM: M5-91

Loudos GK, Giokaris ND, Mainta K, Sakelios N, Stiliaris E, Karabarbounis A, Papanicolas CN, Spanoudaki V, Nikita KS, Uzunoglu NK, Archimandritis SC, Varvarigou AD, Stefanis KN, Majewski S, Weisenberger A, Pani R, Maintas D (2004) High-resolution and high-sensitivity SPECT imaging of breast phantoms. Nucl Instrum Meth A 527:97–101A

Mackewn JE, Strul D, Hallett WA, Halsted P, Page RA, Keevil SF, Williams SCR, Cherry SR, Marsden PK (2005) Design and development of an MR-compatible PET scanner for imaging small animals. IEEE Trans Nucl Sci NS-52:1376

Massoud TF, Gambhir SS (2003) Molecular imaging in living subjects: seeing fundamental biological processes in a new light. Genes Dev 17: 545

McElroy DP, MacDonald LR, Beekman FJ, Wang Y, Patt BE, Iwanczyk JS, Tsui BMW, Hoffman EJ (2002) Performance evaluation of A-SPECT: a high resolution desktop pinhole SPECT system for imaging small animals. IEEE Trans Nucl Sci 49:2139–2147

Meikle SR, Fulton RR, Eberl S, Dahlbom M, Wong K-P, Fulham MJ (2001) An investigation of coded aperture imaging for small animal SPECT. IEEE Trans Nucl Sci 48:816–821

Meikle SR, Kench P, Weisenberger AG, Wojcik R, Smith MF, Majewski S, Eberl S, Fulton RR, Rosenfeld AB, Fulham MJ (2002) A prototype coded aperture detector for small animal SPECT. IEEE Trans Nucl Sci 49:2167–2171

Metzler SD, Bowsher JE, Tornai MP, Pieper BC, Peter J, Jaszczak RJ (2002) SPECT breast imaging combining horizontal and vertical axes of rotation. IEEE Trans Nucl Sci 49:31

Miyaoka R, Laymon C, Janes M, Lee K, Kinahan P, Lewellen T (2002) Recent progress in the development of a micro crystal element (MiCE) PET system. IEEE Nucl Sci Symp Conf Rec 2: 1287–1291

Moehrs S, Belcari N, Del Guerra A, Herbert DJ, Mandelkern MA, Motta A, Saveliev V (2004) A small-animal PET design using SiPM and anger logic with intrinsic DOI. 2004 IEEE nuclear science symposium conference record, Rome, Italy, October 16–22, vol 6, pp 3475–3479. ISBN 0-7803-8701-5

Moses WW, Derenzo SE (1994) Design studies for a PET detector module using a PIN photodiode to measure depth of interaction. IEEE Trans Nucl Sci NS-41:1441

Murthy K, Aznar M, Bergman AM, Thompson CJ, Robar JL, Lisbona R, Loutfi A, Gagnon JH (2000) Radiology 215:280–285

Nagarkar V, Gordon JS, Vasile S, Gothoskar P, Hopkins F (1996) High resolution X-ray sensor for non-destructive evaluation. IEEE Trans Nucl Sci 43:1559–1563

Olcott PD, Talcott JA, Levin CS, Habte F, Foudray AMK (2005) Compact readout electronics for position sensitive photomultiplier tubes. IEEE Trans Nucl Sci NS-52:21–27

Panetta D, Belcari N, Del Guerra A, Moehrs S (2008) An optimization-based method for geometrical calibration in cone-beam CT without dedicated phantoms. Phys Med Biol 53(14): 3841–3861

Pani R, Pellegrini R, Betti M, De Vincentis G, Cinti MN, Bennati P, Vittorini F, Casali V, Mattioli M, Orsolini Cencelli V, Navarri F, Bollini D, Moschini G, Iurlaro G, Montani L, de Notaristefani F (2007) Clinical evaluation of pixellated NaI:Tl and continuous $LaBr_3$:Ce, compact scintillation cameras for breast tumors imaging. Nucl Instrum Meth Phys Res A 571: 475–479

Paulus MJ, Gleason SS, Kennel SJ, Hunsicker PR, Johnson DK (2000) High resolution X-ray computed tomography: an emerging tool for small animal cancer research. Neoplasia 2(1–2): 62–70

Pichler BJ, Bernecker F, Boning G, Rafecas M, Pimpl W, Schwaiger M, Lorenz E, Ziegler SI (2001) A 4×8 APD array, consisting of two monolithic silicon wafers, coupled to a 32-channel LSO matrix for high-resolution PET. IEEE Trans Nucl Sci 48(4):1391–1396

Pichler BJ, Gremillion T, Ermer V, Schmand M, Bendriem B, Schwaiger M, Ziegler SI, Nutt R, Miller SD (2003) Detector characterization and detector setup of a NaI-LSO PET/SPECT camera. IEEE Trans Nucl Sci 50(5):1420–1427

Pichler BJ, Swann BK, Rochelle J, Nutt RE, Cherry SR, Siegel SB (2004) Lutetium oxyorthosilicate block detector readout by avalanche photodiode arrays for high resolution animal PET. Phys Med Biol 49:4305

Popov V, Majewski S, Weisenberger AG, Wojcik R (2001) Analog readout system with charge division type output. IEEE Nucl Sci Sym Conf Rec 4:1937

Qi J, Kuo C, Huesman RH, Klein GJ, Moses WW, Reutter BW (2002) Comparison of rectangular and dual-planar positron emission mammography scanners. IEEE Trans Nucl Sci NS-49: 2089–2096

Robar JL, Thompson CJ, Murthy K, Clancy R, Bergman AM (1997) Construction and caliberation of detectors for high resolution metabolic breast cancer imaging. Nucl Instrum Meth A 392:402

Rouze CN, Hutchins GD (2003) Design and characterization of IndyPET-II: a high-resolution, high-sensitivity dedicated research scanner. IEEE Trans Nucl Sci NS-50:1491–1497

Saoudi A, Lecomte R (1999) A novel APD-based detector module for multi-modality PET/SPECT/CT scanners. IEEE Trans Nucl Sci 46(3):479–484

Saoudi A, Pepin CM, Dion F, Bentourkia M, Lecomte R, Andreaco M, Casey M, Nutt R, Dautet H (1999) Investigation of depth-of-interaction by pulse shape discrimination in multicrystal detectors read out by avalanche photodiodes. IEEE Trans Nucl Sci NS-46:462–467

Schmand M, Dahlbom M, Eriksson L, Casey ME, Andreaco MS, Vagneur K, Phelps ME, Nutt R (1998) Performance of a LSO/NaI(Tl) phoswich detector for a combined PET/SPECT imaging system. J Nucl Med 39:9P

Schramm NU, Ebel G, Engeland U, Schurrat T, Behe M, Behr TM (2003) High-resolution SPECT using multipinhole collimation. IEEE Trans Nucl Sci 50:315–320

Seidel J, Vaquero JJ, Siegel S, Gandler WR, Green MV (1999) Depth identification accuracy of a three layer phoswich PET detector module. IEEE Trans Nucl Sci 46:485–490

Shao Y, Yao R, Ma T, Manchiraju P (2007) Initial studies of PET-SPECT dual-tracer imaging. IEEE Nucl Sci Symp Conf Rec 6:4198–4204

Singh M, Mumcuoglu E (1997) Design of a CZT based breast SPECT system. IEEE Nucl Sci Sym Conf Rec 2:1150–1154

Stock SR (2008) Micro computed tomography. CRC Press, Boca Raton

Tai Y-C, Chatziioannou AF, Yang Y, Silverman RW, Meadors K, Siegel S, Newport DF, Stickel JR, Cherry SR (2003) MicroPET II: design, development and initial performance of an improved microPET scanner for small-animal imaging. Phys Med Biol 48:1519

Tenney CR, Egger AB, McCurley JW, Dhah HK (2009) A staggered array of pinhole cameras for dedicated breast SPECT. 2009 IEEE nuclear science symposium conference record (NSS/MIC), pp 3586–3588

Tornai MP, Brzymialkiewicz CN, Cutler SJ, Madhav P (2004) Comparison of scintimammography and dedicated emission mammotomography. In: Conference records of 2004 IEEE/NSS/MIC, Roma, pp 2818–2822

US Patent 6303935 Combination PET/SPECT nuclear imaging system. Engdhal JC, Rago A

Volterrani D et al (eds) (2010) Fondamenti di Medicina Nucleare. Springer, Italy

Weinberg IN, Stepanov PY, Beylin D, Zavarzin V, Anashkin E, Lauckner K, Yarnall S, Doss M, Pani R, Adler LP (2002) PEM-2400 – a biopsy-ready PEM scanner with real-time X-ray correlation capability. IEEE Nucl Sci Symp Conf Rec 2: 1128–1130

Weisenberger AG, Wojcik R, Bradley EL, Brewer P, Majewski S, Qian J, Ranck A, Saha MS, Smith K, Smith MF, Welsh RE (2003) SPECT-CT system for small animal imaging. IEEE Trans Nucl Sci 50:74–79

Zanzonico P (2004) Positron emission tomography: a review of basic principles, scanner design and performance, and current systems. Semin Nucl Med 34(2):87

Zavattini G, Del Guerra A (2004) Small animal scanners. In: Del Guerra A (ed) Ionizing radiation detectors for medical imaging. World Scientific, Singapore, pp 385–464

Zavattini G, Del Guerra A, Cesca N, Di Domenico G, Gambaccini M, Moretti E, Sabba N (2005) High Z and medium Z scintillators in ultra-high-resolution small animal PET. IEEE Trans Nucl Sci NS-52(1):222–230

Zhu S, Tian J, Yan G, Qin C, Feng J (2009) Cone beam micro-CT system for small animal imaging and performance evaluation. Int J Biomed Imaging 2009:960573

46 Imaging Instrumentation and Techniques for Precision Radiotherapy

Katia Parodi[1,3] · *Christian Thieke*[2,3]
[1]Heidelberg Ion Beam Therapy Center, Heidelberg, Germany
[2]German Cancer Research Center, Heidelberg, Germany
[3]University Clinic Heidelberg, Heidelberg, Germany

1 Introduction 1154

2 Imaging for Treatment Planning 1155
2.1 Biological Imaging, Dose Painting 1157

3 Imaging for Image-Guided Radiotherapy 1158
3.1 Image Guidance in Photon Therapy 1160
3.2 X-ray-Based Image Guidance in Ion Therapy 1161
3.3 Ion-Based Image Guidance in Ion Therapy 1161
3.3.1 Ion Radiography 1163
3.3.2 Ion Tomography 1165

4 Imaging for Dose-Guided Radiotherapy 1166
4.1 Dose Reconstruction in Photon Therapy 1167
4.2 Range Monitoring and Dose Reconstruction in Ion Therapy 1168

5 Conclusion 1172

References 1174

Further Reading 1177

C. Grupen, I. Buvat (eds.), *Handbook of Particle Detection and Imaging*,
DOI 10.1007/978-3-642-13271-1_46, © Springer-Verlag Berlin Heidelberg 2012

Abstract: Over the last decade, several technological advances have considerably improved the achievable precision of dose delivery in radiation therapy. Clinical exploitation of the superior tumor-dose conformality offered by modern radiotherapy techniques like intensity-modulated radiotherapy and ion beam therapy requires morphological and functional assessment of the tumor during the entire therapy chain from treatment planning to beam application and treatment response evaluation. This chapter will address the main rationale and role of imaging in state-of-the-art external beam radiotherapy. Moreover, it will present the status of novel imaging instrumentation and techniques being nowadays introduced in clinical use or still under development for image guidance and, ultimately, dose guidance of precision radiotherapy.

1 Introduction

The main goal of external beam radiotherapy is to deliver a lethal *dose* (i.e., absorbed energy per mass unit) to the tumor cells while limiting as much as possible the radiation burden to the traversed healthy tissue and nearby critical structures. This task becomes particularly challenging for inoperable, deep seated tumors growing in close proximity of radiosensitive organs. In these cases, excellent conformation of the dose delivery is mandatory in order to ensure successful eradication of the primary tumor without compromising the functionality of the surrounding organs at risk. In less critical cases, precise conformation of the treatment with a limited amount of primary and secondary radiation to the healthy tissue is still highly desirable in order to reduce the risk of secondary cancer induction, which is a topic of recently increasing concern especially with regard to pediatric tumors (Xu et al. 2008).

Over the last decades, several technological advances have dramatically improved the precision of external beam radiotherapy. Three-dimensional (3D) inversely-optimized treatment planning and delivery of intensity-modulated photon irradiation from different beam angles (IMRT – intensity-modulated radiotherapy) have significantly improved the achievable tumor-dose conformality in daily clinical practice at most centers (Webb 2001). Latest innovations and developments aim to improve the precision and speed of fixed-field IMRT (i.e., individual fields delivered at fixed incident directions) via the introduction of continuous rotational irradiation (tomotherapy, arc therapy) or the usage of robotic linear accelerators (Cyberknife, Accuray Inc., Sunnyvale, CA) (Fenwick et al. 2006). In addition to these ongoing efforts based on conventional photon radiation, an exponentially increasing number of new facilities have been recently realized to introduce and investigate proton and carbon ion therapy in the clinic (Sisterson 2005), and also to advance its technology, e.g., via implementation of state-of-the-art scanning beam delivery (Haberer et al. 1993; Pedroni et al. 1999). The main rationale of ion beams for precision radiotherapy is their favorable energy deposition, which can be concentrated in a few millimeter narrow region (the *Bragg peak*) at an adjustable depth. This intrinsically enables superior physical selectivity in comparison to the depth-dose deposition of photon radiation in tissue which exhibits a maximum close to the entrance channel followed by an exponential attenuation. With respect to the most wide-spread used protons, carbon ions additionally offer improved lateral dose conformation because of reduced scattering, at the expense of a minor exit dose beyond the Bragg peak from light fragments. Their higher ionization density (LET – Linear

Energy Transfer) may enable improved treatment of radioresistant tumors in comparison to sparsely ionizing photons and protons due to a selective increase of biological effectiveness as well as a reduced sensitivity to the tumor oxygenation level for cell killing (see ❯ Chap. 47, "Tumor Therapy with Ion Beams").

The enhanced dose conformality attainable with the most modern radiotherapy techniques offers the possibility to achieve steep dose gradients between the tumor and the surrounding healthy tissue. This selectivity promises improved therapeutic outcome via a safe escalation of the dose precisely delivered to the tumor area, without involving nearby healthy tissues and critical structures. However, it can also cause adverse therapeutic results in the case of tumor miss and/or accidental exposure of surrounding organs at risk due to incorrect delivery of the intended dose during the fractionated treatment course, if patient positioning or other geometrical/anatomical variations occur.

The major sources of uncertainties for external beam treatment modalities using conventional radiation are related to the localization of the tumor target volume, the accuracy and reproducibility of the patient setup and immobilization, and the issue of organ motion for specific anatomical sites (e.g., lung, liver and prostate). Besides sharing the same sources of uncertainties as for conventional radiation modalities, ion beams pose more stringent issues due to the concept of *beam range*, i.e., the stopping point of the primary ions in the tumor which determines the position of the Bragg peak and thus the precise localization of the dose delivery.

To account for these uncertainties, cautious *safety margins* are added to the tumor volume when designing the treatment plan (cf. ❯ Sect. 2). However, reduction of these margins is a major goal of radiotherapy since margins are directly associated with excess toxicity and pose limitations to dose escalation for increased tumor control. Therefore, the advances in the achievable selectivity of dose delivery in modern IMRT and ion therapy techniques have been accompanied by an increasing role of imaging in the whole radiotherapy process. This includes precise identification of the target volume at the planning stage, evaluation of the patient geometry directly at the treatment site and, eventually, quantification of the actual dose delivery in comparison to the planned one as well as assessment of the treatment response for adaptive strategies.

2 Imaging for Treatment Planning

The availability of full-computerized, multimodal diagnostic tools for three dimensional (3D) imaging of the tumor target volume has been the prerequisite for the remarkable evolution of radiotherapy in the last decades, moving from classical 2D approaches to more precise 3D techniques that design the treatment based on detailed image-derived 3D models of the patient anatomy.

The possibility to import and co-register different imaging modalities in the treatment planning system (TPS) is essential for the first step of the planning process, namely delineation and segmentation of the patient-specific target volume and critical organs. Nowadays, state-of-the-art TPS are capable to merge volumetric information from the major morphological imaging modalities such as X-ray computed tomography (CT, see ❯ Chap. 36, "CT Imaging: Basics and New Trends") and non-ionizing magnetic resonance imaging (MRI, Lauterbur 1973).

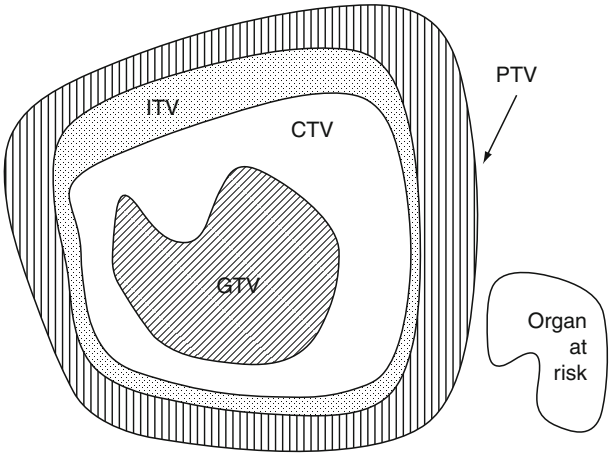


Fig. 1

Representation of the ICRU volumes for treatment planning, as defined in the text. *CTV* clinical target volume, *GTV* gross tumor volume, *ITV* internal target volume, *PTV* planning target volume (source: Podgorsak (2005))

Identification and segmentation of the relevant volumes for treatment planning is based on the recommendations of the ICRU international commission on radiation units and measurements, reports 50 (ICRU 1993) and 62 (ICRU 1999) (*Fig. 1*). The gross tumor volume (GTV) represents the gross demonstrable extent and location of malignant growth (ICRU 1993). The clinical target volume (CTV) is the tissue volume that contains a demonstrable GTV and/or additional sub-clinical microscopic malignant disease which has to be eliminated (ICRU 1993). The internal target volume (ITV) consists of the CTV plus an internal margin designed to take into account variations in size and position of the CTV relative to the patient's reference frame (usually defined by the bony anatomy), e.g., due to organ motions such as breathing or bladder filling (ICRU 1999). The planning target volume (PTV) is a geometrical concept which is designed to select the appropriate beam arrangements, taking into consideration the net effect of all possible geometrical variations including setup errors in order to ensure that the prescribed dose is actually absorbed in the CTV (ICRU 1993). The organ at risk (OAR) is finally an organ whose sensitivity to radiation is such that the dose received from a treatment may be significant compared with its tolerance.

Once the PTV and OARs are defined on the basis of the available images and safety margins according to the clinical experience, incident beam directions of the chosen radiation modality are selected on the basis of the image-based 3D model of the patient anatomy. The main criterion is to provide an optimal compromise between geometrical target coverage and sparing of organs at risk in the beam eye view (BEV). The determination of the treatment plan is then performed via inverse optimization techniques aiming to fulfill the prescribed dose coverage of the PTV without exceeding the dose constraints to the OARs. For this purpose the full-computerized dose calculation is typically performed on the morphological X-ray CT image acquired without contrast agents, so called *planning CT*. This is because the CT numbers or Hounsfield Units

(HU) of a native (i.e., without injected external agents) acquisition give information on the electron density distribution in the patient, which can be readily used for calculation of absorbed dose in photon therapy (Thomas 1999), or for the necessary determination of the ion beam range in the patient using semi-empirical CT-range calibration curves (Schaffner and Pedroni 1998; Rietzel et al. 2007, see ❯ Chap. 47, "Tumor Therapy with Ion Beams").

Evaluation of the treatment plan is based on the visual inspection of the calculated dose distributions superimposed onto the patient planning CT, as well as on the quantitative estimation of the target coverage and radiation burden to the organs at risk via histograms of the percentage of volume receiving a given percentage of the maximum dose for all the segmented image-based structures (*dose–volume histograms* DVH) (Drzymala et al. 1991). The finally approved treatment plan specifies the parameters of the therapeutic beam delivery for direct transfer to the control system of the linear accelerator or of the ion therapy unit. In precision radiotherapy, each individual patient plan is typically verified prior to the first day of treatment by means of dosimetric measurements in water or water-like phantoms in the framework of the medical physics quality assurance program.

2.1 Biological Imaging, Dose Painting

Although not yet supported by all planning systems, functional and molecular imaging modalities using different acquisition sequences (MRI-based imaging) or radioactive tracers (nuclear medicine imaging) can provide additional useful information for treatment planning. In particular, already clinically established techniques like functional MRI and nuclear medicine PET (Positron Emission Tomography, see ❯ Chap. 38, "PET Imaging: Basics and New Trends") and SPECT (Single Photon Emission Computed Tomography, see ❯ Chap. 37, "SPECT Imaging: Basics and New Trends") can help visualizing microscopic disease outside the macroscopic tumor as shown by the anatomical images. Moreover, PET imaging using standard and novel tracers can gain valuable information on the tumor at the molecular level, e.g., on its metabolism (tracer: fluorodeoxyglucose[^{18}F]-FDG), cell proliferation (tracer: 3′-18F-fluoro-3′-deoxy-L-thymidine [^{18}F]-FLT) and oxygenation (tracer: e.g., Fluoromisonidazole-[^{18}F]-FMISO). Advanced MR techniques like spectroscopy, dynamic perfusion and diffusion, as well as novel approaches like ultra-high magnetic fields up to 7 T and 9.4 T are currently being investigated with regard to their potential for visualizing biological information like cellular energy and lipid metabolism, blood perfusion and pathological damages of cell membranes (Olsen and Thwaites 2007). Functional imaging can be used to complement morphological imaging not only for improved delineation of the tumor volume with reduced inter- and intra-observer variability, but also for identification of more radioresistant tumor areas which may require increased dose delivery for successful tumor control (*dose painting*) (Ling et al. 2000; Olsen and Thwaites 2007). An example of a clinical case where tumor delineation may benefit from multimodal imaging is illustrated in ❯ *Fig. 2*. It should be noted that for reduction of geometrical uncertainties in treatment planning all images are acquired with the patient already immobilized in the treatment position.

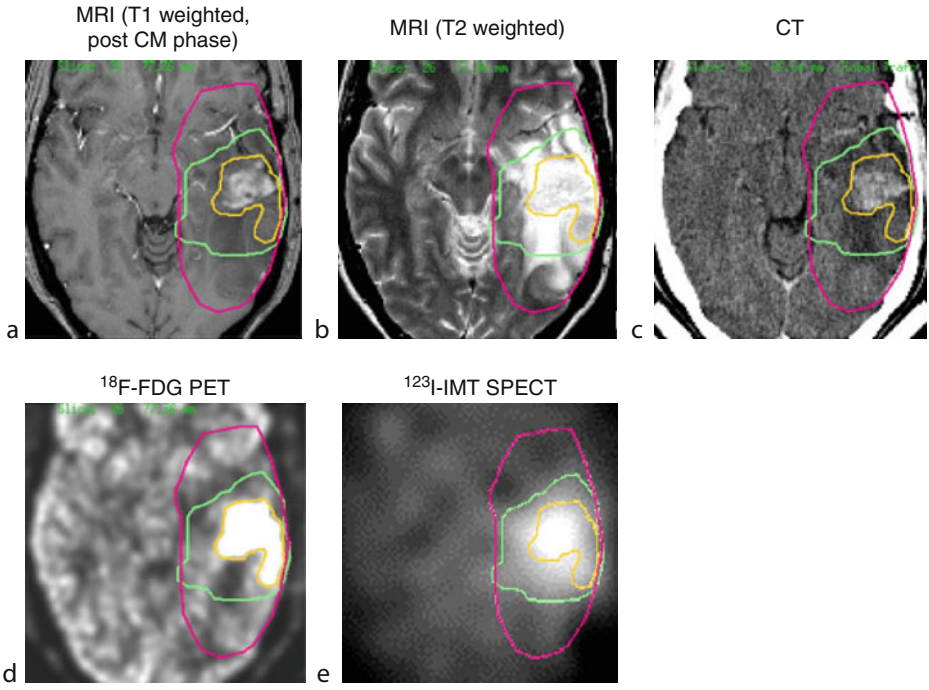


Fig. 2

Example of a patient suffering from a recurrent astrocytoma in the left temporal lobe. The target volume defined in the SPECT image (e) (*green outline*) represents the area of an elevated amino-acid transport and corresponds to the exact histological extension of the whole tumor. The smaller volume (*yellow outline*), showing an elevated glucose metabolism in the PET scan (d) corresponds to the most aggressive parts of the astrocytoma. In contrast, an exact definition of the target volume based on MRI is difficult, because it is underestimated in the T1 weighted image (a) as well as overestimated in the T2 weighted scan (b) (*red outline*). From Schlegel and Mahr (2001)

3 Imaging for Image-Guided Radiotherapy

The goal of image-guided radiotherapy (IGRT) is to reduce geometrical uncertainties and optimize the definition of the safety margins to enable trustworthy application of escalated doses tightly sculpted to the tumor volume for improved tumor control with reduced toxicity.

The most important sources of geometrical uncertainties include target delineation, phantom transfer errors, set-up errors and physiological changes. Indeed, identification of the target volume at the planning stage is still the major source of geometrical uncertainty, with delineation errors rarely below 3 mm (Korreman et al. 2010). In this respect, the potential of multimodal imaging for improved tumor target delineation has been addressed in the previous section.

The remaining challenge of IGRT is to ensure that the patient position at the daily treatment situation is a correct replica of the patient position and anatomy at the time of treatment planning, as well as to provide population-based estimation of geometrical uncertainties for optimization of the clinical safety margins in treatment planning.

Differences between the planning and treatment position can be first of all due to setup errors, which can be classified as *systematic* and *random*. Systematic errors are reproducible consistent errors which go in the same direction and are of similar magnitude over the complete course of fractionated treatment which can extend up to several weeks. Random errors unpredictably vary from day to day in both direction and magnitude. In addition to errors in the patient set-up, physiological changes with respect to the planning situation can occur during the course of fractionated radiotherapy. These physiological modifications may include *inter-fractional* changes between different days of treatment (e.g., tumor shrinkage), or *intra-fractional* changes within the same day of treatment (e.g., breathing motion).

Conventional patient positioning only based on the alignment of skin markers and in-room laser is indeed insufficient to guarantee the sub-millimeter level of precision and accuracy requested for high-precision radiotherapy. Moreover, it does not provide any information on internal anatomical changes. Therefore, in-room imaging of the patient has been increasingly introduced for image guidance of precision photon therapy to reduce errors from incorrect patient positioning or from anatomical modifications and displacements, as it will addressed in the next subsection.

In addition to the geometrical/anatomical verification of the patient at the treatment site, ion beam therapy poses even more stringent quality assurance requirements for IGRT. In particular, the critical quantity to be assessed is the finite ion beam range in the patient. This is established at the treatment planning stage using a facility- and CT-scanner-dependent calibration curve between HU numbers and water-equivalent ion range, basically reflecting the ion stopping power dependence in different tissues relative to water (cf. previous section and ❯ Chap. 47, "Tumor Therapy with Ion Beams"). Although this relationship is carefully determined via experimental range measurements using tissue-equivalent substitutes or real tissue samples, the achievable accuracy of the ion range calculation in the patient is estimated to be within 1–3% (Schaffner and Pedroni 1998; Rietzel et al. 2007), corresponding to approximately 1–3 mm uncertainty at typical depths of 10 cm. This is due to the non-unique correspondence between HU numbers and materials, in addition to unavoidable experimental uncertainties of the CT image (e.g., because of X-ray spectral changes depending on the imaged object size, known as *beam hardening effects*, Rietzel et al. 2007) and of the HU-range calibration curve (e.g., due to sample heterogeneities and HU partial volume effects). Besides intrinsic inaccuracies of the calibration curve, HU values can also be improperly assigned along the actual ion beam path due to artifacts in the planning CT (e.g., because of metallic implants, Jäkel and Reiss 2007) or the mentioned modifications of the patient position/anatomy in the treatment situation. Hence, deviations from the planned range can further increase up to ≈5–20% in the most unfavorable situations. Range uncertainties can be accounted for in clinical routine of ion therapy by using cautious safety margins and beam directions, e.g., avoiding to place the sharpest distal-dose fall-off in front of critical organs. However, image guidance of the actual ion beam range can be beneficial for promoting safe full clinical exploitation of the utmost tumor-dose conformality achievable with ion beams. Therefore, the ❯ Sect. 3.3 will also address unconventional imaging techniques in ion beam therapy for pre-treatment in-vivo verification of the patient position and beam range. Additional techniques based on the detection of a surrogate signal formed as by-product of the therapeutic irradiation will be described in the ❯ Sect. 4.



Fig. 3
Examples of X-ray-based image guidance in photon therapy. *Left panel*: In-Room CT scanner (kV, fan collimation) from Siemens. *Middle Panel*: Onboard CT (kV, cone beam) from Varian. *Right Panel*: Integrated CT (MV, fan collimation) from TomoTherapy (pictures taken from Korreman et al. (2010))

3.1 Image Guidance in Photon Therapy

In its most basic form, image guidance inside the treatment room is performed by acquiring single or orthogonal 2D images using either the treatment beam (portal images) or stationary X-ray sources. While being perfectly adequate for a broad range of clinical applications, such solutions do not provide the highest level of tumor localization accuracy necessary for critical challenges like curative treatment of moving targets or paraspinal tumors.

The current state of the art in in-room image guidance can be characterized as 3D/4D imaging of the patient using ionizing radiation. This group can be divided in solutions that use kV versus MV beams and cone-beam versus fan-beam collimation (Korreman et al. 2010). Solutions using kV sources typically provide better soft-tissue contrast than those using MV sources due to different absorption characteristics, and solutions using fan-beam collimation typically provide better image quality than cone-beam-based solutions due to less scatter and more efficient detector technologies. The highest image quality is provided by dedicated CT scanners installed in the treatment room sharing the same treatment couch with the linear accelerator (Thieke et al. 2006). Cone-beam solutions are mounted directly on the treatment gantry, eliminating the need for couch movement between imaging and treatment (Jaffray et al. 2002). Fan-beam collimated MV imaging is provided by the TomoTherapy treatment device (Mackie et al. 2003). Examples are shown in *Fig. 3*.

The decision for or against a particular imaging solution cannot be based on image quality alone. Besides technical aspects like different quality assurance requirements or maximum size of the treatment room, the specific clinical applications are the crucial factors that demand or rule out certain imaging techniques. For example, in-room CT imaging is best suitable for treating tumors with interfractional variations (e.g., head and neck tumors), but is by design not able to detect intrafractional motion like breathing motion.

In addition to the most widely available volumetric imaging by ionizing radiation, there are a number of specialized imaging and non-imaging solutions under investigation and commercially available that address specific treatment problems.

For certain skin-deep lesions like the breast, optical systems for 3D surface imaging can offer a valid alternative for checking the patient set-up without any additional exposure to ionizing

radiation (Bert et al. 2006). Similar setup corrections for internal displacement like in the case of the prostate can be evicted from non-ionizing ultrasound imaging prior to the treatment (Boda-Heggemann et al. 2008) or from radiofrequency transponders implanted in the patient and monitored during the treatment (Willoughby et al. 2006). For continuous, real-time monitoring of respiratory motion, non-ionizing external surrogates based on optical imaging like infrared reflectors/cameras (Saw et al. 2007) or pressure sensors placed on the patient abdomen/chest (Li et al. 2006) as well as air-volume measuring devices (spirometers) (Lu et al. 2005) are being used in clinical practice. These tools can be synchronized with appropriate beam delivery strategies like gating (Keall et al. 2006a) or beam tracking (Keall et al. 2006b) in order to ensure adequate tumor coverage over the whole breathing cycle. Correlation of the external surrogate with the internal tumor motion can be established prior to the treatment via synchronized dynamic acquisition of several CT images sorted out in different breathing phases (so called 4D CT, Rietzel et al. 2005) and verified during treatment via fluoroscopic imaging or oblique X-ray projections (Berbeco et al. 2005).

One of the most promising current developments in in-room image guidance is the integration of MR imaging in the treatment room. MRI has significant advantages over CT-based imaging like superior soft-tissue contrast, absence of ionizing radiation and the potential of imaging not only anatomy, but also biological function. The high magnetic field inside the treatment room presents several major technical challenges, e.g., for generating the treatment beam and calculating the correct treatment dose distribution. However, several institutions and vendors actively pursue this development (e.g., Raaymakers et al. 2004), and first clinical applications can be expected within the next few years.

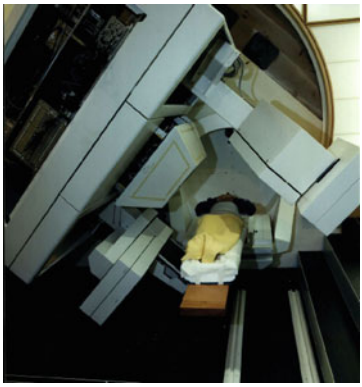
3.2 X-ray-Based Image Guidance in Ion Therapy

The increased sensitivity of ion beam therapy to density changes along the beam path makes the verification of the patient position and anatomy at the treatment site even more crucial than with photon therapy. Therefore, IGRT solutions for 2D and/or 3D X-ray kilovoltage imaging of the patient in the treatment position have been also implemented at ion beam therapy centers, similar to the systems already described for photon therapy. Examples of planar and volumetric imaging equipment integrated in ion beam therapy facilities are illustrated in ❯ *Fig. 4*. Dedicated solutions tailored to ion beam therapy include the development of vertical CT systems for volumetric in-room imaging of patients treated in seated position (Kamada et al. 1999), or the introduction of a hole in the last bending magnet of a novel proton therapy gantry in order to accommodate BEV X-ray portal imaging simultaneous with the proton beam (Pedroni 2009). In some implementations, time-resolved operation is supported to enable 4D capabilities, e.g., for fluoroscopic imaging of moving targets during ion beam delivery.

3.3 Ion-Based Image Guidance in Ion Therapy

Besides sharing similar X-ray kilovoltage IGRT instrumentation as established in photon therapy, the specific needs of in-vivo range verification have also prompted investigations on dedicated imaging techniques for ion beam therapy. Similar to portal imaging in photon therapy, the possibility of exploiting the same radiation source as used for treatment has been


a


b


c

Fig. 4

Examples of X-ray-based image guidance in ion beam therapy. The *top left* and *middle panel* (a and b) illustrate the clinical workflow implemented at the first proton gantry of the Paul Scherrer Institute (PSI website 2010): the patient is prepared and imaged with a dedicated CT system outside of the treatment room (*left panel*) and then transported into the proton treatment room, where additional checks can be performed by taking X-ray images of the patient in his treatment position using integrated instrumentation with retractable mechanical holders (*middle panel*). The *lower panel* (c) illustrates the C-arm imager (Siemens AG) installed at the new Heidelberg Ion Beam Therapy Center (Haberer et al. 2004) for in-room patient position verification via radiographic and maybe, eventually, tomographic (cone-beam CT) X-ray imaging. This illustrative picture does not depict the patient fixation device used in clinical practice

explored in ion beam therapy as well. However, primary therapeutic ions are completely stopped in the patient, in contrast to the more penetrating photons. Therefore, ion-based imaging methods can only rely on the detection of transmitted beams having higher energies than those used for therapeutic treatment.

The possibility of using energetic ion beams capable to traverse the patient has been proposed since the late 1960s (Koehler 1968; Sommer et al. 1978) for obtaining low-dose radiographic imaging at high density resolution, complementary or alternative to X-ray imaging for

localization of the tumor and verification of the patient position at the treatment site. However, it soon became clear that the technique has also the potential to provide pre-treatment information on the patient-specific stopping properties for indirect in-vivo assessment of the therapeutic ion beam range and, possibly, individualized refinement of the HU-range calibration curve. Ultimately, tomographic reconstruction of the ion stopping power map in the patient is envisaged to replace the X-ray CT for treatment planning calculations, thus directly eliminating the uncertainties connected with the usage of semi-empirical HU-range calibration curves. However, due to remaining technical challenges ion radiography and tomography are still methods under investigation and not yet established in clinical routine. Basic ideas and ongoing developments are summarized in the following.

3.3.1 Ion Radiography

Radiographic imaging with energetic ion beams provides a 2D map of the mean residual range of the transmitted ions behind the object of interest. Due to the weak energy dependence of the ion stopping power ratio in an arbitrary medium relative to water, information on the mean residual ion range gained at higher initial beam energies than the therapeutic ones can still be used for indirect verification of the planning HU-range calibration curve integrated all along the path in the patient (Schneider and Pedroni 1995; Schneider et al. 2005). Moreover, the radiographic measurement can also convey information on the actual distribution of ion beam residual ranges, exhibiting a natural broadening around the mean value due to the initial beam momentum spread, the statistical fluctuations of the energy loss as well as the spread (due to small lateral deflections) of multiple ion paths in the traversed tissue yet sharing the same stopping location. Whereas the first two processes can be approximated by Gaussian distributions, the latter is clearly influenced by the different stopping properties of the medium encountered along the different ion paths. Therefore this broadening, also referred to as *range dilution*, provides a quantitative measure of tissue inhomogeneities which can be exploited, e.g., at the stage of treatment planning for selection of beam incidence angles more robust against ion range uncertainties (Schneider et al. 2004). Finally, the higher density resolution of ion beams enables obtaining morphological information on the patient geometry with higher contrast and lower (e.g., by a factor of 50–100 as reported in Schneider et al. 2004) doses of irradiation than X-ray radiography. This translates in an improved capability of soft tissue differentiation, e.g., for tumor localization and patient positioning, though likely at the expenses of inferior spatial resolution depending on the used ion beam source and imaging implementation.

From the technological point of view, the method requires an accelerator source capable to deliver sufficiently high ion beam energies in order to completely traverse the patient. This condition is typically met for the anatomical locations of interest by most operational ion therapy facilities, offering highest beam energies corresponding to approximately 30 cm penetration in water (Rinaldi et al. 2010). A large irradiation area can be obtained for both delivery techniques with scattered (broad-beam or cone-beam configuration) or scanned ion beams. The detection system must be capable of measuring the residual ion beam range or energy. This is typically achieved either with a range telescope, i.e., a stack of detector channels like scintillation plates (Schneider et al. 2004) or ionization chambers interleaved by absorber plates (Brusasco et al. 2000) for a direct measurement of the Bragg-peak positions, or a thick energy detector like a plastic scintillator (Shinoda et al. 2006) or a crystal calorimeter (Petterson et al. 2007) to directly

score the residual beam energy. Both these approaches postulate the capability of the detection system to completely stop the primary and initially mono-energetic ion beams. In some situations this may require the usage of additional absorbers like range shifters in order to compensate for too thin areas of the sample to be traversed. Investigations on alternative methods using thinner detectors like scintillation screens read out by cameras for intensity measurements of properly energy-modulated impinging beams (Ryu et al. 2008) or amorphous silicon detectors for recovery of the emerging ion energies from energy loss measurements of initially mono-energetic beams (Engelke et al. 2010) are ongoing. For improved imaging performances, information of the ion tracks can be obtained by complementing the chosen detector system with the identification of the entrance and exit coordinates and even corresponding directions (especially for cone-beam configurations) of the beam – ideally at the single ion level – in two or four planes in front and behind the patient. This can be realized for example with position-sensitive scintillating fiber hodoscopes (Schneider et al. 2004) or silicon strip detectors (Petterson et al. 2007). The positional and directional information recorded in coincidence with the range or energy measurement can be then used for reconstructing the radiographic images at a selected depth in the object with improved spatial resolution. In addition, some tracking detectors like the silicon strips may also enable retrieving information on the actual single ion energy at the entrance and exit of the patient for improved density resolution. Regardless of the chosen implementation, the radiographic imaging systems typically require a calibration procedure by means of a reference scan, i.e., a measurement acquired without the object to be imaged (blank scan) or in pure water. The mean range loss in the patient (absolute or relative to water) can be finally quantified by comparison of the actual transmission measurement with the patient in place with respect to the calibration one.

For proton beams the major complication arises from multiple Coulomb scattering, making the detection of the individual protons at the entrance and exit of the patient essential for more accurate reconstruction in the selected image plane. In particular, usage of the most likely proton trajectory instead of a simplified straight line approximation of the proton path is strongly recommended in order to improve the achievable spatial resolution.

For heavier ions like carbon the problem of lateral deflection from multiple Coulomb scattering is almost negligible. However, fragmentation of the high-energy beam in the patient or the detector can introduce a background to the measurement due to the production of additional emerging radiation besides the transmitted primary ion beam. Nevertheless, this background signal typically does not affect Bragg-peak identification in range telescopes. Moreover, it can be reduced or suppressed by anticoincidence techniques for set-ups using thick energy detectors. The basic idea is to add an anticoincidence detector surrounding the energy detector for elimination of undesirable events with projectile fragments carrying some of the primary beam energy out of the energy detector (Shinoda et al. 2006). For other implementations, e.g., exploiting energy loss measurements in thin detectors, the feasibility of fragment separation for background reduction or suppression still needs to be investigated.

So far, pioneering pre-clinical studies of ion radiography have been reported for scanned proton beams at PSI, Switzerland, using the retractable apparatus of ❯ *Fig. 5* completely integrated into the first operational proton beam gantry. Although the method has not yet reached clinical application, the feasibility of indirect in-vivo range verification and of customization of the HU-range calibration curve could be demonstrated for animal patients as reported in Schneider et al. (2004). Currently, efforts are being pursued by several institutions worldwide through the realization and optimization of small-scale prototypes using new-generation detectors in preparation of larger-scale systems for likely future clinical use.

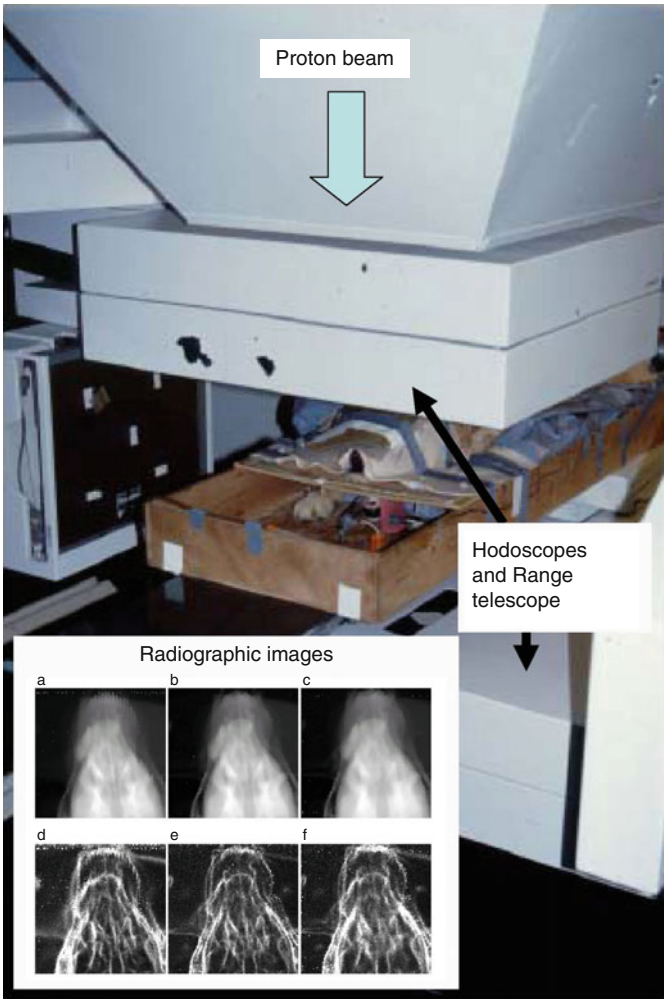


Fig. 5
Proton radiography prototype developed at the Paul Scherrer Institute, consisting in two position-sensitive hodoscopes (above and beneath the patient) and the distal range telescope beneath. The system, completely integrated into the proton beam gantry, has been successfully tested with animal patients. The insert shows images of proton range (*top* row) and range dilution (*bottom* row) reconstructed at 5%, 30% and 50% (from *left* to *right*) of the object thickness for a dog patient treated for a nasal tumor (Schneider et al. 2004)

3.3.2 Ion Tomography

Ion radiographic imaging is a promising technique to control the range of the ion beams, to detect range dilution and to verify the correct positioning of the patient for a selected projection.

The natural extension to 3D by proper acquisition and reconstruction of different radiographic projections has the appeal to retrieve volumetric information on the electronic density and ion stopping power properties of the imaged object. Indeed, investigations on proton computer tomography (pCT) started over four decades ago showing certain advantages over diagnostic X-ray CT, such as higher accuracy of electron density reconstruction and lower dose at the same density resolution (Hanson et al. 1981). However, the limitations of image quality due to energy loss straggling and multiple Coulomb scattering, together with the inherent higher costs of dedicated proton accelerators and beam transport systems, slowed down the diagnostic development of pCT in favor of the less expensive X-ray CT imaging. Similar considerations apply to the usage of the even less available and more expensive energetic heavier ions.

Recently, the worldwide rapid spreading of ion beam therapy facilities with rotating beam delivery systems (proton gantries) together with the advances in detector technologies and reconstruction techniques have renewed the interest in ion tomography, however shifting the application field from diagnostics to replacement of the X-ray CT for ion treatment planning as well as IGRT in the actual treatment position. First very encouraging experimental results could be already demonstrated by different groups for application of small-scale radiographic prototypes to tomography, mostly via rotation of the phantom rather than the beam source and detector system (e.g., Shinoda et al. 2006; Petterson et al. 2007). In particular, the advances in the detector technology including ion tracking (even at the single particle level) and the implementation of modern iterative or algebraic reconstruction techniques combined with the concept of most likely path (MLP, especially important for the more scattering proton beams) have demonstrated the potential to overcome the traditional imaging limitations encountered in the pioneering investigations back to the end of the 1960s. Moreover, the increasing availability of ion beam therapy sources, including the new developments of heavy-ion beam gantries (Weinrich 2006) will likely further promote the technique to eventual clinical application.

4 Imaging for Dose-Guided Radiotherapy

Whereas the main goal of IGRT is to reduce geometrical/anatomical uncertainties and, for ion beam therapy, also to enable pre-treatment range verification, the ultimate goal of imaging for precision radiotherapy is to enable an in-vivo non-invasive verification of the actual dose delivery with respect to the planned one. This implies an integral check of the whole chain from treatment planning to patient positioning and beam delivery, independent from all the pre-treatment quality assurance programs. Assessment of the actual dose delivery in the patient could be then directly used during treatment to prevent gross irradiation errors (*online monitoring*), and immediately after each fraction for adaptation of the subsequent dose delivery in order to guarantee correct application of the total dose prescription over the entire treatment course (*dose-guided radiotherapy* DGRT). Being related to the interaction of the primary radiation with the irradiated tissue, DGRT can only be performed using the primary therapeutic beam or a secondary surrogate signal induced by the therapeutic irradiation. The underlying basic principles together with the main implementations and ongoing investigations for precision photon and ion beam therapy are described in the following.

4.1 Dose Reconstruction in Photon Therapy

The intrinsic physical properties of the penetrating photon beams enable in-treatment detection of the radiation traversing the patient via electronic portal imaging devices EPIDs (Antonuk 2002). Although the original purpose of portal imaging was confirmation of the treatment position via analysis of 2D projections of the transmitted irradiation field with respect to visible landmarks like implanted markers or bony structures, novel attempts to validate and even extrapolate the dose delivered to the patient from the detected residual radiation have been proposed and pursued over the last few years (van Elmpt et al. 2008). A major step forward in these developments has been the ability to convert portal images into portal-dose measurements (Nijsten et al. 2007a) via non-trivial understanding and calibration of the signal acquired by a new generation of EPIDs (typically amorphous silicon), including the possibility to separate the transmitted primary radiation from the radiation scattered in the patient and in the detector.

The basic verification strategies investigated so far involve either forward or backward dose calculation approaches. The first so-called "2D transit dose verification" is based on the comparison of the two-dimensional portal-dose image measured at the EPID level beyond the patient with respect to its forward calculation. Different implementations use either the treatment planning system or an independent computational engine (e.g., Monte Carlo) for the forward calculation of the transmitted portal dose, starting from the initial nominal or actually measured photon fluence maps prior to the entrance in the patient (typically modeled by the planning CT). Although almost no commercial system supports any of these functionalities yet, valuable clinical results could be already demonstrated by different institutions on the basis of in-house developments, as reported for example in (Nijsten et al. 2007b). However, these 2D data can only serve for a patient-specific verification of the degree of agreement between planned and applied treatment, but do not yet convey any direct feedback on the actual dose delivery to the patient. A straightforward extension of the forward approach towards a 3D dose calculation in the patient involves propagation of the actual photon fluence map measured by an EPID placed in front of the patient (non-transit dosimetry, Steciw et al. 2005), again with different possible implementations with respect to both the used calculation algorithm (TPS or independent engine) and patient model (from the planning CT or in-room megavoltage cone-beam CT). However, the ultimate goal of 3D in-vivo dose verification using direct information on the interaction of the beam with the patient necessitates transmission measurements (transit EPID dosimetry) complemented by a more cumbersome backward approach. The latter requires the extraction of the transmitted primary energy fluence from the portal image and its backprojection to create a 3D primary photon energy fluence map within the irradiated object for the dose estimation. An implementation completely independent from the treatment plan can be achieved by combining a dose calculation engine different from the TPS with anatomical information of the patient in the treatment position as acquired by a cone-beam CT shortly before irradiation. Very promising results of this EPID-based 3D in-vivo dosimetry have been recently reported in the work of van Elmpt et al. (2009), see *Fig. 6*, thus opening new avenues in the quality of adaptive radiation therapy (ART) based not only on detection of anatomical changes (IGRT) but also dose delivery differences (DGRT) in intensity-modulated photon therapy.

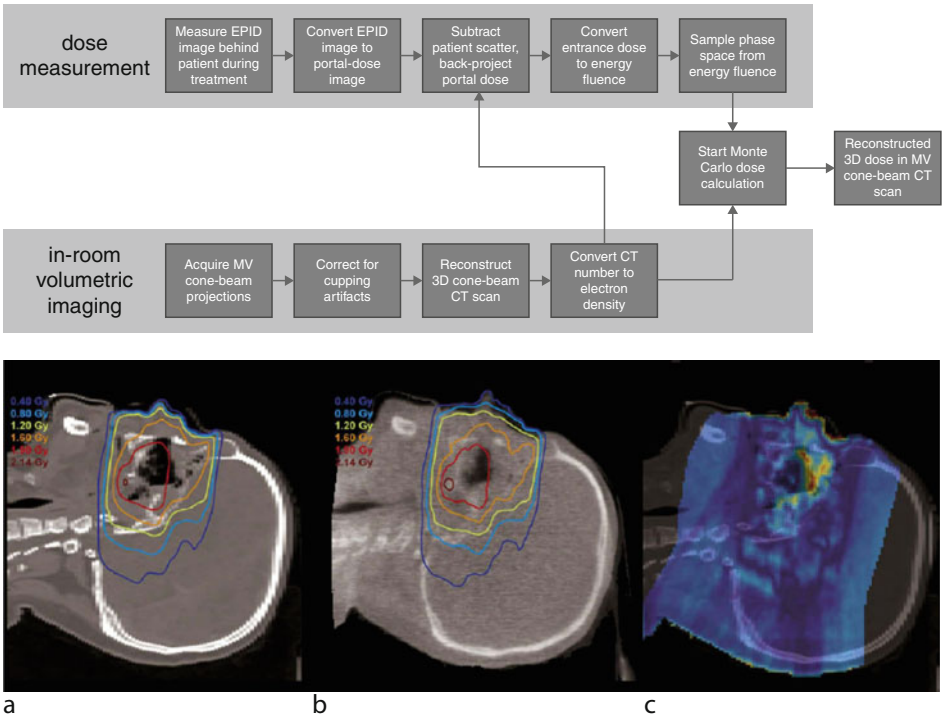


Fig. 6

***Top panel*: Workflow of EPID- and CBCT-based 3D in-vivo dose reconstruction of IMRT. *Bottom panel:* first promising results for a clinical case. Panel (a) shows the dose calculated by the treatment planning system. Panel (b) illustrates the MV cone-beam CT scan with the reconstructed dose distribution. Panel (c) depicts a comparison of the dose distributions from (a) and (b) via gamma evaluation. Adapted from van Elmpt et al. (2009)**

4.2 Range Monitoring and Dose Reconstruction in Ion Therapy

The stopping of the primary ion beam in the tumor and the limited amount of secondary radiation (e.g., neutrons, photons as well as light fragments for heavier ions) exiting the patient downstream of the beam direction during therapeutic dose applications prohibits the use of the portal imaging techniques which are already well established and even under further development in photon therapy (cf. previous subsection). Indeed, pre-treatment transmission imaging of energetic non-therapeutic ion beams could already provide valuable information for verification of the patient position and ion range at the treatment site (cf. subsection on ion-based IGRT). However, the utmost tumor-dose conformality potentially offered by ion beam therapy makes a 3D in-situ confirmation of the in-vivo beam range and of the irradiated volume highly desirable. Being the primary ions completely stopped in the patient, techniques for in-vivo verification of the actual treatment delivery can only be based on the detection of a surrogate signal of emerging secondary radiation during or shortly after therapeutic irradiation.

Nowadays, PET is the only technically feasible method fulfilling the requirement for a 3D, non-invasive, in-vivo monitoring of the delivered ion treatment and, in particular, of the beam range in the patient during or shortly after the therapeutic irradiation. The unconventional application of this nuclear imaging technique (see Chap. 38, "PET Imaging: Basics and New Trends") to ion beam therapy is based on the detection of the transient pattern of β^+ activation (mainly from ^{15}O and ^{11}C with half-lives of approximately 2 and 20 min) which is produced in nuclear fragmentation reactions between the incoming ions and the irradiated tissue. The mechanism of β^+-activity production can be single or twofold, depending on the primary ion beam. This mainly influences the shape of the irradiation-induced activity signal, more uniformly distributed along the beam path for proton beams due to target fragmentation only, while peaked shortly before the Bragg peak for carbon ion beams due to the additional β^+-active projectile fragments. Regardless of the target or projectile activation mechanism, the different physical processes behind nuclear interaction and electromagnetic energy deposition indeed complicate the correlation between the measurable surrogate signal and the sought actual dose deposition. Nevertheless, the comparison of the measured PET images with an expectation calculated on the basis of the planned treatment and the actual time course of irradiation and imaging (e.g., using dedicated Monte Carlo techniques) or deduced by a reference activation measurement (e.g., taken at the first day of treatment) can allow detection of unpredictable discrepancies between prescribed/reference irradiation and actual delivery (Enghardt et al. 2004a; Parodi 2004; Nishio et al. 2010). This offers an independent verification of the whole therapy chain from treatment planning to beam delivery, opening the possibility for intervention prior to the application of the next therapeutic session in fractionated radiotherapy and thus potentially enabling the desired safe reduction of the planning margins as well as the exploitation of the distal-dose fall-off (rather than the less sharp lateral beam penumbra) for more conformal ion treatments and related dose escalation studies. Moreover, it promises to play an important role in assuring the correct treatment delivery in the increasingly considered application of ion beams to high-dose hypo-fractionated therapy, where less or even no subsequent fractions are available for compensation of errors. Finally, it can improve confidence of the successful implementation of special beam delivery strategies like gating or tracking in the challenging presence of organ motion, as supported by first encouraging phantom experiments reported in Parodi et al. (2009).

From the instrumentation point of view, the transient β^+ activation of the patient can be detected *in-beam* (i.e., during irradiation) by means of customized systems fully integrated in the dose delivery environment (Enghardt et al. 2004b; Nishio et al. 2010), or *in-room/offline* (i.e., at remote detectors after irradiation) using conventional nuclear medicine PET scanners located inside or nearby the treatment room (Hishikawa et al. 2002; Parodi et al. 2007), as shown in *Fig. 7*. The former approach demands the development of dedicated detector geometries with typical dual-head configuration to avoid interference with the beam as well as to enable flexible patient positioning. Furthermore, they require customization of the data acquisition system for synchronization with the beam delivery and rejection of the undesired prompt-gamma radiation background during the beam-on time (Crespo et al. 2005; Nishio et al. 2010). Differently, in-room/offline imaging can rely on commercially available full-ring tomographs, but it requires accurate replication/fixation of the treatment position and it is more subject to the degradation of the measurable signal due the physical and biological decay in the time elapsed between irradiation and imaging (Parodi et al. 2007).

The idea of using the irradiation-induced transient pattern of β^+ activation to visualize the ion treatment delivered to the patient was already envisaged since the late 1970s

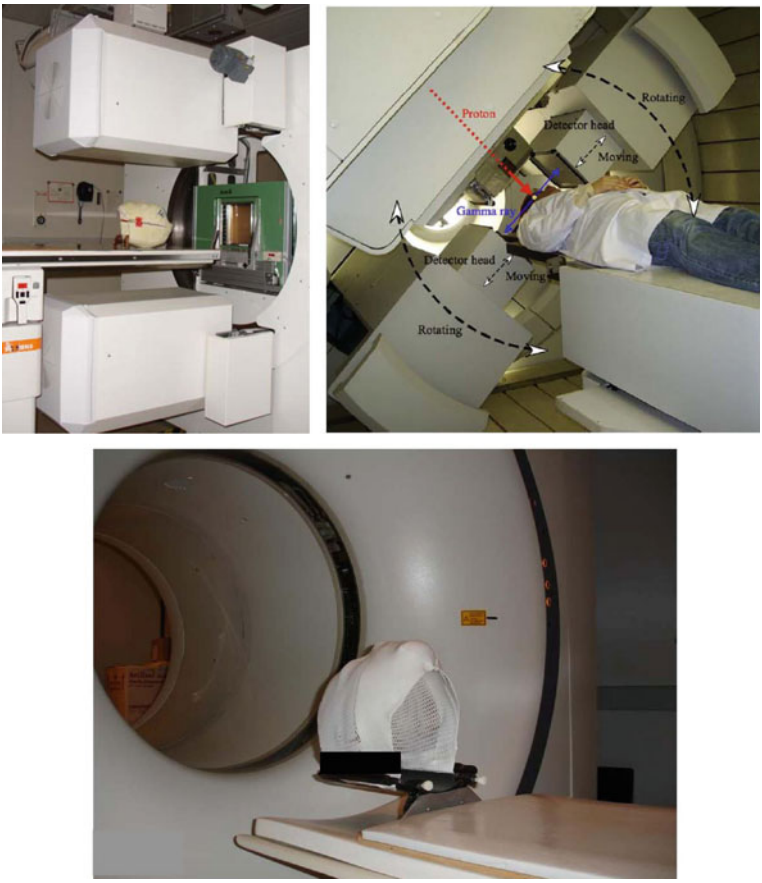


Fig. 7

Examples of instrumentation for PET verification of ion therapy. *Top left*: tomographic double-head PET scanner used to monitor carbon ion therapy during irradiation *(in-beam)* at GSI Darmstadt, Germany (Enghardt et al. 2004b). *Top right*: planar on-line PET system recently installed at the proton beam gantry of the Kashiwa National Cancer Center, Japan, for acquisition immediately after end of irradiation (*in-room,* Nishio et al. 2010). *Bottom*: commercial remote PET/CT scanner used to measure proton-induced activation minutes after irradiation (*offline*) at Massachusetts General Hospital, USA (Parodi et al. 2007)

(Tobias et al. 1977; Bennett et al. 1978). However, almost 30 years elapsed before PET verification of ion beam therapy could be investigated in thorough clinical trials. This was mainly due to the technical challenges for realization of dedicated in-beam PET instrumentation integrated into the clinical treatment units and the limitations (especially co-registration issues) of post-radiation imaging using commercial standalone PET devices. To date, first clinical trials have been reported for more than 400 carbon ion patients and more than 50 proton patients imaged either with dedicated tomographic or planar in-beam double-head PET instrumentation in the treatment room (Enghardt et al. 2004b; Nishio et al. 2010), or with remote commercial PET/CT

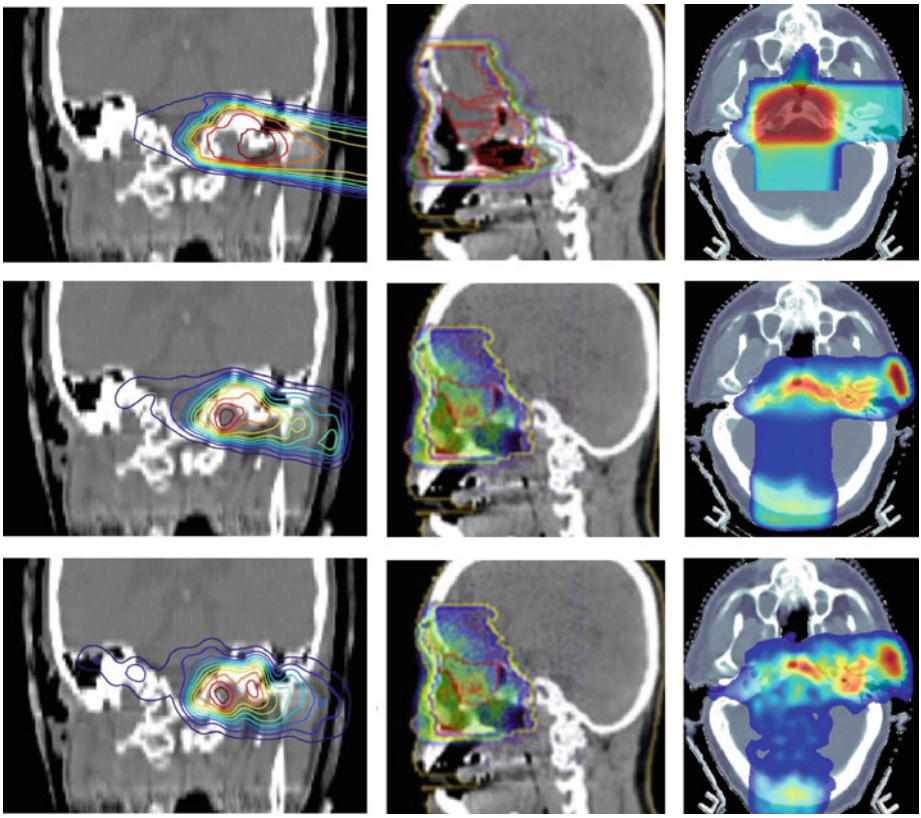


Fig. 8

Clinical implementations of PET verification using the instrumentation of *Fig. 7*. *Top* row: planned carbon ion (*left*) and proton (*middle*, *right*) dose distributions superimposed onto the planning CT for cases of skull-base tumors. *Bottom* row: measured PET images (*left*: in-beam PET at GSI, *middle*: in-room PET in Kashiwa, *right*: offline PET at MGH). *Middle* row: images used for comparison, obtained from Monte Carlo calculation (*left* and *right*) or measured as reference in the first day of treatment (*middle*) (adapted from Parodi et al. (2007, 2008), Nishio et al. (2010))

scanners (Parodi et al. 2007; Hsi et al. 2009; Knopf et al. 2011). Examples of different implementations are shown in *Fig. 8*. The results reported so far indicate the promise of the technique for in-vivo range verification, establishment of population-based as well as patient-specific margins and assessment of unpredictable deviations between planned and actual treatment delivery. This especially applies to favorable anatomical locations of intracranial and cervical spine tumors. However, the clinical studies also highlighted limitations especially in extracranial anatomical locations due to remaining major issues connected to motion and biological washout, in addition to the general challenge of extremely low-counting statistics in comparison to standard tracer imaging in nuclear medicine.

Indeed, the not yet accomplished ultimate goal would be the extrapolation of the dose delivery to the patient from the measured PET signal. Approximate indirect solutions to this very

ill-posed problem have been proposed and clinically implemented for in-beam PET monitoring of carbon ion beam therapy, already demonstrating a valuable clinical feedback (Enghardt et al. 2004a; Parodi 2004). Additional recent works have suggested mathematical formulations for tackling the problem of direct dose reconstruction from PET data in proton therapy (Parodi and Bortfeld 2006; Fourkal et al. 2009). However, satisfactory solutions to the challenging task of accurate PET-based quantification of the actual dose delivered to the patient are still under investigation and might likely benefit from ongoing efforts of several groups and projects, aiming to improve the imaging methodology and technology with realization of new generation, dedicated in-beam PET scanners.

In addition to the research aiming to advance unconventional PET imaging for optimal clinical application in ion beam therapy, new efforts are also being invested to explore alternative or complementary techniques. Among the other possible surrogate signals to be imaged, the most promising one is the prompt-gamma emission from excited nuclear states induced by ion-nucleus interactions (Min et al. 2006; Testa et al. 2010). This radiation, which introduces an undesired background for in-beam PET detection (Crespo et al. 2005), is emitted almost isotropically with a broad energy spectrum in a very short time (<1 ns) after nuclear fragmentation reactions along the beam path. Thus, detection of the prompt gammas preserving spatial information on the place of emission (e.g., via a collimator) may offer the possibility of a *real-time* verification of the irradiated volume insensitive to the degradation of biological processes, in opposition to the PET signal which is intrinsically delayed according to the typical 2–20 min half-life time of the most abundant β^+-emitter products.

To date experimental investigations have been limited to the detection of collimated right-angled prompt gammas, scanned along the proton or carbon ion beam penetration depth in phantoms for 1D assessment of the spatial correlation between the dose delivery and the ion beam range (❯ *Fig. 9*). Detectors investigated so far include, e.g., scintillators combined with lead collimators (Min et al. 2006; Testa et al. 2010) or Compton cameras (Kabuki et al. 2009), with suppression of neutron background based on moderation and absorption or time-of-flight discrimination. Nevertheless, optimal instrumentation fulfilling the conflicting requirements of high detection efficiency, good spatial resolution, effective suppression of the considerable background from secondary neutrons as well as scattered photons and, possibly, enabling recovery of 3D spatial information on the treated volume is still at the research and development phase. Therefore, prompt-gamma imaging cannot be considered yet mature for clinical application, but it is indeed an emerging promising area of research for in-vivo real-time visualization of treatment delivery and beam range in ion beam therapy.

5 Conclusion

This chapter has provided a broad overview on the emerging role of imaging instrumentation and techniques in precision radiotherapy. Indeed, over the last years the constant improvements in the attainable tumor-dose conformality offered by modern radiotherapy techniques like IMRT and ion therapy have demanded an increasing role of imaging in the whole therapeutic process, from the tumor diagnosis to the dose delivery and, finally, the evaluation of the treatment response. This demand has been translated into an increasing multimodal usage of morphological and functional imaging for more accurate treatment planning and biologically adapted radiotherapy, as well as into constant efforts for advancement and integration

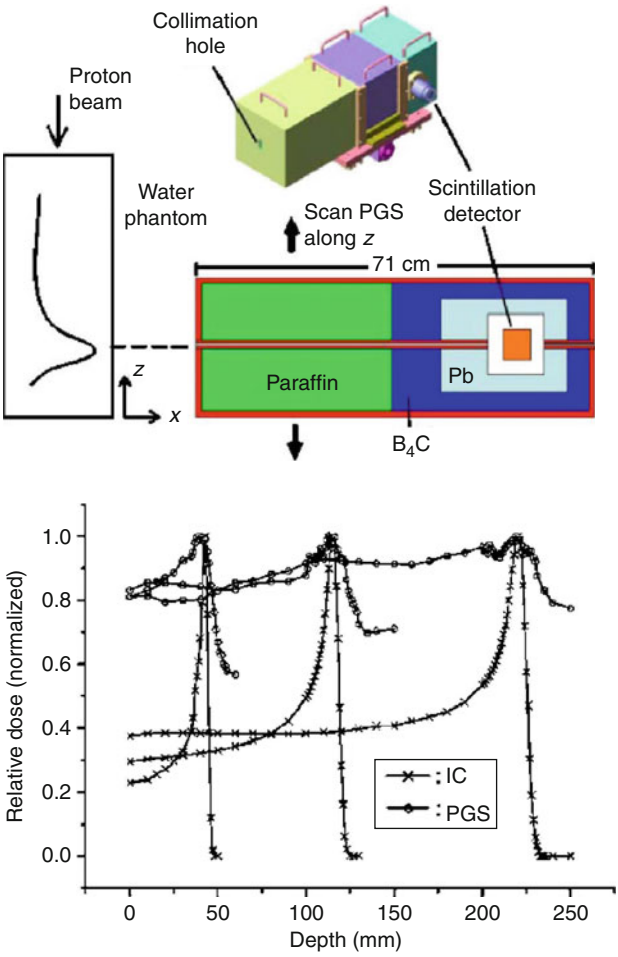


Fig. 9

Experimental set-up (*top panel*) used by (Min et al. 2006) for right-angled measurement of prompt gammas induced by proton beams slowing down in water. A collimator system consisting of lead, paraffin and B_4C powder is used to suppress the considerable background from scattered photons and neutrons, respectively. The gamma detector is a CsI(Tl) scintillator. The resulting prompt-gamma scans (PGS) along the beam penetration in water are compared in the *bottom panel* to depth-dose measurements taken with an ionization chamber (IC) to illustrate the promising correlation with the Bragg-peak location (Min et al. 2006)

of imaging instrumentation in the treatment site for *image-guided radiotherapy*. Moreover, it has prompted new developments towards the final goal of in-vivo treatment verification and *dose guidance*, either resorting to unconventional application and further development of established imaging modalities or to the investigation of novel concepts and detector solutions, often profiting from technological advances of other fields like high energy physics. Overall, it can

be foreseen that in the multidisciplinary world of radiotherapy, a wide spectrum of state-of-the-art imaging instrumentation and techniques will increasingly play a major role towards therapeutically effective, high-precision treatments.

References

Antonuk LE (2002) Electronic portal imaging devices: a review and historical perspective of contemporary technologies and research. Phys Med Biol 47:R31–65

Bennett GW, Archambeau JO, Archambeau BE, Meltzer JI, Wingate CL (1978) Visualization and transport of positron emission from proton activation in vivo. Science 200:1151–1153

Berbeco RI, Neicu T, Rietzel E, Chen GT, Jiang SB (2005) A technique for respiratory-gated radiotherapy treatment verification with an EPID in cine mode. Phys Med Biol 50:3669–3679

Bert C, Metheany KG, Doppke KP, Taghian AG, Powell SN, Chen GT (2006) Clinical experience with a 3D surface patient setup system for alignment of partial-breast irradiation patients. Int J Radiat Oncol Biol Phys 64:1265–1274

Boda-Heggemann J, Köhler FM, Küpper B, Wolff D, Wertz H, Mai S, Hesser J, Lohr F, Wenz F (2008) Accuracy of ultrasound-based (BAT) prostate-repositioning: a three-dimensional online fiducial-based assessment with cone-beam computed tomography. Int J Radiat Oncol Biol Phys 70:1247–1255

Brusasco C, Voss B, Schardt D, Krämer M, Kraft G (2000) A dosimetry system for fast measurement of 3D depth-dose profiles in charged-particle tumor therapy with scanning techniques. Nucl Instrum Meth Phys Res B 168:578–592

Crespo P, Barthel T, Frais-Kölbl H, Griesmayer E, Heidel K, Parodi K, Pawelke J, Enghardt W (2005) Suppression of random coincidences during in-beam PET measurements. IEEE Trans Nucl Sci 52:980–987

Drzymala RE, Mohan R, Brewster L, Chy J, Goitein M, Harms W, Urie M (1991) Dose-volume histograms. Int J Radiat Oncol Biol Phys 21: 71–78

Engelke J et al (2010) Investigation of an amorphous silicon detector for ion radiography. In: Book of abstract of the 49th annual meeting of the particle therapy co-operative group PTCOG 49, Gunma, Japan, 20–22 May

Enghardt W, Parodi K, Crespo P, Fiedler F, Pawelke J, Pönisch F (2004a) Dose quantification from in-beam positron emission tomography. Radiother Oncol 73:S96–S98

Enghardt W, Crespo P, Fiedler F, Hinz R, Pawelke J, Parodi K, Pönisch F (2004b) Charged hadron tumour therapy monitoring by means of PET. Nucl Instrum Methods A 525:284–288

Fenwick JD, Tomé WA, Soisson ET, Mehta MP, Mackie TR (2006) Tomotherapy and other innovative IMRT delivery systems. Semin Radiother Oncol 16(4):199–208

Fourkal E, Fan J, Veltchev I (2009) Absolute dose reconstruction in proton therapy using PET imaging modality: feasibility study. Phys Med Biol 54:N217–N228

Haberer T, Becher W, Schardt D, Kraft G (1993) Magnetic scanning system for heavy ion therapy. Nucl Instrum Methods Phys Res A 330:296–305

Haberer T, Debus J, Eickhoff H, Jäkel O, Schulz-Ertner D, Weber U (2004) The Heidelberg ion therapy center. Radiother Oncol 73(S2):186–190

Hanson KM, Bradbury JN, Cannon TM, Hutson RL, Laubacher DB, Macek RJ, Paciotti AM, Taylor CA (1981) Computed tomography using proton energy loss. Phys Med Biol 26:965–983

Hishikawa Y, Kagawa K, Murakami M, Sakai H, Akagi T, Abe M (2002) Usefulness of positron-emission tomographic images after proton therapy. Int J Radiat Oncol Biol Phys 53:1388–1391

Hsi WC, Indelicato DJ, Vargas C, Duvvuri S, Li Z, Palta J (2009) In vivo verification of proton beam path by using post-treatment PET/CT imaging. Med Phys 36:4136–4146

ICRU-International Commission on Radiation Units and Measurements (1993) ICRU report 50: prescribing, recording, and reporting photon beam therapy. International Commission on Radiation Units and Measurement, Bethesda, pp 3–16

ICRU- International Commission on Radiation Units and Measurements (1999) ICRU report 62: prescribing, recording, and reporting photon beam therapy. International Commission on Radiation Units and Measurement, Bethesda, pp 3–20

Jäkel O, Reiss P (2007) The influence of metal artefacts on the range of ion beams. Phys Med Biol 52:635–644

Jaffray DA, Siewerdsen JH, Wong JW, Martinez AA (2002) Flat-panel cone-beam computed tomo graphy for image-guided radiation therapy. Int J Radiat Oncol Biol Phys 53:1337–1349

Kabuki S, Ueno K, Kurosawa S, Iwaki S, Kubo H, Miuchi K, Fujii Y, Kim D, Kim J, Kohara R, Miyazaki O, Sakae T, Shirahata T, Takayanagi T, Terunuma T, Tsukahara Y, Yamamoto E, Yasuoka K, Tanimori T (2009) Study on the use of electron-tracking Compton gamma-ray camera to monitor the therapeutic proton dose distribution in real time. In: IEEE nuclear science symposium conference record 2009, Orlando, pp 2437–2440

Kamada T, Tsujii H, Mizoe JE, Matsuoka Y, Tsuji H, Osaka Y, Minohara S, Miyahara N, Endo M, Kanai T (1999) A horizontal CT system dedicated to heavy-ion beam treatment. Radiother Oncol 50:235–237

Keall P, Vedam S, George R, Bartee C, Siebers J, Lerma F, Weiss E, Chung T (2006a) The clinical implementation of respiratory-gated intensity-modulated radiotherapy. Med Dosim 31:152–162

Keall P, Cattell H, Pokhrel D, Dieterich S, Wong KH, Murphy MJ, Vedam SS, Wijesooriya K, Mohan R (2006b) Geometric accuracy of a real-time target tracking system with dynamic multileaf collimator tracking system. Int J Radiat Oncol Biol Phys 65:1579–1584

Knopf AC, Parodi K, Paganetti H, Bortfeld T, Daartz J, Engelsman M, Liebsch N, Shih H (2011) Accuracy of proton beam range verification using post-treatment positron emission tomography/computed tomography as function of treatment site. Int J Radiat Oncol Biol Phys 79:297–304

Koehler AM (1968) Proton radiography. Science 160:303–304

Korreman S, Rasch C, McNair H, Oelfke U, Maingon P, Mijnheer B, Khoo V (2010) The European Society of Therapeutic Radiology and Oncology-European Institute of Radiotherapy (ESTRO-EIR) report on 3D CT-based in-room image guidance systems: a practical and technical review and guide. Radiother Oncol 94:129–144

Lauterbur PC (1973) Image formation by induced local interactions: examples of employing nuclear magnetic resonance. Nature 242: 190–191

Li XA, Stepaniak C, Gore E (2006) Technical and dosimetric aspects of respiratory gating using a pressure-sensor motion monitoring system. Med Phys 33:145–154

Ling CC, Humm J, Larson S, Amols H, Fuks Z, Leibel S, Koutcher JA (2000) Towards multidimensional radiotherapy (MD-CRT): biological imaging and biological conformality. Int J Radiat Oncol Biol Phys 47:551–560

Lu W, Low DA, Parikh PJ, Nystrom MM, Naqa IM, Wahab SH, Handoko M, Fooshee D, Bradley JD (2005) Comparison of spirometry and abdominal height as four-dimensional computed tomography metrics in lung. Med Phys 32:2351–2357

Mackie TR, Kapatoes J, Ruchala K, Lu W, Wu C, Olivera G, Forrest L, Tome W, Welsh J, Jeraj R, Harari P, Reckwerdt P, Paliwal B, Ritter M, Keller H, Fowler J, Mehta M (2003) Image guidance for precise conformal radiotherapy. Int J Radiat Oncol Biol Phys 56:89–105

Min CH, Kim CH, Youn MY, Kim JW (2006) Prompt gamma measurements for locating the dose falloff region in the proton therapy. Appl Phys Lett 89(183517):1–3

Nijsten SM, van Elmpt WJ, Jacobs M, Mijnheer BJ, Dekker AL, Lambin P, Minken AW (2007a) A global calibration model for a-Si EPIDs used for transit dosimetry. Med Phys 34:3872–3884

Nijsten SM, Mijnheer BJ, Dekker AL, Lambin P, Minken AW (2007b) Routine individualised patient dosimetry using electronic portal imaging devices. Radiother Oncol 83:65–75

Nishio T, Miyatake A, Ogino T, Nakagawa K, Saijo N, Esumi H (2010) The development and clinical use of a beam ON-LINE PET system mounted on a rotating gantry port in proton therapy. Int J Radiat Oncol Biol Phys 76:277–286

Olsen DR, Thwaites DI (2007) Now you see it…Imaging in radiotherapy treatment planning and delivery. Radiother Oncol 85:173–175

Parodi K (2004) On the feasibility of dose quantification with in-beam PET data in radiotherapy with ^{12}C and proton beams. PhD thesis, Dresden University of Technology. In Forschungszentrum Rossendorf Wiss-Techn-Ber FZR-415

Parodi K, Bortfeld T (2006) A filtering approach based on Gaussian-powerlaw convolutions for local PET verification of proton radiotherapy. Phys Med Biol 51:1991–2009

Parodi K, Bortfeld T, Enghardt W, Fiedler F, Knopf A, Paganetti H, Pawelke J, Shakirin G, Shih H (2008) PET imaging for treatment verification of ion therapy: implementation and experience at GSI Darmstadt and MGH Boston. Nucl Instrum Methods A 591:282–286

Parodi K, Paganetti H, Shih HA, Michaud S, Loeffler JS, DeLaney TF, Liebsch NJ, Munzenrider JE, Fischman AJ, Knopf A, Bortfeld T (2007) Patient study on in-vivo verification of beam delivery and range using PET/CT imaging after proton therapy. Int J Rad Oncol Biol Phys 68:920–934

Parodi K, Saito N, Chaudhri N, Richter C, Durante M, Enghardt W, Rietzel E, Bert C (2009) 4D in-beam

positron emission tomography for verification of motion-compensated ion beam therapy. Med Phys 36:4230–4243

Pedroni E, Böhringer T, Coray A, Egger E, Grossmann M, Lin S, Lomax A, Goitein G, Roser W, Schaffner B (1999) Initial experience of using an active beam delivery technique at PSI. Strahlenther Onkol 175(II):18–20

Pedroni E (2009) Proton beam delivery technique and commissioning issues: scanned protons. Presented at PTCOG Educational meeting, Jacksonville, http://ptcog.web.psi.ch/PTCOG47/presentations/1_Education_Monday/EPedroni.pdf

Petterson M, Blumenkrantz N, Feldt J, Heimann J, Lucia D, Seiden A, Williams DC, Sadrozinski HF-W, Bashkirov V, Schulte R, Bruzzi M, Menichelli D, Scaringella M, Talamonti C, Cirrone GAP, Cuttone G, Lo Presti D, Randazzo N, Sipala V (2007) Proton radiography studies for proton CT. In: IEEE nuclear science symposium conference record 2006, San Diego, pp 2276–2280

Podgorsak EB (technical ed) (2005) Radiation oncology physics: a handbook for teachers and students, IAEA International Atomic Energy Agency, Vienna, p 220

PSI website (2010) The PSI Proton Therapy Facility: Proton radiography on the PSI gantry. http://radmed.web.psi.ch/asm/gantry/handl/n_handl.html

Raaymakers BW, Raaijmakers AJ, Kotte AN, Jette D, Lagendijk JJ (2004) Integrating a MRI scanner with a 6 MV radiotherapy accelerator: dose deposition in a transverse magnetic field. Phys Med Biol 49:4109–4118

Rietzel E, Pan T, Chen GT (2005) Four-dimensional computed tomography: image formation and clinical protocol. Med Phys 32:874–889

Rietzel E, Schardt D, Haberer T (2007) Range accuracy in carbon ion treatment planning based on CT-calibration with real tissue samples. Radiat Oncol 23:2–14

Rinaldi I, Ferrari A, Jäkel O, Mairani A, Parodi K (2010) Novel imaging and quality assurance techniques for ion beam therapy: a Monte Carlo study. In: Proceedings of the 12th international conference on nuclear reaction mechanisms, Varenna, Italy, 15–19 June 2009. CERN-Proceedings-2010-001:575–580

Ryu H, Song E, Lee J, Kim J (2008) Density and spatial resolutions of proton radiography using a range modulation technique. Phys Med Biol 53:5461–5468

Saw CB, Brandner E, Selvaraj R, Chen H, Saiful Huq M, Heron DE (2007) A review on the clinical implementation of respiratory-gated radiation therapy. Biomed Imaging Interv J 3:40

Schaffner B, Pedroni E (1998) The precision of proton range calculations in proton radiotherapy treatment planning: experimental verification of the relation between CT-HU and proton stopping power. Phys Med Biol 43:1579–1592

Schlegel W, Mahr A (2001) 3D Conformal Radiation Therapy: A multimedia introduction to methods and techniques. Springer, Berlin

Schneider U, Pedroni E (1995) Proton Radiography as a tool for quality control in proton therapy. Med Phys 22:353–363

Schneider U, Besserer J, Pemler P, Dellert M, Moosburger M, Pedroni E, Kaser-Hotz B (2004) First proton radiography of an animal patient. Med Phys 31:1046–1051

Schneider U, Pemler P, Besserer J (2005) Patient specific optimization of the relation between CT-Hounsfield units and proton stopping power with proton radiography. Med Phys 32:195–199

Shinoda H, Kanai T, Kohno T (2006) Application of heavy-ion CT. Phys Med Biol 51:4073–4081

Sisterson J (2005), Particles Newsletter 36: 10–11 (http://www.ptcog.com/particles/ptles36.doc)

Sommer FG, Capp MP, Tobias CA, Benton EV, Woodruff KH, Henke RP, Holly W, Genant HK (1978) Heavy-ion radiography: density resolution and specimen radiography. Investig Radiol 13:163–170

Steciw S, Warkentin B, Rathee S, Fallone BG (2005) Three-dimensional IMRT verification with a flat-panel EPID. Med Phys 32:600–612

Testa M, Bajard M, Chevallier M, Dauvergne D, Freud N, Henriquet P, Karkar S, Le Foulher F, Létang JM, Plescak R, Ray C, Richard MH, Schardt D, Testa E (2010) Real-time monitoring of the Bragg-peak position in ion therapy by means of single photon detection. Radiat Environ Biophys 49:337–343

Thieke C, Malsch U, Schlegel W, Debus J, Huber P, Bendl R, Thilmann C (2006) Kilovoltage CT using a linac-CT scanner combination. BJR 79(Spec No 1):S79–S86

Thomas SJ (1999) Relative electron density calibration of CT scanners for radiotherapy treatment planning. BJR 72:781–786

Tobias CA, Benton EV, Capp MP, Chatterjee A, Cruty MR, Henke RP (1977) Particle radiography and autoactivation. Int J Radiat Oncol Biol Phys 3:35–44

van Elmpt WJ, McDermott L, Nijsten S, Wendling M, Lambin L, Mijnheer B (2008) A literature review of electronic portal imaging for radiotherapy dosimetry. Radiother Oncol 88(3):289–309

van Elmpt WJ, Nijsten SM, Petit S, Mijnheer BJ, Lambin P, Dekker AL (2009) 3D in vivo dosimetry using megavoltage cone-beam CT and EPID dosimetry. Int J Rad Oncol Biol Phys 73:1580–1587

Webb S (2001) Intensity modulated radiation therapy, Series in medical physics. Institute of Physics Publishing, Bristol. ISBN 0-7503-0699-8

Weinrich U (2006) Gantry design for proton and carbon hadrontherapy facilities. In: Proceedings of EPAC 2006, Edinburgh, Scotland, pp 964–968

Willoughby TR, Kupelian PA, Pouliot J, Shinohara K, Aubin M, Roach M, Skrumeda LL et al (2006) Target localization and real-time tracking using the Calypso 4D localization system in patients with localized prostate cancer. Int J Radiat Oncol Biol Phys 65:528–534

Xu XG, Bednarz B, Paganetti H (2008) A review of dosimetry studies on external-beam radiation treatment with respect to second cancer induction. Phys Med Biol 53:R193–R241

Further Reading

Fiedler F, Shakirin G, Skowron J, Braess H, Crespo P, Kunath D, Pawelke J, Pönisch F, Enghardt W (2010) On the effectiveness of ion range determination from in-beam PET data. Phys Med Biol 55:1989–1998

Grosu AL, Molls M, Zimmermann FB, Geinitz H, Nüsslin F, Schwaiger M, Nieder C (2006) High-precision radiation therapy with integrated biological imaging and tumor monitoring: evolution of the Munich concept and future research options. Strahlenther Onkol 182: 361–368

Grosu AL, Nestle U, Weber WA (2009) How to use functional imaging information for radiotherapy planning. Eur J Cancer 45(1):461–463

Miyatake A, Nishio T, Ogino T, Saijo N, Esumi H, Uesaka M (2010) Measurement and verification of positron emitter nuclei generated at each treatment site by target nuclear fragment reactions in proton therapy. Med Phys 37:4445–4455

van Elmpt W (2009) 3D dose verification for advanced radiotherapy. PhD thesis, Maastricht University, the Netherlands

47 Tumor Therapy with Ion Beams

Gerhard Kraft[1] · *Uli Weber*[2]
[1]GSI Helmholtzzentrum für Schwerionenforschung GmbH, Darmstadt, Germany
[2]Universitätsklinikum Gießen/Marburg, Germany

1 Introduction ***1180***

2 Physical Basics for Particle Therapy ***1180***
2.1 Energy Deposition and Depth–Dose Distribution of Particle Beams 1180
2.2 Lateral and Longitudinal Scattering 1181
2.3 Nuclear Fragmentation and PET Verification 1184

3 Clinical Beam Application Systems ***1186***
3.1 Passive Beam Spreading 1186
3.2 Active Beam Delivery 1187

4 Detectors and Quality Assurance ***1189***
4.1 Therapy Online Monitors 1189
4.2 Detectors for Permanent Recording: Films and Nuclear Track Detectors . . . 1191
4.3 Ionization Chamber Dosimetry 1192

5 Biological Properties of Heavy Ions Relevant for Therapy ***1195***
5.1 Definition of RBE and Its Dependence on Dose or Effect Level 1196
5.2 The RBE Dependencies on Physical and Biological Parameters and the Molecular Understanding 1197

6 The Planning of the Biological Effective Dose ***1199***

7 Quality Assurance and RBE Detectors ***1201***

8 Conclusions ***1204***

References ***1204***

Further Reading ***1205***

C. Grupen, I. Buvat (eds.), *Handbook of Particle Detection and Imaging*,
DOI 10.1007/978-3-642-13271-1_47, © Springer-Verlag Berlin Heidelberg 2012

Abstract: From 1954 when the first patient was treated at Berkeley to now, tumor therapy using ion beams has developed to high-technology application. In order to achieve an extreme tumor conform irradiation a fine pencil beam is guided over a three-dimensional grid of pixels that fills the target volume. A main problem is the quality assurance before, during, and after patient irradiation where different types of detectors and monitors are used. In this chapter, the basic principles of ion beam therapy are given and the monitor systems are described more in their functionality rather than in the individual specifications that differ between the various therapy units.

1 Introduction

In the developed countries, half a percent of the population is diagnosed with cancer every year. For more than 40% of these patients, a 5 year tumor control can be reached meaning that after 5 years these patients are free of tumors and are cured. The different therapy modalities contribute as follows: surgery 22%, radiotherapy 12%, and the combination of both 6% (De Vita et al. 1997). Chemotherapy is also frequently used but mostly in a palliative way, i.e., with the expectation to better the quality of the remaining life span and to extend it. Most of the finally cured patients have been early diagnosed with one solid tumor without metastases. But there are nearly another 20% of the patients with no metastases that have a poor prognosis: their tumor cannot be removed completely by surgery nor by radiation because of critical structures nearby. These patients are the candidates for a high-precision therapy like the intensity-modulated particle therapy (IMPT) (Durante and Loeffler 2010).

Using ion beams, especially heavy ions like carbon, a millimeter precision can be reached in principle everywhere in the body and target volumes can be inactivated with high doses but sparing the critical structures and therefore minimizing complications (Kraft 2000; Schulz-Ertner and Tsujii 2007). Prerequisite for the success of such a treatment is the appropriate quality assurance (QA) of all relevant parameters starting in the diagnostics, the patient positioning, the planning, and the patient immobilization. In this chapter, however, the beam application and its quality control will be discussed. This is the online beam monitoring, the retrospective PET analysis, and the physical and biological effectiveness verification. To do so, the basic principles of ion beam therapy are outlined first.

2 Physical Basics for Particle Therapy

2.1 Energy Deposition and Depth–Dose Distribution of Particle Beams

The main reason to use ions in therapy instead of photons is the inverse dose profile, i.e., the increase of energy deposition with penetration depth up to a sharp maximum at the end of the particle range, the Bragg peak, named after William Bragg, who measured this characteristic behavior for alpha particles.

In 1946, in measurements at the Berkeley cyclotron Robert Wilson recognized the potential of ions like protons or carbon for tumor therapy (Wilson 1946). Their favorable depth–dose

distribution is a direct consequence of the interaction mechanism of heavy charged particles with the penetrated material and is different from that of electromagnetic radiation as reported in detail in Chap. 1, "Interactions of Particles and Radiation with Matter." The therapy-relevant properties will be briefly summarized here:

Within the energy range used for therapy, heavy charged particles interact predominantly with the target electrons and the interaction strength is directly correlated with the interaction time. At high projectile energies, the interaction time is short and the energy transfer to the target is small. When the particles are slowed down and close to the end of their range the interaction time becomes larger and the value of the energy transfer is at its maximum.

The energy loss as function of particle energy and atomic number is given in the Bethe–Bloch formula (Bethe 1930; Bloch 1933):

$$\frac{\mathrm{d}E}{\mathrm{d}x} = -\frac{4\pi e^4 Z_{\mathrm{eff}}^2 N}{m_e v^2} \times \left[\ln \frac{2m_e v^2}{I(1-\beta^2)}\right] + \text{relativistic terms}, \tag{1}$$

where m is the electron mass, v the projectile velocity, N the density of the electrons of the target material, e the elementary charge, and I the mean ionization potential. Finally, Z_{eff} is the effective charge interacting with the target electrons. Z_{eff} was approximated by Barkas in an empirical formula (Barkas 1963). For high energies, all electrons are stripped off from the projectile and the effective charge equals the atomic number.

The maximum of interaction between projectile and target occurs at low energies at around 20 MeV/u for carbon ions when the atoms are still fully stripped but have a large interaction time. Then many electrons are liberated from the target atoms by the ions impact forming a very dense track of the biologically important low-energy electrons. When the distance between two ionization events either from the ions or from the subsequent delta electrons becomes comparable to the distance between the DNA strands a simultaneous cleavage of both strands, a double-strand break is more likely and the quality of the biological lesion changes dramatically. For heavier ions clusters of double-strand breaks are formed which overcome the cellular repair capacity (*Fig. 1*). This enhanced reaction is the main reason for the use of the heavier ions. The coincidence of the maximum biological action with the maximum energy loss is the reason to select carbon ions for therapy (Kraft 2000).

At smaller energies, electrons are collected from the target and the effective charge decreases, approaching zero when the particles stop. This results in a sharp decrease of the energy loss at low energies and in consequence the finite range of particle beams which is an essential criterion for the use of heavy particles in therapy.

2.2 Lateral and Longitudinal Scattering

The range $R(E)$ of the ions can be obtained by integrating the energy loss $\mathrm{d}E/\mathrm{d}x$ over the energy,

$$R(E) = \int_0^E \frac{\mathrm{d}E'}{\mathrm{d}E/\mathrm{d}x(E')}. \tag{2}$$

When the energy loss according to the Bethe–Bloch relation is plotted over the penetration depth a depth–dose profile (Bragg curve) for a single particle results. This theoretical single-particle curve has a much larger dose ratio from plateau to peak than the measured Bragg curves. Because of the statistics of the energy loss process, the interaction of the projectile with the electrons yields a small straggling of the individual particle ranges (Molière 1948; Gottschalk et al. 1993).

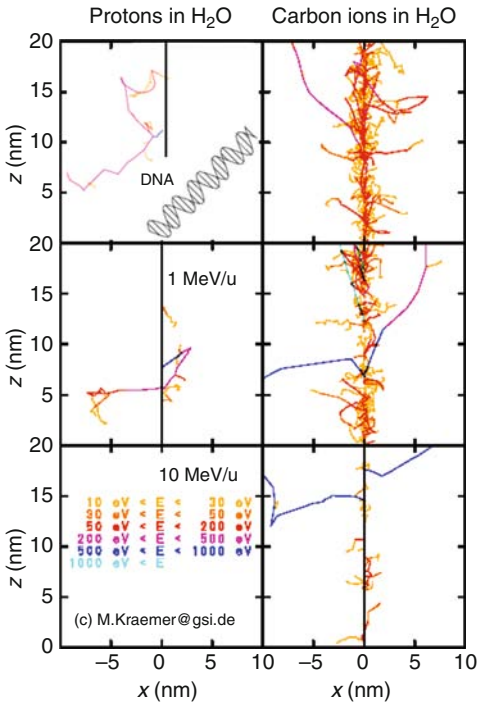


Fig. 1

The microscopic structure of proton and carbon tracks in water is compared to a schematic representation of a DNA molecule. For protons, mostly independent electron tracks are formed, for the heavier carbon ions at low energies many electron tracks are produced that can cause clusters of locally damaged sites within the DNA that cannot be repaired (Krämer and Kraft 1994)

The range straggling broadens the individual Bragg curve and decreases the peak to plateau ratio. Depending on the atomic number of the projectiles, the Bragg peak is always broader for protons than for carbon ions (*Fig. 2*). Because of the much larger size of the target volumes the intrinsic width of the Bragg maxima is not important. In the practical application, the treatment planning the energy loss data as well as the range data cannot be calculated with a sufficient accuracy (< 0.1%) for therapy and the formulas are fitted to precision measurements (Haettner et al. 2006).

More important than the longitudinal straggling for therapy is the lateral scattering. Because of general range uncertainties, a target volume in the proximity of a critical structure will not be treated "in a head on irradiation" where the beam stops in front of the critical structures. These tumors are treated in a way that the beam passes the critical sites. Then, the lateral scattering determines the closest approach possible. *Figure 2* shows that for protons the beam broadening is less than 2 mm for a penetration depth up to 10 cm in tissue but increases then rapidly for larger penetration depth. For carbon ions, the broadening is smaller than for photon beams up to a penetration depth of 20 cm, which will allow precision treatments at any depth.

The lateral scattering determines also the overall quality of the treatment of usually inhomogeneous target volumes: normally the planned range of a beam is calculated for the density

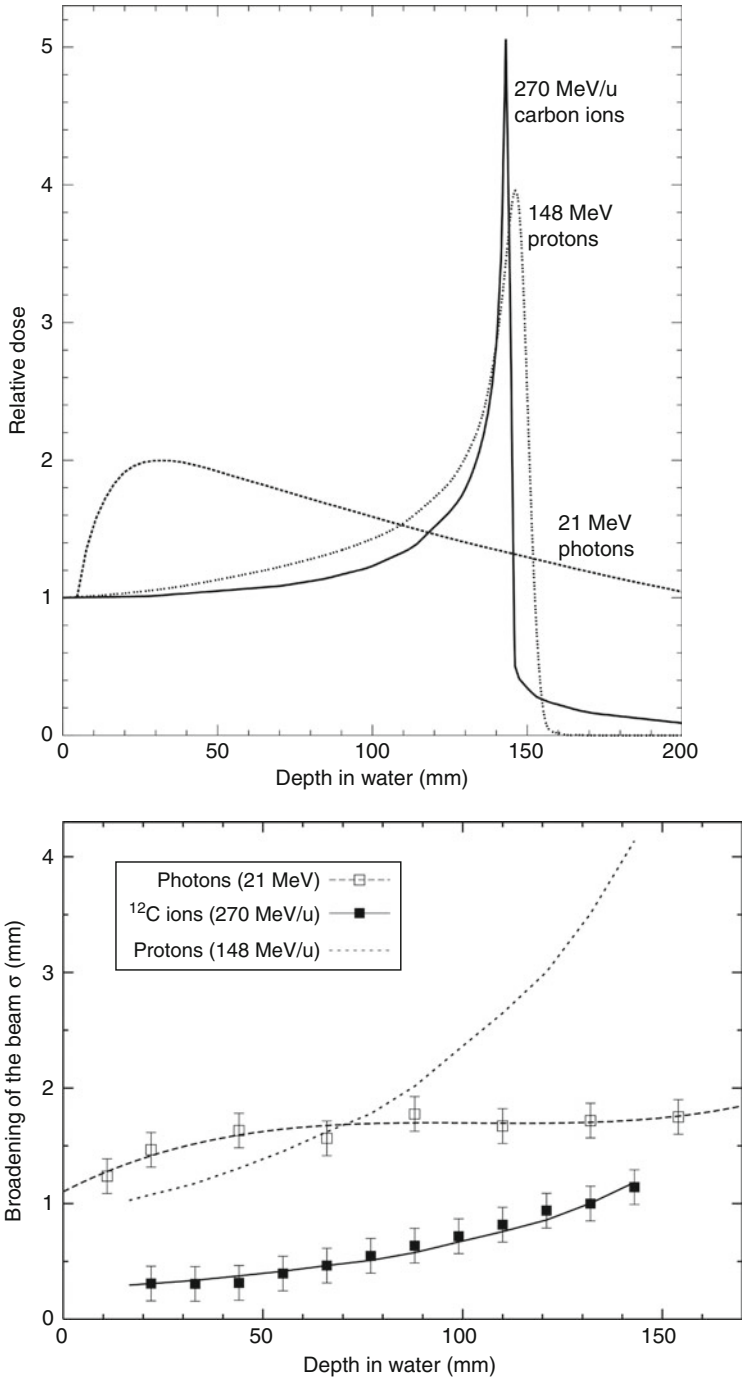


Fig. 2

The measured dose as function of penetration depth is compared for photons, protons, and carbon ions (*top*). The lateral scattering of these beams is given for the same penetration depth (*bottom*) (Krämer et al. 2000)

of the central ray. When the beam is enlarged because of lateral scattering, the outer part will penetrate tissues of different densities and will therefore have a different range. This is frequently the case for head tumors where dense structures like bones and low-density structures like vacuoles are close together then these density variations translate into range variations at the distal side of the target field, which exceed the influence of the longitudinal scattering (Weber and Kraft 2009).

In consequence, the heavier beams having low lateral scattering produce the smallest penumbra of the beam and the best overall confinement of the treatment plans.

2.3 Nuclear Fragmentation and PET Verification

Another difference between protons and heavy ions in their penetration profile becomes visible in *Fig. 2* where the carbon ions produce a small tail of dose at the distal side. In nuclear collisions, lighter fragments are produced that have the same velocity as the primary ions at the collision (Hüfner 1985). The range R of these fragments is given by the formula

$$R_{\text{fragment}} = R_{\text{projectile}} \frac{Z_{\text{pr}}^2}{M_{\text{pr}}} \frac{M_{\text{frag}}}{Z_{\text{frag}}^2} \tag{3}$$

with Z the atomic number and M the masses of fragments and projectiles, respectively. The fragments with a lower atomic number have a longer range forming a tail of dose beyond the Bragg maximum of the primary beam.

A special case represents the isotopes of the projectile ions that have lost one or two neutrons. These neutron-deficient isotopes ^{10}C and ^{11}C are positron emitters and their stopping point can be monitored over the coincident emission of the two annihilation quanta of the positron decay. As the stopping point of the radioactive carbon isotopes is close to the range of the primary beam, the range of the primary beam can be calculated from the measurement of the positron activity. Another positron emitter is ^{15}O, which is produced by nuclear reactions inside the patient's body by any projectile including protons (Enghardt et al. 1992). The ^{15}O distribution represents a broad continuum but having a sharp decay close to the range of the primary projectiles. Therefore, also the range of protons can be recorded inside the patient by positron emission tomography (PET) measurements (Parodi et al. 2007). The positron-emitting carbon isotopes have been routinely used for beam verification in the GSI therapy (*Fig. 4*). Although the pattern obtained by positron-emitting isotopes is not identical with the dose distribution, it is possible to monitor the range of the primary beam (Enghardt et al. 1999).

In the pilot therapy project, PET imaging was an extremely useful tool because of a noninvasive verification of the particle range inside the patient's body. In the human body, there are density differences between fat, bones, muscles of some 30%. In addition, air-filled vacuoles as well as metal implants have a much greater density variation, which have to be considered in treatment planning. For the recording of the density distribution a CT scan is taken without contrast drug. From the gray values, the Hounsfield numbers the electron densities can be calculated using a calibration curve that has been produced before on the basis of fresh animal samples (Geiss et al. 1999). With this empirical calibration, good agreement has been found in the patient therapy between PET-measured ranges of the carbon beams and the distribution calculated before. Disagreement as observed in a few percent of patients were found for instance

when vacuoles filled with liquids during treatment time. In such a case, the planning has to be redone before the treatment can be continued.

In the GSI pilot project, the online PET imaging of the beam-stopping distribution developed very rapidly to an important tool of quality assurance, which was extremely valuable because of having a daily control for all the novel techniques of beam application. Limitations of the online PET arose from the poor statistics of 30–50,000 true coincidences that could be analyzed after the treatment of a single field. This small number was due to geometrical reasons of the limited angle of the camera (*Fig. 3*) and because of the limited sampling time: during treatment only in the beam extraction pauses coincidences could be analyzed. In addition the time of approx. 1 min for the access of the physician could be used for data acquisition because it was not intended to let the patient in his or her very unpleasant fixation longer than necessary.

At the new facilities at Heidelberg, Marburg, and Kiel online PETs around the patient are not planned in the same or similar way as tested at GSI: at these new facilities, the patients are carried by a robotic device. Therefore, it is possible to install an off-line PET. Within a



Fig. 3

PET camera at the carbon therapy at GSI. The carbon beam exits from the window in the center and hits the patient head in the mask attached to the patient couch. The two camera heads contain 32 BGO crystals that are used to monitor the coincidently emitted gamma quanta from the positron decay (Enghardt et al. 1999)

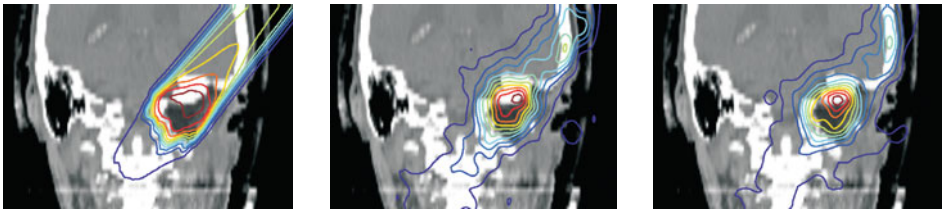


Fig. 4
Comparison of the treatment plan (*left*) with the expected (*middle*) and measured (*right*) distribution of the beta activity (Enghardt et al. 1999)

few seconds the patient is transferred from the exposure position at the beam exit into a PET device with a full detector ring having a higher counting rate and therefore better resolution. See also Chaps. 37, "SPECT Imaging: Basics and New Trends" and 38, "PET Imaging: Basics and New Trends."

3 Clinical Beam Application Systems

3.1 Passive Beam Spreading

Essential for the treatment success is the conformity of the delivered dose to the target volume. The pristine beam from the accelerator has a small energy spread in order of a few per mill and a small divergence resulting in a beam spot of a few millimeters in diameter. In order to cover an extended target volume of typically 100 cm^2 the beam has to be enlarged, both in lateral and longitudinal direction. At present, most proton and heavy-ion facilities use passive beam-forming techniques, which are very close to the dose-shaping systems of the conventional X-ray therapy (*Fig. 5*). In these techniques, the primary mono-energetic and sharp "pencil" beam is widened laterally by complex scattering systems that produce a profile having a flat top. Then using apertures like collimators the outer contours are shaped to the projected tumor-contours (in "beams eye view"). In addition, the beam is modulated in depth with ridge filters according to the maximum extension of the target volume: particles penetrating through the valleys between the ridges experience a smaller energy loss than particles that penetrate through the full ridge. In this way, energy spread and consequently range variations are introduced to cover the target volume in depth with a homogenous dose distribution. The spread out Bragg peak (SOPB) may have a flat top as in the case of protons or may have a decreasing top in order to compensate for the RBE variation as it is the case for heavy ions. The shape of the wanted dose profile determines the shape of the ridge filter. So each filter gives not only a certain range variation but has also an intrinsic RBE acceptance.

In general, with these passive techniques, treatment volumes can be produced that contain the complete tumor volume. But also a large amount of normal tissue is inside the high-dose area and is hit by the stopping particles of great effectiveness limiting the dose to the tumor by unwanted side reactions. In addition, the passive beam-shaping materials in front of the patient produce neutrons that are scattered in forward direction into the patient which can induce

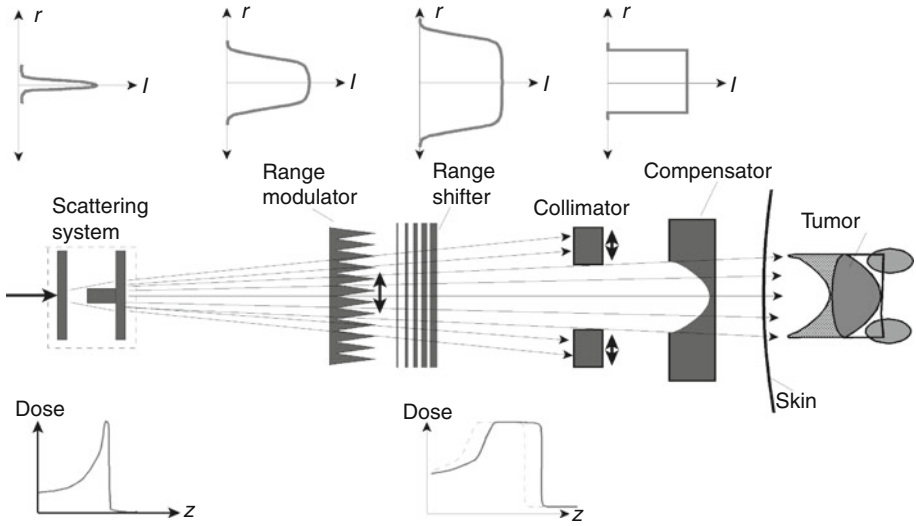


Fig. 5

Principles of passive beam application: the pristine ion beam is widened laterally by a set of inhomogeneous foils that provide a flat top in the lateral profile. Range modulators produce the extension in depth. Collimators restrict the outer contours to the largest cross section of the target and compensators can be used to for the distal contours. A general review of the methods used for beam shaping is given in (Chu et al. 1993)

secondary tumors. Therefore, more conform beam delivery techniques have been developed in order to minimize the amount of normal tissue exposed to the radiation.

3.2 Active Beam Delivery

Active shaping systems use the possibility of deflecting charged particles by applying magnetic fields and were introduced into clinical practice at paul scherer institut (PSI) in Villigen for protons (Pedroni et al. 1991) and at GSI in Darmstadt for carbon ions (Haberer et al. 1993).

The principles of active beam shaping are illustrated in *Fig. 6*. The target volume is dissected into layers of equal particle range, iso-energy layers. Using two deflecting magnets driven by fast power supplies, the "pencil beam" is scanned in a raster-like pattern over the layers, starting with the most distal one. After one layer has been painted, the energy of the beam and consequently the range is reduced and the next layer is treated. The beam may be delivered in discrete "spots" with minimal overlap along the raster path (PSI technique) or nearly continuously in largely overlapping pixels as in a raster scan technique (GSI method). The intensity of the beam is continuously monitored and the dose delivered at each location is controlled by monitors installed directly in front of the patient.

In treatment planning, the number of stopping particles for each pixel has to be calculated, which is a nonlinear problem because of the shape of the Bragg maximum. The more proximal layers are already partly covered with dose when the distal layers are treated. Consequently, these layers have to be covered in the next steps with a smaller, usually nonuniform particle

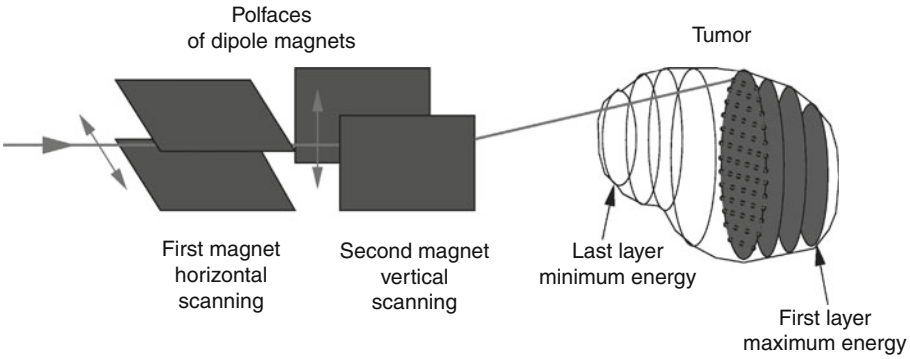


Fig. 6

Principle of active scanning: the target volume is divided in layers of equal particle range and each layer is covered by a grid of individual pixels that are treated sequentially. In practice, 30–60 iso-energy layers are used filled with 5,000–50,000 (all layers) pixels that are delivered with beam in 2–6 min

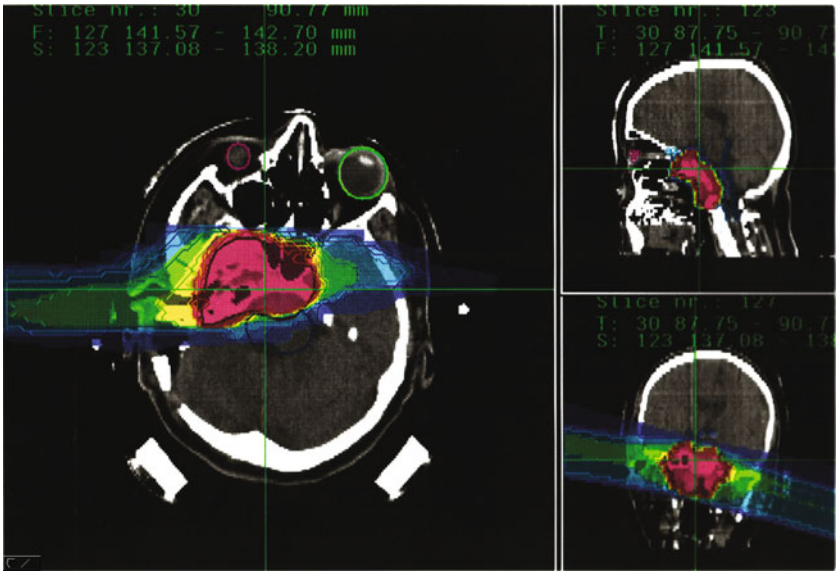


Fig. 7

Treatment plan of a clival chordoma as produced by IMPT: Critical targets like the eye balls, optical nerves, chiasm, and brainstem can be spared and the dose conformed to the target (Jäkel et al. 1999)

distribution in order to reach a uniform dose or a uniform biologically effective dose over the overall target volume.

Although each layer may differ in contour, the shape of a target volume can be "filled" with high precision. Mostly, the different layers are of great complexity as is shown in *Fig. 7*, which depicts the transverse map of beam intensities in each of a set of energy slices for the

treatment with carbon ions of a patient at GSI. The small spots of high energy or of low energy in the beginning and at the end of scanning are due to the density inhomogeneities of the tissue. Small areas of high density like bones have to be treated with higher energies. This accounts for the small spots of higher-energy irradiation. Vice versa, areas of low density have to be filled with low-energy particles.

Using cyclotrons, the energy variations have to be performed using passive degraders after which the beam has to be cleaned from the large energy spread. Then more than 95% of the primary beam is lost producing a large amount of neutrons. Using synchrotron the energy can be requested step by step from the accelerator. At the GSI system 255 different energies are stored in a library and could be requested from pulse to pulse. In addition, for the scanning procedure intensity (five steps) and beam diameter (seven steps) could also requested pulse by pulse from the accelerator. So the beam delivery and the accelerator form an interactive system for intensity-modulated particle therapy (IMPT). A separate optimization of the individual components is not sufficient to get the system to the necessary performance. At the end, a fast interplay between all components has to be reached.

4 Detectors and Quality Assurance

The main challenge of the tumor conform irradiation is the quality control before, during, and after the beam delivery. The intensity-modulated therapy applies a large dose at the end of the beam in the short time where the beam is moved over the individual pixels of the target volume with a velocity in the range of 10 m/s. Any dislocation to normal tissue outside the target volume would cause severe side effects. Therefore, the beam has to be controlled during application in real time but also the planned dose distributions have to be verified before exposure. See Chap. 12, "Tracking Detectors."

To fulfil these very different tasks, a variety of detectors are used: combinations of ionization and multiwire chambers as online detectors in front of the patient that control the performance, the online PET system as a retrospective range control, and finally external dose monitor systems like films, thermoluminescent detectors, nuclear track detectors mostly CR 39, but most important ionization chambers in water phantoms for the dose verification before treatment. See Chap. 12, "Tracking Detectors."

4.1 Therapy Online Monitors

The energy of each pencil beam is controlled by the settings of the magnets of the beam line from the accelerator to the treatment area. In this "quasi-spectrometer," only the beams of correct mass and energy are transmitted. Therefore, these parameters are not controlled during exposure but the magnet settings are calibrated before.

In contrast, it is necessary to monitor the beam position and intensity for each pixel online. From the treatment plan a number of stopping particles is assigned to each pixel. When this number is reached the beam has to be moved to the next pixel of the same iso-energy layer. When the distance of the centers of the pixels is much smaller than the beam width, the beam has not to be turned off when moving. This was realized at the GSI system. When the pixels are separated by a beam half width then the beam has to be turned off when moving from one to the next pixel (PSI system).

For the local control of the beams intensity and location two possibilities exist: position-sensitive ionization chambers and the combination of a large parallel plate ionization chamber with a multiwire chamber.

The position-sensitive ionization chamber was designed and tested at the Berkeley therapy system (Chu et al. 1993), but a position-sensitive ionization chamber is always limited in its spatial resolution by the number of pixels in which the detector is segmented. The main problem seems to be to connect the inner pixels to an amplifier separately for each position because these wiring introduces a non-tolerable inhomogeneous absorber layer for the beam.

For a scanning system, a spatial resolution of 1 mm is the goal that can be reached when the function of intensity and location monitoring are separated. This can be easily done because it can be assumed that the beam is more or less Gaussian shaped and the target will not be hit by two beams in parallel at the same time. So the intensity measured in a broad parallel plate ionization chamber can be attributed to the one beam location determined in a position-sensitive wire chamber. In the GSI pilot project both detectors had a sensitive area of 20 cm × 20 cm and were located in front of the patient (❯ *Fig. 8*).

The parallel plate ionization chamber consists of a counting anode foil und two cathode foils (connected to HV) with an active gap of 2 × 5 mm where a mixture of Ar (80%) and CO_2 (20%) circulates at normal atmospheric pressure. The chamber was operated up to a voltage of 2,000 V in order to minimize saturation effects of the high ion track ionization densities. But the actual pressure and the gas filling and the active counting length depend on the particles energy, atomic number, and the pulse intensity. In order to reach a good signal above the background, the beam intensity cannot be too much lowered. For instance, rescanning techniques where the beam is scanned many times over the target area with diminished intensity are frequently beyond these limits of operation.

The sampling time of the ionization chamber was with ca. 10 μs 10 times faster than the wire chambers where analysis times of ca. 100 μs were necessary. The spacing between the wires was



Fig. 8

Patient set up for therapy: The patient is immobilized with a thermoplastic mask fixed to the patient couch. At the *left* side, the beam exit window is visible followed by three ionization chambers and two wire chambers for online beam localization (photo: Otto, GSI)

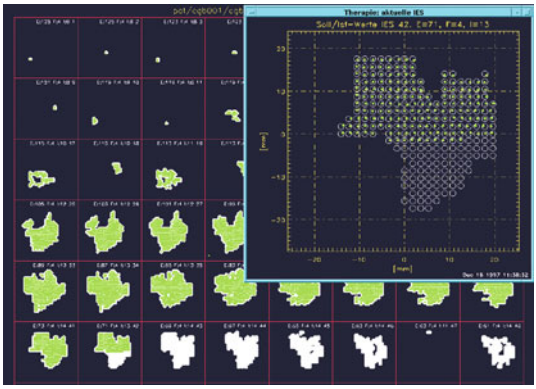


Fig. 9

Compilation of the different range slices of a treatment volume. In each panel one slice is shown; in the magnified panel, the circles represent the calculated center positions of the beam that are filled with the measured center of the beam. The beam diameter is larger than the circles, and overlap over many positions yields a homogeneous distribution

1 mm yielding a spatial resolution of the center-of-mass position of the beam profile of 0.5 mm. In 5 mm distance from the anode grid, the x and y cathode wires were positioned 5 mm on opposite sides to the anode and perpendicular to each other. They collect the electron avalanche produced by the high electric field close to the wires. High voltages of a ca. 2 kV produced an amplification of the signal in the same gas mixture as used for the ionization chamber but 4% heptan added.

The ionization chamber signal is used to monitor the beam intensity and to shift the beam to the next spot when enough intensity is given to the previous one. The beam intensity was regulated before to a level that at least three position measurements should be performed at each pixel. Because of Poisson statistics at least two of them have to be at the correct center region of the pixel to allow continuing with irradiation, see Chap. 5, "Statistics." If the position measurements and the intensity were out of these limits the beam delivery has to be interrupted within less than half a millisecond. Such an interlock and its reason are indicated at the control panel and a decision is asked from the operator to reset the interlock and to continue at the pixel where the failure occurred. The information of the ionization and the wire chamber can be also combined and displayed in two dimensions showing in real time the actual location of the beam in comparison to the planned beam positions. In *Fig. 9* the therapy online monitor (TOM) is shown.

4.2 Detectors for Permanent Recording: Films and Nuclear Track Detectors

In conventional therapy, the most frequent used passive detectors are films. Films have a good spatial resolution, so field size and shape are recorded. Specially manufactured X-ray films have a large intensity range where the film blackening is proportional to the dose. In consequence, the applied dose can be quantified by measuring gray values. In addition, films represent a permanent recording tool because the developed silver grains in a film emulsion are stabile over

years and do not faint. For legal reasons, films of the clinical exposition of a patient have to be stored over 30 years as record of the treatment. This official regulation has the consequence that even all digital exposure data finally have to be converted to films and stored.

For particle therapy, the situation is more complicated because of saturation effects (quenching) in each particle track. As before, films can be used as a fast tool to record and visualize the irradiation field and the contours of structures inside the field. But the blackening of the film is not proportional to the particle dose, i.e., the energy loss of the ions. With decreasing particle energy the local electron density increases, ❯ *Fig. 1*. Then more local energy depositions occur in a grain than necessary to initiate the transition from silver bromide to silver. This local "overkill" effect is similar as observed in the case of DNA breaks after densely ionizing irradiation.

In ❯ *Fig. 10* the recording of a particle beam in a film is compared to the depth–dose curve (Bragg curve). In the film measurement, the Bragg maximum nearly disappears and the measured signal is no longer proportional to the energy deposition, the dose. In the lower panel of ❯ *Fig. 10* the film efficiency is given as function of the particle energy for both protons and carbon ions. Already below 100 MeV/u the efficiency significantly decreases and values < 1 are measured. The theoretical curves in the figure are calculations according to the local effect model (LEM) that was developed for biological systems but can be also applied for nonbiological systems when the parameters of the X-ray response and the size of the critical target (the silver grains) are correctly incorporated into the calculation. LEM will be discussed in some detail in the biology part.

In ion beam therapy, films are frequently used because of a fast recording of the relevant geometries. But films cannot be used to determine the applied dose. However, when the particle field is known in its composition in energy and atomic number, the expected film response can be calculated and compared to measurement. In this case a quality assurance is possible in retrospective.

A second type of detectors frequently used in ion beam therapy are nuclear track detectors like CR 39. These detectors record single-particle tracks: after exposure, the Cr 39 plates are etched for a few hours in 6 m Na-OH solution in order to produce visible holes along the tracks of individual particles. CR 39 is a great standard because the individual holes can be counted under a microscope and the number of particles per square unit can be measured and used for an absolute calibration of other detectors, see ❯ Chap. 12, "Tracking Detectors."

But the great advantage of CR 39 is that the diameter of these holes depends on the energy deposition in the material. Therefore, higher doses at the end of the particle range become visible as greater scattering centers for light. In this way the patterns of different dose distributions in a target volume can be visualized although the distribution of the reflected light is not strictly proportional to dose. In ❯ *Fig. 11* a spheroid dose distribution is shown as produced in a water tank as tissue equivalent.

In contrast to the films, the CR 39 stacks are much more elaborate to produce and are not used routinely because of the long time needed for the development. Films are used routinely for an overall check of the patient setup and also as a retrospective quality control.

4.3 Ionization Chamber Dosimetry

The only dosimeters that are widely independent from quenching effects are ionization chambers operated at sufficient high voltage. For the quality control of the applied dose of a treatment

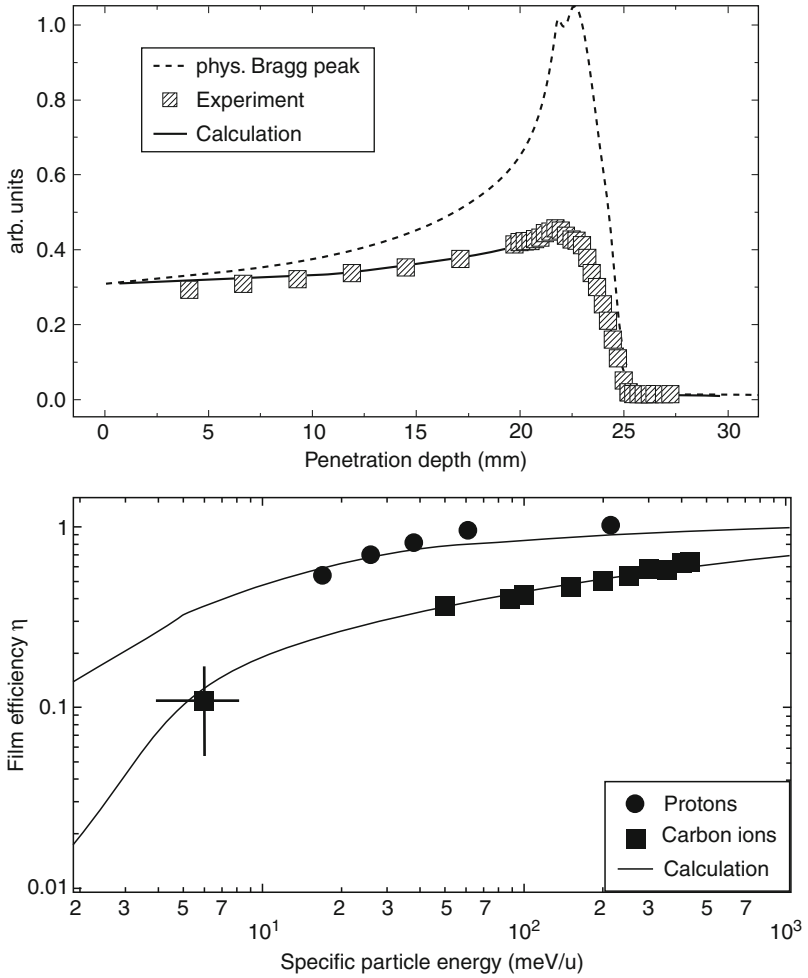


Fig. 10

Film response to particle irradiation: at the *top* a Bragg curve (upper curve) is recorded with films and compared to LEM calculations (lower curve). At the *bottom* the film effectiveness is given for protons and carbon ions as function of the particles' energy (Spielberger et al. 2003)

plan the critical points were verified in a water-phantom measurement: critical for instance are the gradients between target volume and sensitive structures, the range of the most distal part, the ratio between entrance and target dose. For these tests the treatment plan has to be transferred to a water phantom, which means that density variations in the patient are translated to different water-equivalent lengths along the path of the beams, calculating the corresponding dose in a cubic water phantom. Then the "fingers" of a 24 finger thimble ionization chamber (so-called pinpoint chambers) are located in the water phantom at these critical points using a

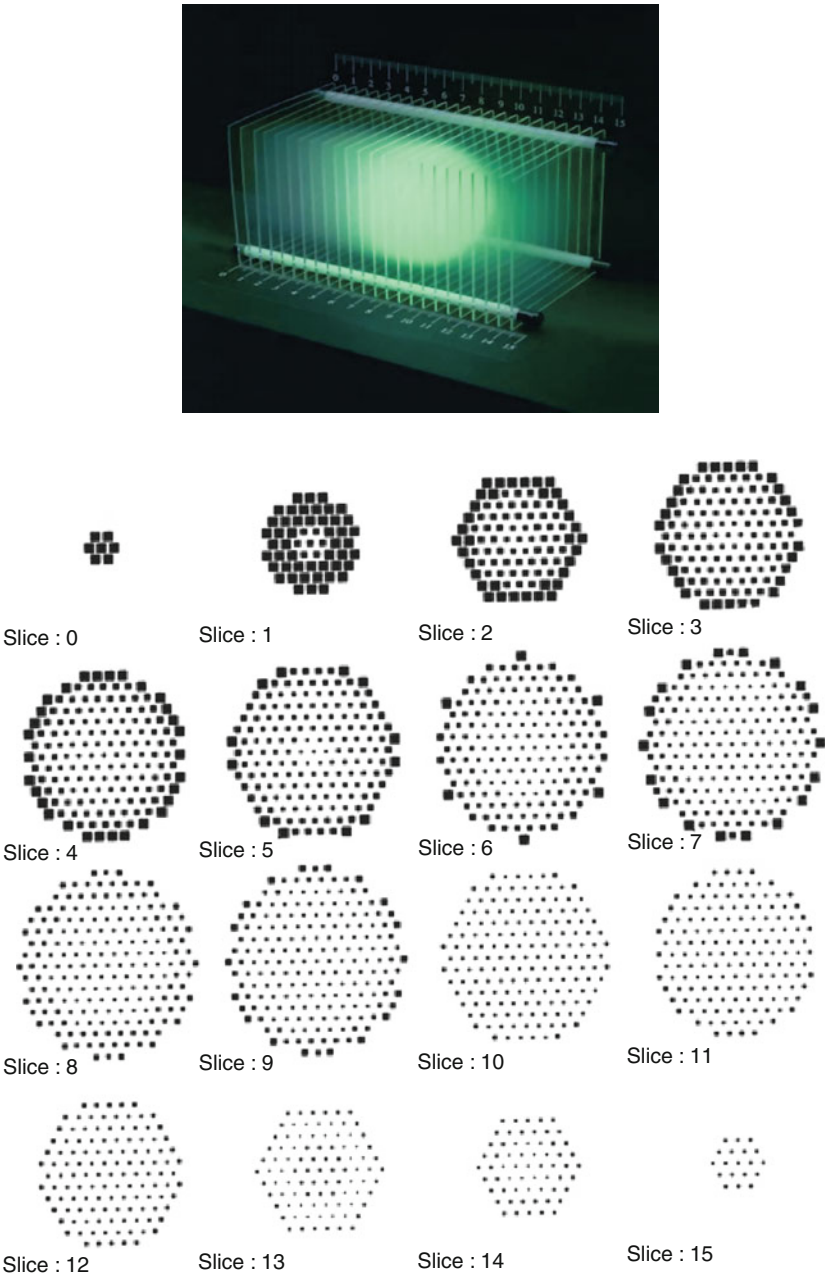


Fig. 11

3-D dosimetry: A spheroid of 6 cm in diameter and in a mean water depth of 11 cm is shown at the *top*. The 16 iso-energy layers were independent from the position of the CR 39 plates. The pencil beam was applied from the left side only with approx. 5 mm diameter. At the bottom a fluence distribution of the different iso-energy layer is given for a 2 cm sphere where the structure becomes more visible

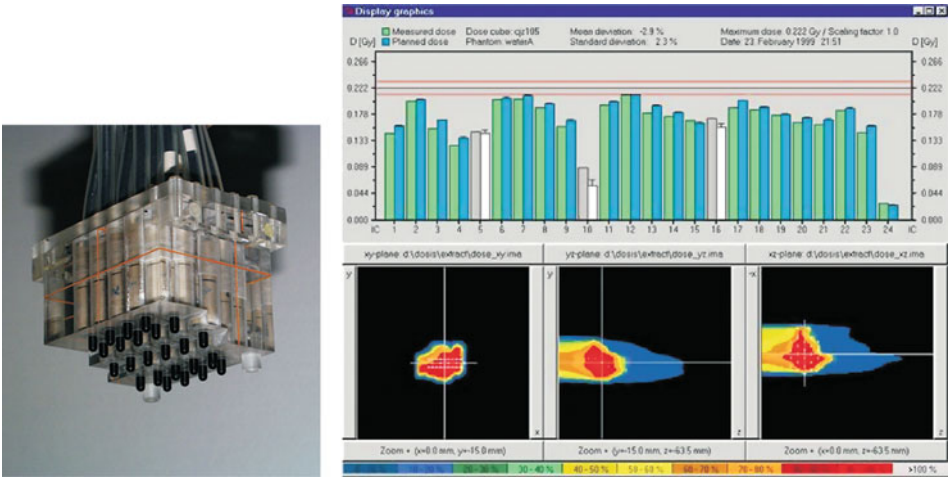


Fig. 12

Dosimetry setup: *Left*: a set of thimble ionization chambers are used for the control of a treatment plan, *right*: measurement of critical points in a treatment plan: Position of the measurements as well as the comparison to the planned doses are given (Karger et al. 2006)

3D drive. Because of the sequential way of the beam application with the raster system a complete plan has to be applied and the doses in the ionization chambers have to be integrated in order to obtain the full dose in every sampling point. Therefore, for this QA procedure, an exposure time similar to the patient treatment later on is necessary. In *Fig. 12* a photograph of the dosimeters as well as the position of the measurement point are given. According to legal regulations maximum dose deviations of plus 5% and minus 3% are allowed compared to the planning. In case of larger deviations the beam application plan has to be redone and controlled by a new measurement.

In addition, the ionization chamber measurements have to be a part of the plan verification where a standard plan has to be verified to control the performance of the system. The results of these measurements together with the irradiation protocols have to be stored for 30 years after patient treatment (Jäkel et al. 1999).

5 Biological Properties of Heavy Ions Relevant for Therapy

The problem of any radiation therapy is to find the optimum way to kill all tumor cells and to spare normal tissue as far as possible which is the problem of exact targeting and the effective biological interaction. Carbon ions are most close to a perfect solution of this problem: as demonstrated before their dose distribution is superior compared to other modalities. But carbon ions exhibit also a greater biological effectiveness in the target. In contrast to the previously used neutron beams where the relative biological effectiveness RBE is increased over the total beam range, for carbon ions the RBE variation is a differential effect: in the entrance channel

the effectiveness stays close to sparsely ionizing radiation like photons or protons but at the end of the particles range it can be drastically enhanced.

5.1 Definition of RBE and Its Dependence on Dose or Effect Level

Therapy-relevant information on RBE originates from so-called survival experiments where the clonogenic capacity of the cells is measured as a function of dose. In the experiments aliquots of cells are exposed to different doses of different radiation and the number of cells as determined by their colony-forming ability is plotted in a semilogarithmic way as function of the dose.

Experimentally it is found that densely ionizing radiation like α particles or heavier ions generate a greater biological effect for the same dose than X-rays. In ❯ *Fig. 13* a schematic dose-effect curve for cell inactivation is compared to that of X-rays.

The X-ray dose response is characterized by a bi-phase behavior: at low doses the response curve has a large shoulder and at high doses a steep exponential tail. In practice, the graph is approximated by a linear–quadratic response curve

$$S = S_0 \exp\left[-(\alpha D + \beta D^2)\right]. \tag{4}$$

S/S_0 is the normalized survival, α and β are the coefficients for the linear and quadratic term of the dose D. For particles, the quadratic compound decreases with increasing ionization density and the "survival curve" is characterized for great linear energy transfer (LET) values by a pure exponential decay in dose. In order to compare the different radiation qualities the relative biological effectiveness RBE was defined as the ratio of X-ray dose to particle dose that is necessary to achieve the same biological effect,

$$RBE = \frac{D_X}{D_{ion}} \tag{5}$$

with D_X and D_{ion} being the X-ray and ion doses, respectively (Hall 1994).

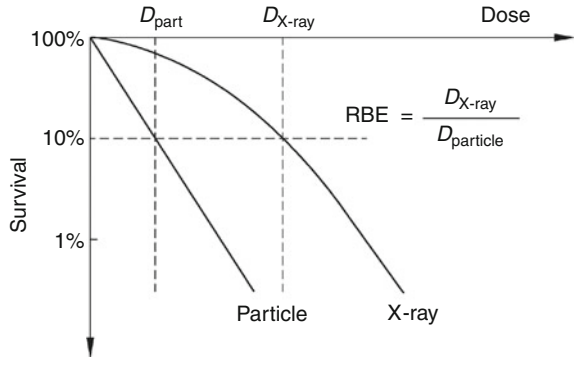


Fig. 13

RBE Definition: The cell survival is given as a function of dose in a schematically way for X-rays and particle radiation together with the definition of the relative biological effectiveness

Because of the nonlinearity of the X-ray curve as reference, RBE strongly depends on the effect level: At very high doses, X-ray and particle response curves usually parallel each other and RBE approaches a value close to one. Toward lower doses, i.e., greater survival RBE increases and reaches its largest value in the limit of zero dose. This value is called RBE_{α} because it is the ratio of the initial slopes of the ion curve divided by the X-ray curve.

For patient treatment, the RBE–dose dependence is an important fact and has to be taken into account in each iteration step of the planning when the general dose and even the local doses are changed. It is not possible to double the physical dose in order to double the wanted biological effect. For the greater dose the RBE decreases and the final effect becomes smaller than the double of the primary effect. Therefore, in an optimization process where doses are changed, the RBE calculations have to be iterated too. Other important dependencies are those from physical factors such as particle energy and LET.

5.2 The RBE Dependencies on Physical and Biological Parameters and the Molecular Understanding

As DNA is the main target for cell inactivation by ionizing radiation all the dependencies of RBE on the various parameters become at least qualitatively evident from the mechanisms of DNA damage: at low local doses (X-rays and protons) mainly isolated damage such as single-strand breaks (SSBs) are produced (*Fig. 1*). But mammalian cells have a very efficient repair system for this type of damage that occurs very frequently in daily life and is not only caused by ionizing radiation. Even simultaneous damage at both DNA strands, i.e., double-strand breaks (DSBs) can be repaired by the cell rather quickly with high fidelity. But if the local damage is enhanced by higher local doses more complex damage, i.e., clusters of double-strand breaks are produced which are less reparable.

For particles, the amount of irreparable damage depends very much on the local dose, which is created by the energy deposition of the electrons inside the tracks, which varies over many orders of magnitude from the inner to the outer parts of a track but which also depends on the atomic number and the particles energy.

In general, there is no difference whether an electron is created by photon impact or by the impact of heavy charged particles, but there is a big difference in the spatial distribution between electrons being created along a track of a heavy charged particle and the random distribution of electrons from photon (or proton) beams and there is also a difference in ionization density between particles of different energies and atomic numbers. In general, it can be said that tracks with greater local ionization densities produce a larger biological effect until a saturation value of ionizations is reached where "overkill" is produced, i.e., more damage than necessary for cell killing.

As a consequence, the biological effectiveness depends at the physical side from the local ionization density: higher densities and consequently ions with higher atomic number produce greater damage. Because heavier particles have in general greater energy loss values, they are more efficient than lighter particles.

When an ion along its path is studied it will have a lower effectiveness when the energy is larger because of the lower ionization densities. When it slows down the energy loss, i.e., the dose but also the effectiveness increases until a point where too much damage is produced and the local dose might increase further but the effectiveness decreases (overkill effect).

Finally, the RBE depends on the biological repair capacity of the cells. Cells that are very sensitive to radiation injury do not show a significant difference in the radiation response between densely and sparsely ionizing radiation and therefore no enhanced RBE for ion beams. But for cells with high repair capacities the greater local damage produces also a potentiated effect. Therefore, the very radio-resistant tumors are most suited as candidates for heavy-ion therapy.

In *Fig. 14* the Bragg curve for protons, carbon and neon ions are compared to the RBE dependence. For protons, the RBE increases at the distal side, for neon ions already before the Bragg maximum. Only for carbon ions the Bragg maximum coincides with the maximum of

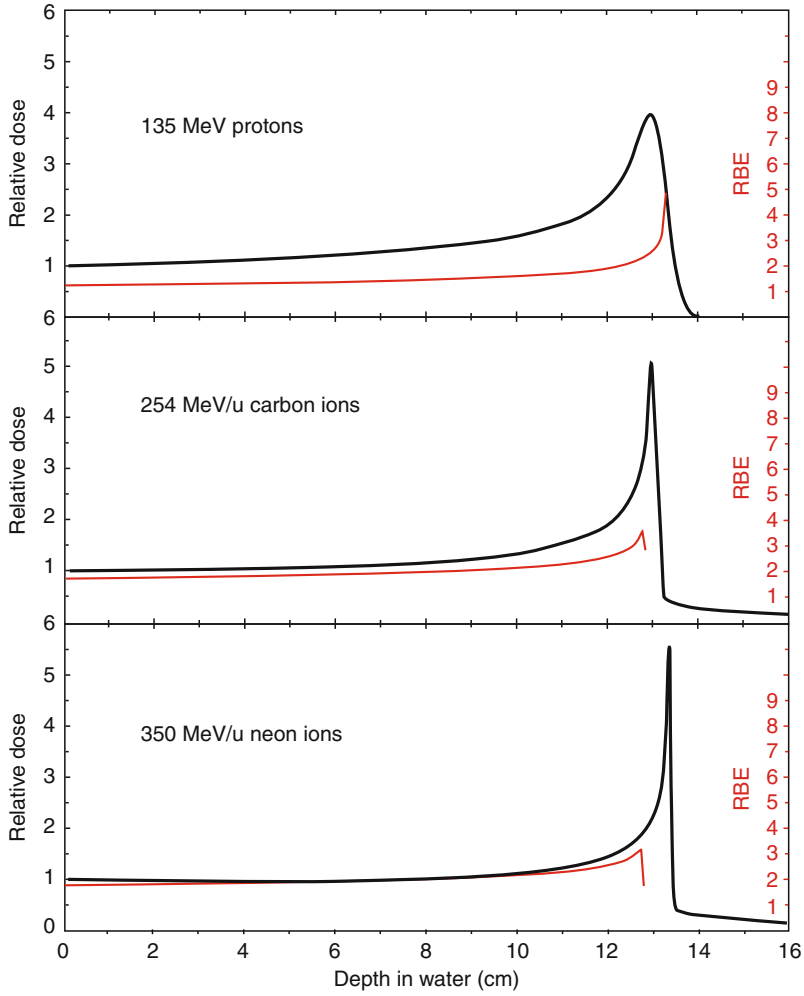


Fig. 14

Comparison of the depth–dose distribution/Bragg curve (left scale) of protons, carbon and neon ions with the local RBE value (right scale). Only for carbon the Bragg maximum coincides with the maximum of the RBE. For protons the RBE increases at the distal side, for Neon already before the Bragg maximum (Weber and Kraft 2009)

the RBE. Therefore, the overall RBE contribution in proton therapy has been approximated by a global dose factor of 1.15 although in cell experiments RBE value up to 6 have been measured. For neon ions the high RBE in front of the Bragg maximum causes severe side effect in the normal tissue. But carbon ions are the optimal candidates for therapy because the high dose in the Bragg maximum is potentiated by a high RBE.

6 The Planning of the Biological Effective Dose

It is the finite goal to achieve a homogenous inactivation over the complete target volume independent from the variation of the physical composition of the local radiation field. In practice the biological effective dose, which is defined as BED [GyE] = dose [Gy] × RBE should be constant over the complete target volume. (Although RBE has no dimensions, BED is mostly given Gray equivalent GyE or Cobalt Gray equivalent CGyE. According to new ICRU recommendations Gy(RBE) should be used.) In order to compensate for the physical field variations local RBE values have to be applied. In principle also the sensitivity variation of the biological tissue should be taken into account, but these variations cannot be detected with the present diagnostic techniques.

According to the various application methods different ways are used to assign local RBE values: Using passive beam-shaping systems the RBE dependence has to be inbuilt in the shape of the range modulators and is fixed (Chu et al. 1993). For active beam-shaping systems, the higher conformity with the target field has the consequence that RBE varies over the complete target volume in three dimensions (Weyrather and Kraft 2004).

At Berkeley, in vitro data from cell cultures were used for RBE determination. In many experiments, cell survival was determined in pristine and also in spread out Bragg peaks of variable depth and width. These measurements were used to design a set of ridge filters for therapy of the tumor sites in the corresponding depth and extension. However, the assumption is implicit that these in vitro data of special human cell lines should be valid for all the human tissue affected by irradiation (Blakely et al. 1980).

At the National Institute for Radiological Science (NIRS) in Chiba, the basis for the RBE incorporation was again cell experiments using human salivary gland HSG cells (Kanai et al. 1997, 1999). It could be shown, that the RBE in the mid of the spread out Bragg peak of carbon ions (at a mean LET of 80 keV/μm) was close to neutron cell irradiation. Consequently, the clinical dose prescription and the clinical RBE values of carbon ions were taken from the large clinical experience of neutron therapy and not from the cell experiments (Tsujii et al. 2004).

At GSI and also for HIT and the other upcoming units the RBE is calculated separately for each voxel of the treatment volume. The basis for the calculation is the local effect model (LEM). In this model, the biological response to particle radiation is the convolution of the induction probability of lethal damage as function of the dose measured with X-ray irradiation with the different values of the radial dose distribution integrated over the size of a cell nucleus (Scholz and Kraft 1994),

$$S_{\text{ion}} = \exp\left(-N_{\text{lethal}}\right) \tag{6}$$

with

$$N_{\text{lethal}} = \int_V \frac{\ln S_X(D)}{V}\,\mathrm{d}V, \tag{7}$$

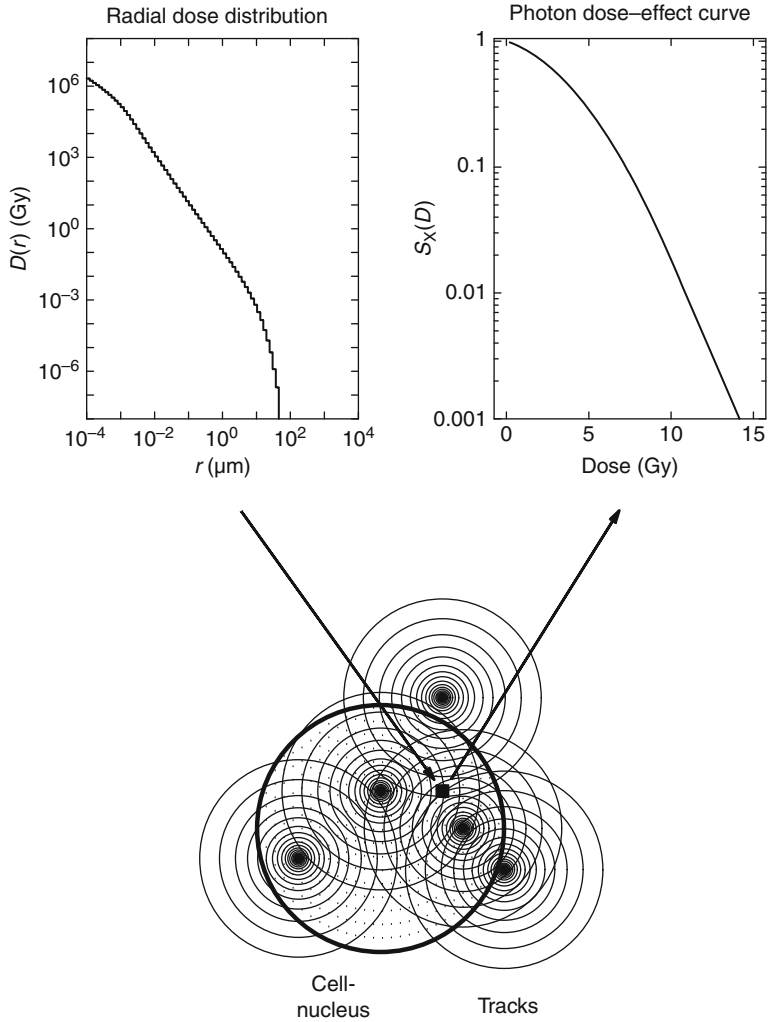


Fig. 15

In the local effect model (LEM) the cell nucleus is exposed to a number of carbon ions necessary for the macroscopic dose, and local doses are calculated according the radial dose distribution of the particle tracks. Then the dose of small compartment of the nuclear volume is calculated and the probability to create a lethal lesion in this area is calculated according to the X-ray dose effect taken at the dose point. Finally the probabilities are summed (integrated) over the complete cell nucleus (Scholz and Kraft 1994)

being the number of lethal lesions, $S_X(D)$ the X-ray dose–effect curve for cell killing, V the nuclear volume, N_{lethal} the number of lethal lesions, and S_{ion} the calculated survival after ion exposure (*Fig. 15*).

It should be explicitly noted that in this method the X-ray dose–effect curve is transformed in particle dose–effect curves: With this formula, cell inactivation by carbon ions can

be calculated from the corresponding X-ray curve. But the cell RBE data are not transferred to therapy directly. It is the X-ray response of the tumor that is transferred to the clinical situation: Tumor response to carbon ions are calculated from the corresponding clinical dose response curve for the same type of tumor against photons.

This type of calculation was used for the tumor therapy at GSI and it worked well in terms of tumor control for more than 400 patients: There were no recurrent tumors and no necrotic areas in the tumors. Evidently the RBE for the tumors was calculated with good precision, but it was also evident that the effect at the low doses in the entrance channel was overestimated because the RBE was too large. Therefore, LEM was refined in several steps: first the radical diffusion from the inner part of the track to the outer regions was included (LEM II) , then the energy variation of the dose cut off in the track center and the interaction of single-strand breaks forming double-strand breaks (LEM III) and finally the interaction of the double-strand breaks were included that form a clustered lesion (LEM IV). With these refinements LEM could be improved to produce the correct RBE values in the entrance channel and also for light ions as for instance protons (Elsässer et al. 2010).

7 Quality Assurance and RBE Detectors

For the conventional therapy using photons it is believed that the biological effect is proportional to the dose and the dose is monitored for quality control within limits of +5% and −3%. For particle therapy the same dogma is believed for protons except that a RBE factor of 1.15 is applied as global dose correction and the dose again is controlled using the standard equipment of ionization chambers or films with an accuracy of a few percent. But for heavy ions the biological effective dose is the essential parameter which is the product: dose x RBE. The RBE depends critically on many variables as shown before. Whatever model is used for the RBE prediction, the accuracy of this calculation can only be verified in specific biological experiments before it can be used in therapy.

The closest approach to the patient is experimental tumors in animals. But these experiments are very expensive and time consuming and mostly of limited analytic power because of the low statistics and also some variability within an animal population. A large-animal experiment for testing the tolerance of the spinal cord has been carried out using rats that are close to the neural situation in humans (Karger et al. 2006). Because LEM can be used to convert an X-ray dose–effect curve to particles, in a series of experiments the influence of the radiation quality and the fractionation on the induction of paresis has been measured in rats and compared with the prediction of LEM. The experiments showed a satisfactory agreement with the LEM calculation. But the experimental effort necessary to perform the experiment was very high when for instance 50 rats got one daily fraction over 18 consecutive days. This was the fractionation scheme for the patients, but the experiment demonstrated that more simple tests have to be developed as a standard routine.

A first, one-dimensional experiment was set up with a stack of plastic sheets that are immersed in a tank completely filled with cell medium (*Fig. 16*). The cells are attached to the sheets in a sufficient density and the beam should penetrate perpendicular through the cell layers from one or two sides so the cell layers form equidistant range slices of beam penetration. The stack in the medium tank should be longer than the beam range but the lateral extension of the beam should be greater than the area covered by the cells. After exposure,

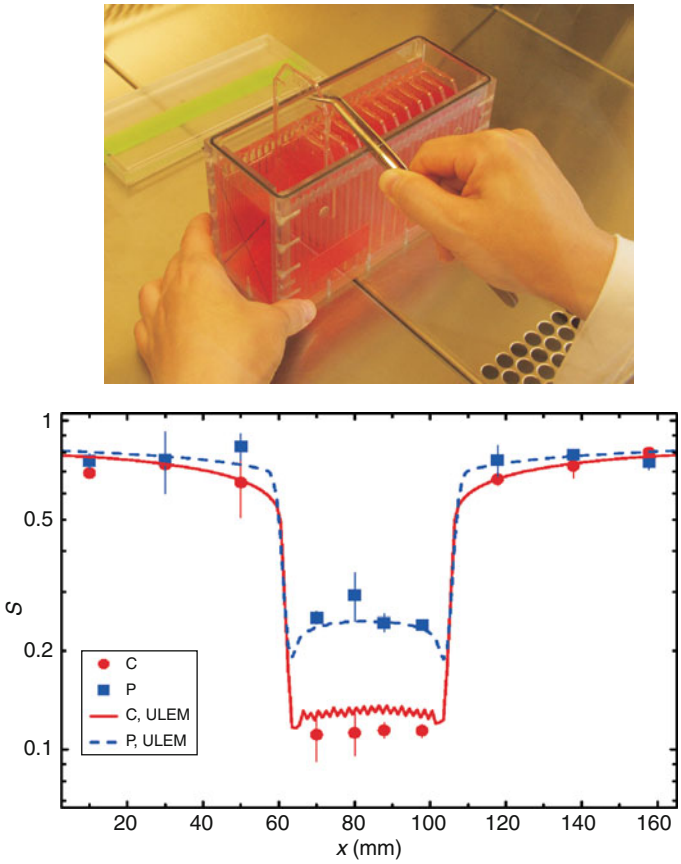


Fig. 16

One dimensional RBE test system: *Top*: for the experiment plastic sheets are immersed in a tank filled with medium and exposed to the particle field from left and right side simulating the exposure of a tumor in the center of the head. At the *bottom* the survival as function of depth is given: for the same effect in the entrance channel, the carbon ions are twice effective in the target region (Elsässer et al. 2010)

the cells are processed each slice separately in the usual way for the clonogenic survival test. The measured survival as a function of depth can be directly compared to the theoretical prediction. In *Fig. 16* the survival as function of depth was measured for a proton and a carbon field simulating the exposure of a tumor in the center of the head and compared to LEM IV calculations using the same set of parameters for both the carbon exposure and the proton exposure. The calculations are in very good agreement with the reality (Elsässer et al. 2010).

This method of using cultured cells for measuring the effectiveness of therapeutic beams can be extended to two dimensions when the cells are attached to thin rods instead of the plastic sheets (Krämer et al. 2003a). The spatial resolution for this setup is given by the extension of the rods and can reach 4 mm. *Figure 17* shows the experimental set up and two rods with cells growing at the surface. *Figure 18* shows the treatment plan and the results of an experiment

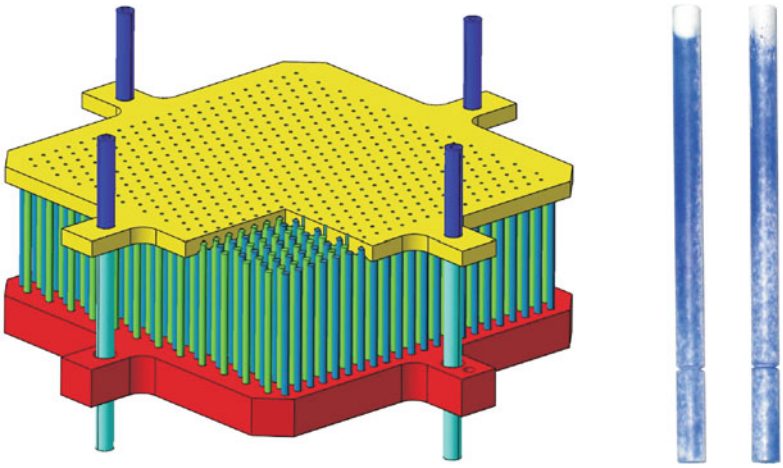


Fig. 17

Detector for the biological verification of a treatment plan in two (three) dimensions, which consists of up to 400 rods fixed between two plates. These rods of tissue equivalent material are distributed in a medium-filled tank. The granularity of the measurement depends on the diameter of the rods. As shown at the left side, cells are grown at the surface of the rods and have to be worked up for each rod independently after exposure. Patent: PCT/EP 2007/002156, Cläre von Neubeck et al.

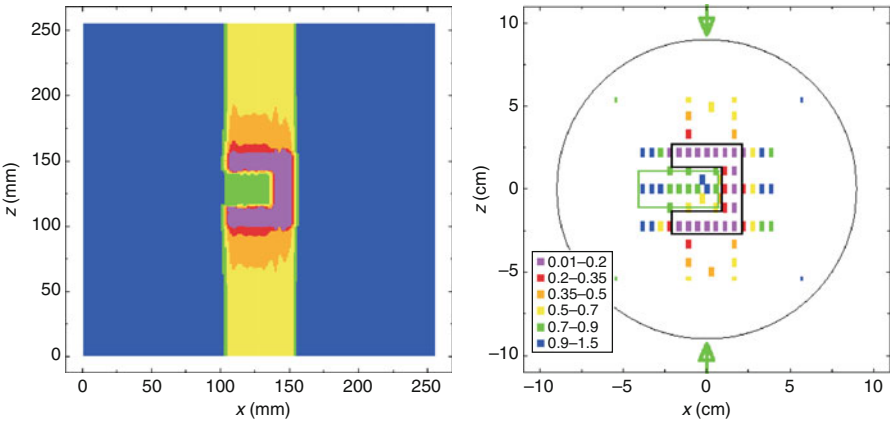


Fig. 18

Two-dimensional test of a treatment field where a critical volume in the center part should be spared. The beams are applied from the top and the bottom, color coding in the planed (*left*) and measured (*right*) are the same. The granularity is about 4 mm

where a complex pattern simulates the geometry of an irradiation of a tumor from two sides but sparing a critical volume in the center part that was adapted from the irradiation of a brain tumor sparing the brain stem. The difficulty of experiments of this type is that all approx. 150 rods have to be processed in the same standard procedure in parallel. But the average time

necessary for the clone formation is 1 week, which means that the result of the experiment is available after that time but much faster than in any animal experiment (Krämer et al. 2003b).

8 Conclusions

Intensity-modulated particle therapy (IMPT) exploits the physical and biological advantages of carbon beams to a maximum extent: The increase of the dose toward the end of the range and its high biological effectiveness enable to inactivate tumor cells with a high probability. Because of the small scattering and the high precision of application normal tissue and critical structures can be spared to a large extent. Low side effects are combined with high tumor control rates reaching up to 100% for instance in the case of chordomas in the base of skull (Elsässer et al. 2010).

But all procedures of IMPT have to be subjected to higher standard of quality control than in conventional therapy: The treatment plan has to be verified in three dimensions before application to the patient, the beam delivery has to be monitored online in real time, i.e., in a sampling time of a few microseconds for the particle fluence and in 100 μs with a spatial resolution of 1–2 mm, the distribution of the stopping particles inside the patient can be monitored from outside in a noninvasive way by means of PET techniques. Finally, the models for the RBE calculations used to determine the biologically active dose have to be verified using biological dosimetry.

For these tasks the arsenal of dosimeters used in conventional therapy has to be enlarged and novel techniques have to be taken over from other fields of physics. But this development is not at its end regarding many new techniques in IMPT that will be developed like the treatment of moving targets for lung and heart (Kraft and Kraft 2009).

References

Barkas HW (1963) Nuclear research emulsions, vol I. Academic Press, New York and London

Bethe H (1930) Ann Phys (Leipzig) 5:325

Blakely EA, Tobias CA, Ngo FQH, Curtis SB (1980) Physical and Cellular Radiobiological Properties of Heavy Ions in Relation to Cancer Therapy, In: Pirucello MD, Tobias CA (eds) Biological and medical research with accelerated heavy ions at the Bevalac. Lawrence Berkeley Laboratory, CA Berkeley, pp 73

Bloch F (1933) Zur Bremsung rasch bewegter Teilchen beim Durchgang durch Materie. Ann Phys (Leipzig) 5:285

Chu WT, Ludewigt BA, Renner TR (1993) Instrumentation for treatment of cancer using proton and light-ion beams. Rev Sci Instrum 64:2055

De Vita VT, Hellmann S, Rosenberg SA (1997) Cancer: principles and practice of oncology. Lippincott-Raven, Philadelphia

Durante M, Loeffler JS (2010) Charged particles in radiation oncology. Nat Rev Clin Oncol 7: 37–43

Elsässer Th, Weyrather WK, Friedrichs T et al (2010) Quantification of the relative biological effectiveness for ion beam radiotherapy: direct experimental comparison of proton and carbon ion beams and a novel approach for treatment planning. Int J Radiat Oncol Biol Phys 78:1177–1183

Enghardt WK, Fromm WD, Geissel H, Keller H, Kraft G, Magel A, Manfrass P, Mäunzenberg G, Nickel F, Pawelke J, Schardt D, Scheidenberger C, Sobiella M (1992) The spatial distribution of positron-emitting nuclei generated by relativistic light ion beams in organic matter. Phys Med Biol 37:2127

Enghardt WK, Debus J, Haberer Th, Hasch B, Hinz R, Jäkel O, Krämer M, Lauckner K, Pawelke J (1999) The application of PET to quality assurance of

heavy-ion tumor therapy. Strahlenther Onkol 175:S33
Geiss O, Kramer M, Kraft G (1999) Efficiency of thermoluminescent detectors to heavy-charged particles. Nucl Instrum Methods Phys Res B 142:592
Gottschalk B, Koehler AM, Schneider RJ, Sisterson JM, Wagner MS (1993) Multiple Coulomb scattering of 160 MeV protons. Nucl Instrum Methods A 330:467
Haberer Th, Becher W, Schardt D, Kraft G (1993) Magnetic scanning system for heavy ion therapy. Nucl Instrum Methods Phys Res A 330:296
Haettner E, Iwase H, Schardt D (2006) Experimental fragmentation studies with 12C therapy beams. Radiat Prot Dosim 122:485
Hall E (1994) Radiobiology for the radiologist. Lipincott Company, Philadelphia
Hüfner J (1985) Heavy fragments produced in proton-nucleus and nucleus-nucleus collisions at relativistic energies. Phys Rep 125:129
Jäkel O, Hartmann G, Karger C, Heeg P (1999) A quality assurance program for heavy ion treatment planning. Radiother Oncol 51:13
Kanai T, Furusawa Y, Fukutsu K, Itsukaichi H, Eguchi-Kasai K, Ohara H (1997) Irradiation of mixed beam and design of spread-out Bragg peak for heavy-ion radiotherapy. Radiat Res 147:78
Kanai T et al (1999) Characteristics of HIMAC clinical irradiation system for heavy-ion radiation therapy. Int J Radiat Oncol Biol Phys 44:201
Karger CP, Peschke P, Sanchez-Brandelik R, Scholz M, Debus J (2006) Radiation tolerance of the rat spinal cord after 6 and 18 fractions of photons and carbon ions: experimental results and clinical implication. Int J Radiat Oncol Biol Phys 66:1488
Kraft G (2000) Tumor therapy with heavy charged particles. Prog Part Nucl Phys 45:473
Kraft G, Kraft SD (2009) Research needed for improving heavy-ion therapy. New J Phys Focus Heavy Ions Biophys Med Phys 11:025001
Krämer M, Kraft G (1994) Calculations of heavy track structure. Radiat Environ Biophys 33:91
Krämer M, Jäkel O, Haberer Th, Kraft G, Schardt D, Weber U (2000) Treatment planning for heavy-ion radiotherapy: physical beam model and dose optimization. Phys Med Biol 45:3299
Krämer M, Wang JF, Weyrather WK (2003a) Biological dosimetry of complex ion eradiation fields. Phys Med Biol 48:2063
Krämer M, Weyrather WK, Scholz M (2003b) The increased biological effectiveness of heavy charged particles: from radiobiology to treatment planning. Tech Cancer Res Treat 2:427
Molière G (1948) Theorie der Streuung schneller geladener Teilchen II, Mehrfach- und Vielfachstreuung. Z Naturforsch 3a:78
Parodi K et al (2007) Patient tomography and computed tomography imaging after proton therapy. J Radiat Oncol Biol Phys 68:920
Pedroni E, Blattmann H, Böhringer T, Coray A, Lin S, Scheib S, Schneider U (1991) Voxel scanning for proton therapy. In: Itano A, Kanai T (eds) Proceedings of the NIRS international workshop on heavy charged particle therapy and related subjects, Anagawa, Japan
Scholz M, Kraft G (1994) Calculation of heavy ion inactivation probabilities based on track structure, x-ray sensitivity and target size. Radiat Prot Dosim 52:29
Schulz-Ertner D, Tsujii H (2007) Particle radiation therapy using proton and heavier ion beams. J Clin Oncol 25:953
Spielberger B, Kramer M, Kraft G (2003) Three-dimensional dose verification in complex particle radiation fields based on x-ray films. Phys Med Biol 48:497
Tsujii H et al (2004) Overview of clinical experience on carbon therapy at NIRS radiotherapy and oncology. Radiother Oncol 73:S41
Weber U, Kraft G (2009) Comparison of carbon ions versus protons. Cancer J 15:325
Weyrather WK, Kraft G (2004) RBE of carbon ions: experimental data and the strategy of RBE calculation for treatment planning. Radiother Oncol 73:S161
Wilson RR (1946) Radiological use of fast protons. Radiology 47:325

Further Reading

Kraft G (2000) Tumor therapy with heavy charged particles. Prog Part Nucl Phys 45:473
Schardt D, Elsässer Th, Schulz-Ertner D (2010) Heavy-ion tumor therapy: physical and radiobiological benefits. Rev Mod Phys 82:383
Schulz-Ertner D, Tsujii H (2007) Particle radiation therapy using proton and heavier ion beams. J Clin Oncol 25:953
Suit H, DeLaney T, Goldberg S et al (2010) Proton versus carbon ion beams in the definitive treatment of cancer patients. Radiother Oncol 95:3

Index

A

AB. *See* Atomic bremsstrahlung (AB)
Absolute quantitation, 1046
Absorption
  –color center, 536, 539
    –deep, 546
    –density, 537, 538, 543, 546
    –depth, 542
    –formation, 536, 539, 553
  –recovery, 542–543
  –shower counters, 536
Abundance sensitivity, 689, 694, 698
Accelerator
  –as interactive system, 1189
  –luminosity, 94
  –measurements, 603, 624
Accelerator-based muon systems, 478–487
Accelerator mass spectrometry (AMS), 654–665, 690
  –applications
    –archaeology, 654–665
    –environmental, 654–665
    –geology, 654–665
Acceptance, 92, 95, 97
Acceptance–rejection sampling method, 1102–1103, 1114
Accidental coincidences, 947–948, 952
AC coupled silicon sensor, 271
Accuracy, 1045, 1052
Acollinearity, 950, 966
Activated scintillator, 434
Activation, 225
Active beam shaping, 1187, 1199
Active contour, 1052–1054
Active mm-wave systems, 638–639
Adaptive filter, 282, 283
Adaptive radiation therapy (ART), 1167
Adaptive vertex filter, 291
Adaptive vertex reconstruction (AVR), 283
ADC. *See* Analog-to-digital converters (ADCs)
AEC. *See* Automatic exposure control (AEC)
AERA. *See* Auger engineering radio array (AERA)
Aerogel, 457, 459, 464, 465, 469
AGASA, 605–607
Aging effects, 241
AIC. *See* Akaike information criterion (AIC)
Air Cherenkov telescope, 466, 468, 469
Air density, 596, 600, 608
AIRES, 597
AIROBICC, 609
Akaike information criterion (AIC), 1078
ALARA principle, 227, 228, 234, 906
ALEPH Experiment, 87, 93
ALICE, 465, 479, 485
Alkali halide, 549–551
Alpha
  –decay, 189, 209
  –particle, 716
Alternating gradient, 143
Americium-241, 563
AMIGA. *See* Auger muons and infill for the ground array (AMIGA)
Amplifier
  –bandwidth, 39, 40, 46
  –charge sensitive, 44–48
  –gains, 692
  –input impedance, 41, 44, 47, 48, 50, 51
AMS. *See* Accelerator mass spectrometry (AMS)
AMS-02 (Alpha Magnetic Spectrometer–02), 581
Analog-to-digital converters (ADCs)
  –aliasing, 78
  –amplifier, 71
  –conversion time, 73, 79
  –differential non-linearity, 71–73
  –flash, 72, 74
  –parameters, 71
  –pipelined, 74
  –successive approximation, 72, 73
  –Wilkinson, 73–75
Analysis of detector system, 60–64
Analytic simulation, 1105, 1107, 1110, 1111, 1115, 1120, 1121
AND gate, 66, 67
Anger camera, 1129, 1138
Anger gamma camera, 918–923
Angiography, 215, 231
Annealing schedule, 283, 291
Annihilation, 87, 93–96, 418, 419, 1108, 1111, 1117, 1118
  –muon, 93, 94
  –observables, 93
Anode photo-current, 539, 540
Antarctic Impulsive Transient Antenna (ANITA), 587, 588
ANTARES experiment. *See* Astronomy with a Neutrino Telescope and Abyss environmental RESearch (ANTARES experiment)
Anti-aliasing, 78
  –filter, 78
Anticoincidence, 562, 563
APD. *See* Avalanche photodiodes (APDs)

Apparent energy, 413, 415, 417
ARGO-YBJ, 607
Array detectors, 691
Arrays, 63, 70
ART. *See* Adaptive radiation therapy (ART)
Arterial input function, 1060, 1068–1070, 1075–1078
Arterial sampling, 1069, 1073
Artificial diamond, 521–529
Artificial neural network, 1052
ASDEX Upgrade, 798, 802–804
Askaryan effect, 587, 614, 615
Astronomy with a Neutrino Telescope and Abyss environmental RESearch (ANTARES experiment), 467
Astroparticle physics, 314, 315, 324, 327
  –detector, 466–468
Astrophysical observations, 564
ATIC, 573, 583
ATLAS, 475–483, 485–487
Atmospheric monitoring, 615, 623
Atmospheric muon detectors, 489–492
Atomic bremsstrahlung (AB), 839, 840
Attenuation, 1047, 1056, 1060, 1109–1113, 1119
  –coefficients, 941, 943, 953–958, 1100, 1112
  –correction, 953–957, 965
  –of X-rays, 838, 839, 841, 848
Auger electrons, 210, 940
Auger engineering radio array (AERA), 623
Auger muons and infill for the ground array (AMIGA), 623
Automatic exposure control (AEC), 906, 907
Avalanche photodiodes (APDs), 304, 308, 309, 943, 1129, 1130, 1138
Average ionization density, 523
Azimuth angle, 292

B

BaBar, 465, 479, 485, 486
Backfill, 674
Background, 413, 417, 418, 422–426, 428, 429, 431, 442
  –combinatorial, 91
  –radiation, 285, 286
  –suppression, 126
Backprojection, 981–989, 991, 992, 997
Backscatter peak, 419, 420
Backsplash, 569, 570
Balloon altitudes, 560
Balloon-borne experiment with a superconducting spectrometer (BESS), 491, 492, 573, 579
Bambynek, W., 835
Band gap, 380–385, 388, 391, 392, 396, 397, 399, 402, 404, 408, 432–434, 440, 443, 444
Bands, 432, 433, 439, 440, 444
Barrel region, 285, 293
Base materials, 902, 904
Basis function method, 1079
Bayes factor, 114, 117–118
Bayesian
  –ideal observer, 1089, 1093
  –intervals, 118–119
  –statistics, 105, 112, 116, 118, 1001, 1087
    –objective, 113
Bayes' theorem, 105, 106, 112, 113, 116, 117, 119
Beam-conditions monitor (BCM), 520, 521, 525, 526
Beam
  –delivery, 1187–1189, 1191, 1204
  –hardening artifacts, 893, 899, 903, 904, 909
  –intensities, 843, 844
  –kaon, 214, 217
  –muon, 214
  –pion, 214
  –range, 1155, 1157, 1159, 1163, 1168, 1169, 1172
  –spot, 843, 844
Becquerel (Bq), 189, 203, 204, 228, 232
Becquerel, H. A., 203
BED. *See* Biological effective dose (BED)
Beer–Lambert law, 885, 886
Belle, 479, 485
  –detector, 278
  –experiment, 464
  –II experiment, 466
Bentonites, 673–677
BESS. *See* Balloon-borne experiment with a superconducting spectrometer (BESS)
$Beta^+$ activation, 1169
Beta decay, 190, 193, 209, 1107–1109
Beta distribution, 107
Bethe–Bloch formula, 9, 523
BGO. *See* Bismuth germanate (BGO)
Bilinear scaling method, 956
Binding
  –energy, 940, 941
  –potential, 1071, 1074
Binning, 1118
Binomial data, 104, 122–123
Binomial distribution, 104, 107, 123
Biological effective dose (BED), 1199–1201
Biological imaging, 1157–1158
Biomonitoring, 697
BIPM. *See* International Bureau of Weights and Measures (BIPM)
Bipolar transistors, 61, 63, 64
Bismuth germanate (BGO), 430, 433, 435, 536–540, 542, 544, 545, 551, 583, 805
BLANCA, 609
Blood flow, 1066–1069
BMD. *See* Bone mineral density (BMD)
BNCT. *See* Boron neutron capture therapy (BNCT)
Bolometers, 801–802, 808
Bone architecture, 907

Bone mineral density (BMD), 902, 904
Bonner sphere spectrometer (BSS), 761, 777–779
Border security, technology, 634–651
Boron-lined proportional counters, 770
Boron neutron capture therapy (BNCT), 782
Bragg
–peak, 129, 210, 1154, 1155, 1163, 1164, 1169, 1173
–scattering, 447
Branching ratio, 84
Breast cancer
–positron emission mammography (PEM), 1146
–single photon emission mammography (SPEM), 1146
Bremsstrahlung, 4, 11, 13, 15, 16, 18, 475, 478, 800, 801, 1109, 1119, 1120
BSS. *See* Bonner sphere spectrometer (BSS)
Bulk Micromegas, 249, 251, 253, 254, 256, 257

C

Cadmium sulfide (CdS), 529
Cadmium telluride (CdTe), 440, 443–444, 520, 529–530
Cadmium zinc telluride (CdZnTe), 529–530, 532
Calculated attenuation correction, 955
CALET, 584
Calibration, 89, 93, 97
Calorimeters, 126–128, 135, 562, 568–570, 580, 581
–energy flow, 500, 502
–hadronic jets, 500
–noble-liquid calorimeters, 502, 503
–segmentation, 127
–water Cerenkov calorimeters, 503
Calorimetric energy, 606
Cancer
–hadron therapy, 214
–leukemia, 233
–thyroid gland, 207, 233
Capacitive matching, 63, 64
Cardiac gating, 1022, 1061
CASA-MIA, 607
Case of direct (DC) acceleration, 141, 145, 146
CBF. *See* Cerebral blood flow (CBF)
CBV. *See* Cerebral blood volume (CBV)
CCD. *See* Charge-coupled device (CCD)
CDF. *See* Cumulative distribution function (CDF)
CdS. *See* Cadmium sulfide (CdS)
CdTe. *See* Cadmium telluride (CdTe)
CdZnTe. *See* Cadmium zinc telluride (CdZnTe)
Cell cultures for RBE determination, 1199
Cell proliferation, 1157
Center-of-gravity (COG) method, 246
Centroid, 413, 415, 416, 420–425, 449
Cerebral blood flow (CBF), 901
Cerebral blood volume (CBV), 901
Chadwick, J., 835
Channel
–number, 413, 414, 416, 420–422, 425, 449
–width, 413, 416, 417
Channeling, 4, 9–11
Channeltrons, 803
Characteristic X-rays, 710, 834, 836, 837, 839, 940
Charge
–carrier concentration, 382–383
–collection, 31, 32, 43
–exchange, 804–805
–integration method, 131
–measurement, 43–48
–sensitive amplifier, 44–48, 57
–transfer, 244, 249
Charge-coupled device (CCD), 565, 711, 713, 804
–cameras, 803, 807
Charged-current interaction, 603
Charged-particle detectors
–implanted junction, 405–407
–surface barrier, 405–407
Chemical techniques, 363
Cherenkov, 126–128, 132–135, 563
–angle, 455, 456, 460, 465, 597, 608
–calorimeters, 504, 511
–camera, 465
–counter, 454–469
–imaging, 464–466
–threshold, 133, 464
–detectors, non-imaging atmospheric, 608
–energy threshold, 608
–lateral distribution, 609
–light, 607–612, 614, 618, 620, 705
–cone, 608
–neutrino telescope, 467
–radiation, 454, 457, 458, 461, 464, 466, 468, 475
–polarization, 614
–radio. *See* Radio Cherenkov radiation
–radiator, 465, 466
–ring radius, 133, 134
–telescope
–air, 466, 468, 469
–ice, 467
–stereoscopic, 466, 468
–water, 457, 469
–threshold, 458, 460, 464
–timing counter, 459
–tracking calorimeter, 461
–yield, 459
Cherenkov, P., 454
Cherenkov telescope array (CTA), 624
Chi-square, 246, 413
–distribution, 107, 110, 116, 123
–filtered, 289, 290
–smoothed, 282, 283, 289–291
Chromaticity, 144
CIEMAT/NIST method, 196

Circuits
–detector equivalent, 48–52
–digital, 57, 66
–equivalent, 60
Classification tasks, 1085–1086
Clay minerals, 670
CLEO-III experiment, 465
CLIC. *See* Compact linear collider (CLIC)
Clinical dose prescription, 1199
Clinical target volume (CTV), 1156
Clocked integration, 78
Cluster(ing), 88, 90, 91, 99, 100, 1052, 1053
–of double-strand breaks, 1182, 1197
Cluster counting method, 131
CMOS, 68, 69
CMS (detector), 94, 276, 476, 477, 479, 482–489, 493
CO_2 capture and sequestration (CCS), 677
CODALEMA, 615, 616
Coded aperture, 1133
Coding error, 1129
Coefficient of variation, 1046
Coincidence
–circuit, 946–948
–counting, 4π β–γ, 193–195
–detection, 946–947, 949, 951, 953, 955, 963
Collider
–beta function, 144
–circular, 275–277
–computing system, 99
–convolution, 98
–hadron, 87, 90, 92–98, 100, 316, 320, 326
–LEP, 87, 93
–linear, 146, 152, 155, 157
–proton–antiproton, 96, 97
–Tevatron, 96
Collimator, 562–564, 919–929, 931, 932, 1133, 1135–1137, 1143–1145
Collisional quenching, 610
Color center. *See* Absorption color center
Colsher filter, 989
Compact linear collider (CLIC), 146, 155, 157
Compartmental methods, 1057–1059, 1065
Compartment model, 1066–1072, 1075, 1078, 1079
Composite kernel, 413, 417
Compound semiconductor detectors, 443–444
Compton
–absorption, 211, 212
–continuum, 419, 430
–edge, 419
–effect, 14–15, 211, 212, 216, 705
–interaction, 940–941, 957
–scatter(ing), 211, 212, 220, 840, 1109, 1110, 1113–1114
–suppression, 430–431
Compton Gamma-Ray Observatory (CGRO), 569
Computed tomography (CT), 975, 976, 979–981, 993, 1004, 1044, 1046, 1047, 1052, 1054, 1057, 1140–1143
Computed tomography dose index (CTDI), 905, 906
Concentration, 413, 414, 418, 420, 422, 428–431, 444, 446, 449
Conduction bands, 433, 439, 522, 523
Conductive glue, 526, 532
Cone beam, 1160, 1162–1164, 1167, 1168
–artifacts, 909
–CT, 908
–geometry, 978, 979, 991, 1140, 1141
CONEX, 597, 613
Confidence intervals, 119–123
Conform beam delivery, 1187
Conformity of the delivered dose, 1186
Contact, electric, 523, 526, 528, 529
Continuous channel number, 413, 416, 420
Continuum, 413, 417–419, 430, 447
Contrast agents, 215, 900, 901
Conversion electrons, 210
Convolution, 1009, 1012
Convolution-forced detection, 1119
Coronal mass ejections (CME), 573
Coronary angiography, 215
Correlation coefficient, 107, 121, 1021, 1023, 1024
CORSIKA, 597, 598, 602, 606, 618
Cosmic muons, 487, 489, 490, 493
Cosmic Ray Electron Synchrotron Telescope (CREST), 589
Cosmic rays, 217–218, 223, 229, 230, 466, 468, 571
–future experiments, 623–625
–measurements, 487–492
–nuclei, 586
–open problems, 596, 623–625
COSMOS, 597
Coster–Kronig probabilities, 837
Coulomb scattering, 795
Counter, 73, 74
Counting statistics, 690, 692, 696
Counting time, 413, 417, 418, 422, 425
Covariance, 106, 107, 109, 111, 112, 120
–matrix, 281–283, 289, 290, 293
CR 39, 1192, 1194
CREAM, 584
Critical angle, 564
Critical energy, 11–12, 16, 475, 599
Cross section, 836–840, 941, 954
–cut, 94
–differential, 96, 97
–inclusive jet, 96, 97
Cryogenic
–calorimeters, 512–513
–detectors, 447
–spectrometers, 446
Crystal calorimeters, 512

CsI, 536, 537
–photosensitive material, 299, 300, 305, 306
CsI(Na), 563
CsI(Tl), 536–538, 540–542, 549–551, 553, 569, 570
CT. *See* Computed tomography (CT)
CTA. *See* Cherenkov telescope array (CTA)
CTDI. *See* Computed Tomography Dose Index (CTDI)
CT projection, 885–887, 889–899, 904, 908, 910, 912
CT-range calibration curves, 1157
CTV. *See* Clinical target volume (CTV)
Cultural heritage
–science, 814, 819, 828, 830
–studies, 814, 828
Cumulative distribution function (CDF), 106, 120, 1100–1105
Cup-factor, 692
Cupping artifact, 893
Curie (Ci), 204
Curie, I., 203
Curie, M., 228, 454
Curie, P., 228, 454
Cyclotron, 1180, 1189
CZT (cadmium zinc telluride), 529–530, 532

D

D0 detector, 308, 479–483
Daly detector, 694, 696, 698, 699
Daly, N.R., 694, 698
Dark currents, erratic, 520, 524, 525
Dark noise, 693
Data
–abstraction, 85, 86, 89, 90, 100
–access, 99
–analysis, 84–100, 850
–detector, 84, 85
–flow, 85
–offline, 85
–online, 85
–raw, 85–92, 99
–reduction, 85, 86, 89, 100
–storage, 85
–workflow, 92
Data acquisition (DAQ), 85
DC. *See* Case of direct (DC) acceleration
Dead time, 693, 695, 698, 1097, 1109, 1117
Decay constant, 189
Decay-level diagram, 209
Decay time, 695
dE/dx. *See* Specific energy loss dE/dx
dE/dx-E technique, 575
Deep color center, 546
Defined solid angle (DSA) counting, 196
Degrees of freedom, 413, 426, 429
Delay, 1077
Delayed coincidence window technique, 948
Deleptonization, 217
DELPHI experiment, 464
–particle identification, 284
Delta
–electron, 4–5
–rays, 210
Densely ionizing radiation, 1192, 1196, 1198
Density
–of air, 596, 599, 600, 608
–effect, 6, 7
–resolution, 1162–1164, 1166
–variations, 1184, 1193
Depth of interaction (DOI), 949, 967, 1129, 1130, 1132
Depth of shower maximum, 599, 601, 603, 606, 607, 612, 613, 615, 616, 620, 623
Detection limits, 838, 840, 842
Detectors, 354, 360, 366–370, 372, 373
–alignment, 279
–central tracking, 284
–crystal, 1115, 1116
–design, 286, 287
–forward, 284
–integration, 274–279
–micromegas, 269, 270
–optimization, 268, 284–287
–performance, 268, 284–287
–resolution, 413, 416, 419, 424, 425, 442, 444
–response, 413, 414, 417–420, 429, 430
–function, 413, 420, 429, 430
–types
–direct, 28, 29
–indirect, 28
Deterministic annealing filter (DAF), 282, 283, 291
Diamond
–chemical vapor deposition, 521–523
–sensor, 521–529
Difference between protons and heavy ions, 1184
Different therapy modalities, 1180
Diffusion, 242–244, 246, 247, 251, 528
Digital, 56, 57, 66–80
Digital circuits, 57, 66, 70
Digital signal processing, 56–59, 74, 76–80
Digitization (DIGI), 54, 71, 73, 75, 79, 80, 99
Digitization noise, 79
Dipole, 476, 484, 485
Dirac delta function, 417
DIRC detector, 128, 465
Direct Fourier reconstruction, 982
Direction resolution, 284
Direct ion storage (DIS), 772–773
DIS. *See* Direct ion storage (DIS)
Discrete channel number, 413, 416, 449
Dispersion, 144, 691, 1077
Displacement energy, 523
DLP. *See* Dose length product (DLP)

DNA as main target, 1197
DOI. *See* Depth of interaction (DOI)
Dopant, 433, 434, 438, 443
–impurities, 383–384
Dose, 885, 894, 905–907, 911, 912, 1154–1159, 1161–1163, 1166–1172
–escalation, 1155, 1169
–painting, 1157–1158
–rate dependent, 536–538, 546, 553
–reconstruction from PET data, 1172
Dose-guided radiotherapy, 1166–1172
Dose length product (DLP), 906
Dosimeter
–albedo, 223, 224, 226
–film badge, 223, 226
–finger-ring dosimeter, 224
–neutron, 223, 224, 226
–pen-type pocket, 223, 224
–personal dosimetry, 205, 223–227
–track-etch dosimeter, 223
Drift chamber, 240, 245, 246, 269, 275
Drift tube chambers (DT), 478, 481–483, 488, 493
Drift velocity, 242–244, 246, 260
DT. *See* Drift tube chambers (DT)
D–T reactions, 794, 795, 806
Dual-energy
–CT, 902, 904
–technique, 215
Dual-layer scintillators, 1144
Dual-readout calorimetry, 511
Dynamic
–CT, 900
–imaging, 1056, 1060
–range, 687, 692, 694
–studies, 1008, 1009, 1015, 1022, 1025
Dynode, 413, 436, 437, 943
–voltage, 35–36

E

EAS. *See* Extensive air showers
EAS-TOP, 607, 609, 620
EDS. *See* Energy dispersion spectrometry (EDS)
Effective dose, 906
EGRET, 569
Electromagnetic calorimeters, 503–506
–electromagnetic shower containment, 504
Electromagnetic showers, 597
Electronegative gas, 243
Electron–hole pairs, 440, 444
Electronic noise, cross coupling, 51
Electronic portal imaging, 1168
Electronics
–baseline, 692, 693, 699
–bubble chamber, 259, 261
–front-end, 85
Electron micro-probe analysis (EMPA), 551
Electron-nuclear double resonance (ENDOR), 551
Electron paramagnetic resonance (EPR), 551
Electron–photon cascades, 16–19
Electron, 1181, 1182, 1184, 1191, 1192, 1197
–attachment, 243, 244
–avalanche, 244, 249, 250, 258
–capture, 190, 195, 209, 210
–density, 793, 798
–mobility, 523, 530
–multipliers, 708
–temperatures, 793, 798, 802, 803
–volt, 141
Electron spin resonance (ESR), 551
Electrostatic analyzer (ESA), 690
Elementary particle physics, 267, 474
Elongation rate theorem, 601
Emission, 538, 541, 545, 546
EMPA. *See* Electron micro-probe analysis (EMPA)
Emulsions, 474
ENC. *See* Equivalent noise charge (ENC)
ENDOR. *See* Electron-nuclear double resonance (ENDOR)
Energy
–center of mass, 150
–confinement time, 795
–deposit, 87–91
–detector, 1163, 1164
–electron, 88
–flow, 91
–loss, 475, 480, 489, 491
–measurement, 54
–missing transverse, 91
–resolution, 26, 28, 30, 31, 36, 416, 419, 424, 425, 435, 437–442, 444–446, 449, 500–509, 511, 516, 536, 549, 942, 943, 957
–of jet, 287
Energy dispersion spectrometry (EDS), 551, 553, 554
Energy-dispersive spectroscopy (EDS), 431, 432, 448
Energy-weighted centroid, 1116
EPID-based 3D in-vivo dosimetry, 1167
Epithermal neutrons, 736, 742–744
E/p method, 499
EPOS, 606
Epotecny E205, 526
Equivalent noise charge (ENC), 40, 60–63
EROC. *See* Estimation receiver operating characteristic (EROC)
Erratic dark currents, 520, 524, 525
Error detector resolution, 279
ESR. *See* Electron paramagnetic resonance (EPR)
Estimation receiver operating characteristic (EROC), 1088, 1092
Estimation tasks, 1086–1089, 1091, 1092

Estimator
–bias, 108, 109
–consistency, 108
–efficiency, 108
–mean, 108–109
–median, 108–109, 113
–variance, 108–109, 112
Event
–builder, 85
–counting, 94–96
–display, 87, 92, 93
–selection, 93–95, 100
Event-based motion correction, 1036, 1038, 1039, 1041
Event reconstruction, pattern recognition (PR), 279–284
Everhart–Thornley (ET) setup, 709
EWLT. *See* Photo-luminescence weighted longitudinal transmittance (EWLT)
Expectation maximization–maximum likelihood algorithm (EM-ML), 975
Expectation value, 106, 108, 110, 116
Experience of neutron therapy, 1199
Experimental tumors in animals, 1201
Exponential distribution, 107, 1100, 1101, 1104, 1105, 1107, 1112
Extensive air showers, 594, 596–603, 614, 623
Extraction fraction, 1068, 1069

F

Factor analysis, 1059
Fan-beam (geometry), 896–897, 978, 979, 1160
Fano factor, 30, 251
Faraday, 697–699
Faraday cup, 689, 691–693, 699, 700
FBP. *See* Filtered backprojection (FBP)
Feldkamp, 899
Feldkamp, Davis and Kress algorithm (FDK), 906, 991
Fermi function, 291
Fermi-LAT, 573
Feynman diagram, 93
Fictitious interaction tracking, 1119
Field programmable gate or logic arrays (FPGAs), 70, 77
Film efficiency, 1192
Filter(ing), 56, 58, 76–80, 1049, 1050
Filtered backprojection (FBP), 889–891, 893, 897, 899, 908, 924, 961–962, 975, 983–984, 986–992
Filter formulae, 289, 290
Finite impulse response (FIR) filter, 77
Fission chambers, 769, 783, 806
Fission track, 699
Fixed target accelerator, 149–150
Fixed target experiments, 126, 131, 135
–central region, 285
–forward and backward regions, 285
–magnet spectrometer, 275
–sensitivity, 275
Flash ADC, 72, 74
Flip-flop, 66, 67, 70
Flow-through reaction chamber, 674
Fluctuations
–induced charge, 37
–of ionization losses, 7–8
–scintillator detector, 27, 29
–signal, 26, 29–31
Fluence (response), 766, 767, 773, 774, 778, 782, 785
FLUKA, 606
Fluorescence, 1109, 1115
–light spectrum, 611
–telescopes, 610–614, 622–624
–yields, 610, 623, 835
Flux
–muon, 217, 230
–neutrino, 215
–photon, 215
Fly's Eye, 612
FODO, 142–144, 148
Forced detection, 1119
FORE. *See* Fourier rebinning algorithm (FORE)
Forward projection, 962, 963
Fourier analysis, 1051
Fourier rebinning algorithm (FORE), 990
Fourier slice theorem, 889, 891, 980–982, 988
FPGA. *See* Field programmable gate or logic arrays (FPGAs)
Fractionated treatment, 1155, 1159
Frank, I., 454
Frank–Tamm equation, 455
Free-electron lasers (FELs), 162, 163, 176, 178, 182–184
Free paths to travel, 1112
Full-energy peak, 419–422, 430, 444, 449
Full width at half maximum (FWHM), 413, 420, 425, 435, 444–446, 1048
Functional and molecular imaging, 1157
Functions, 56, 57, 66, 67, 70, 73, 76, 77, 79–81
Fusion
–monitoring, 761, 783
–plasmas, 793, 798–799, 801, 802, 806
–reactions, 794–797, 799, 806, 808
–reactors, 797, 798, 806
FWHM. *See* Full width at half maximum (FWHM)

G

Gaisser–Hillas function, 612
Galactic cosmic ray (GCR), 585

Gamma
–camera, 861, 864, 918–923
–counting, 4π, 193
–decay, 215, 1107, 1108, 1118
–rays, 413–415, 420, 422, 423, 430–449, 561, 568, 569
Gamma distribution, 107
Gamma-Ray Bursts (GRBs), 567, 569
Gamma-ray detector, 842
G-APD (Geiger-mode APD), 308–310
Gas amplification, 244, 261
Gas detectors, 474, 475, 481, 484, 486
–secondary ionization, 268, 269
–spatial resolution, 269, 284
Gas electron multiplier (GEM), 241, 249–260, 269, 270
Gaseous photomultipliers, 251, 254
Gas fusion, 550
Gated studies, 1015–1016, 1025, 1041
Gating, 1161, 1169
Gaussian, 97, 98, 413, 415, 416, 423, 424, 438, 444
–distribution, 104, 107, 109–111, 120–122
–measurements, 116, 120–122
–standard deviation, 416
Gaussian sum filter (GSF), 282, 283
GCR. *See* Galactic cosmic ray (GCR)
GDMS. *See* Glow discharge mass spectroscopy (GDMS)
GEANT4 package, 287
Geiger counters, 712
Geiger–Müller counter, 240
–gas amplification, 219, 220
–mean free path, 219
Genetic mutations, 233
GENFIT toolkit, 282
Geochronology, 695, 696
Geomagnetic cutoff, 578
Geometric primitives, 1106, 1107, 1110
Geoscientific applications, 669–681
Geo-synchrotron effect, 614
Germanium detector, 197, 440–442
Gibbs distribution, 1002
Glow curve, 531
Glow discharge mass spectroscopy (GDMS), 549, 551
Gluckstern formula, 285, 477
Glucose metabolic rate, 1055, 1058
Golden channel, 478, 479
Graded-depth multilayer coatings, 566
Graded-*Z* shield, 567
Grain, 521, 522, 525, 528
Grangeat's formula, 992
Grazing incidence, 564–566
GRB. *See* Gamma-Ray Bursts (GRBs)
Greisen parametrization, 599, 600
Gross tumor volume (GTV), 1156
Group velocity, 457
GTV. *See* Gross tumor volume (GTV)

H

Hadron blind detector (HBD), 254, 255
Hadron calorimeters
–compensation, 511
–energy deposit profile, 507
–hadronic response function, 508, 509
–hadronic shower development, 507, 508
–hadronic signal linearity, 508
–invisible energy, 507, 509
–nuclear interaction length, 507
Hadron collider, soft QCD background, 276
Hadronic energy resolution, 507–509, 511
Hadronic interactions, 599–601, 603, 606, 612, 618, 623, 624
Hadrons, 536, 538, 547
–jet, 287
–shower, 20, 21, 597, 600–601, 603, 609, 618, 620
–containment, 508
Hahn, O., 203, 228
Half-life, 188
Haverah Park, 605, 607
HAWC. *See* High Altitude Water Cherenkov (HAWC)
HEAT. *See* High Elevation Auger Telescopes (HEAT)
Heavier beams, 1188
Heavier particles, 1197
Heaviside function, 291
Heavy ions, difference to protons, 1184
Heavy-ion therapy, 214
Heitler–Matthews model, 600
Helical CT, 897–898
Helix coordinate system
–cartesian coordinates, 292
–cylindric coordinates, 292
–helix radius, 293
–helix track
–equations, 292
–model, 292
–parameter convention, 293
–perigee point, 293
–spherical polar coordinates, 292
HERA, 151
HERMES experiment, 465
Hermeticity, 502, 503
HERWIG, 603
H.E.S.S. experiment. *See* High energy stereoscopic system (H.E.S.S.) experiment
HFS1, 526
HgCdTe, 529

High Altitude Water Cherenkov (HAWC), 624
High Elevation Auger Telescopes (HEAT), 623
High energy stereoscopic system (H.E.S.S.) experiment, 468, 573, 609, 620–621
High-purity
–Ge (HPGe) detectors, 399–403, 408, 441
–Si detectors, 443
High-resolution imaging, 1126–1147
High-resolution mirror assembly (HRMA), 565
High-sensitivity instrumentation, 1127, 1129
Hillas parameters, 609, 620
HiRes Fly's Eye, 612
Histogram(ming), 1049, 1050, 1056, 1117–1119
–equalization, 1049, 1050
Holes, 413, 439, 440, 445
Hotelling observer, 1090–1092
Hounsfield units (HU), 887, 1047, 1157, 1159, 1163, 1164
HPGe. *See* High-purity Ge (HPGe) detectors
HU. *See* Hounsfield units (HU)
Huber prior, 1002
Humidity, 674
–chamber, 673
Hybrid detectors, 604
Hybrid photodetector (HPD), 299, 303–305
Hybrid pixel detector, bump bonding, 271
Hydration, 674
Hypothesis
–composite, 114
–simple, 114
–testing, 94

I

IACT. *See* Imaging air (atmospheric) Cherenkov telescopes (IACTs)
Ice Cherenkov telescope, 468
IceCube, 607
IceTop, 607
ID-TIMS. *See* Isotope dilution TIMS (ID-TIMS)
ILC. *See* International linear collider (ILC)
ILD detector, 277, 286, 287
Illicit trafficking nuclear materials, 761, 782–783
Ill-posedness, 984
Image
–blurring, 1011, 1013, 1027
–enhancement, 1049
–reconstruction, 937, 961–964, 1085, 1087
–registration, 1014–1017, 1020–1021
–segmentation, 956, 1050–1054
Image-guided radiotherapy (IGRT), 1158–1168, 1172–1173
Imaging air (atmospheric) Cherenkov telescopes (IACTs), 468, 469, 607–610, 620
Imaging Cherenkov counter, 464–466
Immobilization, 1155
IMP-1, 575
Impact parameter
–resolution, 286
–transverse, 286, 293
Importance sampling, 1119
IMPT. *See* Intensity modulated particle therapy (IMPT)
Impurity, 434
–concentrations, 802
–influx, 802
IMRT. *See* Intensity modulated radiotherapy (IMRT)
Induced charge, 31–33, 37, 43, 413, 444
Inductively coupled plasma, 688
Infrared fluorescence, 623
InGrid, 249, 259, 260
Initial-state radiation (ISR), 283
Inorganic mass spectrometry, 686, 687
Inorganic scintillators, 356, 357, 365, 366, 432–436
In-room imaging, 1159, 1161, 1169
Instability, 538
Intensity modulated particle therapy (IMPT), 1180, 1188, 1189, 1204
Intensity modulated radiotherapy (IMRT), 1154, 1155, 1168, 1172
Interaction length
–in air, 596, 600
–nuclear, 20
–hadronic calorimeters, 507
Interactions
–of DSB, 1201
–strong, 90, 96
Inter-fractional changes, 1159
Inter-laboratory comparison, 697
Internal conversion, 190–191, 195
Internal target volume (ITV), 1156
International Bureau of Weights and Measures (BIPM), 189, 197, 198
International linear collider (ILC), 146, 157
International space station (ISS), 578
Interstellar medium (ISM), 571–573
Intra-fractional changes, 1159
Intrinsic carrier concentration, 384, 399
Intrinsic charge-carrier density, 523
Invariant mass, 126
Inverse dose profile, 1180
Inverse problem, 418, 422, 431
Inversion, 415, 418, 427, 1100–1102, 1104
Inverter, 67, 68, 70
In-vivo verification, 1159, 1161, 1164, 1166–1168, 1171, 1173
Ion
–beam, 686–694, 696, 699
–counting, 689, 692, 694–696, 698, 699
–feedback, 250, 253
–mobility, 244
–radiography, 1163–1166

–therapy, 1154, 1155, 1157, 1159, 1161–1166, 1170, 1172
–tomography, 1165–1166
Ionization, 705, 707, 712, 799, 802, 803, 805
–chambers, 188, 196–197, 218, 1189–1195, 1201
–equivalent circuit, 49
–induced charge, 43
–losses, 5–8, 11–13
–primary, 241, 242, 244, 249–251, 254, 260
–sources, 687, 695
–stages, 802, 803
Ionizing radiation, 965
Irradiation protocols, 1195
Irreparable damage, 1197
ISEE-3, 575
ISM. *See* Interstellar medium (ISM)
Isotope. *See* Radioisotopes
Isotope dilution TIMS (ID-TIMS), 688, 695–698
Isotope ratio mass spectrometry, 686–699
ISS. *See* International space station (ISS)
ITER, 807–809
Iterative reconstruction, 961–963

J

Jacobian matrix, 288, 290
Japanese Experiment Module Exposed Facility (JEM-EF), 584
Japanese Exposure Module Extreme Universe Space Observatory (JEM-EUSO), 590, 624
Jeffrey's rule, 113
JET, 805–807
Jet
–algorithm, 87, 90, 96
–analysis, 96, 98
–baryon, 90
–bin, 96, 97
–efficiency, 91, 97
–energy scale, 97
–production, 96–98
–reconstruction, 90, 97
Jitter, 65

K

Kalman filter, 282, 283, 287–291
KASCADE, 596, 604, 607, 615–618, 620
KASCADE-Grande, 618, 625
K-edge subtraction technique, 215
Kerma, 205
Kinetic modeling, 1056–1058
Klein–Nishina distribution, 1114
Klein–Nishina equation, 941
Klystron, 213, 214
Kolmogorov axioms, 104
Kullback–Leibler divergence, 996
kV sources, 1160

L

$LaBr_3$ (Ce), 435
Lagrangian, 84
Lagrangian multipliers, 283
Landau distribution, 8, 130, 131
Landau function, 242
Landau–Pomeranchuk–Migdal (LPM) effect, 600
Large electron–positron collider (LEP), 146, 149, 150, 152, 155, 214, 487
Large Hadron Collider (LHC), 92, 149, 150, 152, 155, 156, 520, 521, 525–529, 532
Laser ablation (LA), 688, 692, 693, 695–697
Latch, 66, 67
Lateral distribution function (LDF), 604
Lateral scattering, 1182–1184
Lattice
–alternating gradient, 143
–magnetic, 142, 143, 148
–magnets, 142–144
–phase advance, 143
Law of total probability, 105
LCG. *See* Linear congruential generator (LCG)
LDT fast simulation tool, 286
Least-squares method, 413, 422, 424–431
LECO, 550
LEM. *See* Local effect model (LEM)
LEP. *See* Large electron–positron collider (LEP)
Lepton collider, 275–277
–beamstrahlung background, 277
LET. *See* Linear energy transfer (LET)
LHAASO, 624
LHC. *See* Large Hadron Collider (LHC)
LHCb, 476, 479, 480, 486, 487
–experiment, 465
Library least squares (LLS), 422, 429–431
Li-drifted silicon (Si(Li)), 576
LiF. *See* Lithium fluoride (LiF)
LiF:Mg,Cu,P, 531, 532
Light
–attenuation length, 536, 543, 546
–collectors, 590
–response uniformity, 536, 538, 547–549
Lighter particles, 1197
Likelihood, 995–997
Limits, 118–123
Linear
–attenuation coefficient, 885–887, 894, 910
–collider, 277
–electro-optic coefficient, 529

–estimator, 282
–expansion, 283
Linear accelerator (LINAC), 214, 215
Linear congruential generator (LCG), 1099
Linear energy transfer (LET), 771, 775, 1154, 1196, 1197, 1199
–energy loss, 210
Linearity, 692, 693, 698
Linearized method, 1079
Line integral model, 976–978, 981, 992, 994
Line of response (LOR), 978, 990, 1032–1039
Line radiation, 800–805
Liquid scintillation counting (LSC), 195, 196
Lithium fluoride (LiF), 531, 532
Local effect model (LEM), 1192, 1193, 1199–1202
Localization of tumor target volume, 1155
Local overkill effect, 1192
Local RBE values, 1198, 1199
Logic, 66–70
Logic symbols, 67
Long-object problem, 992
LOPES, 615, 618, 625
LOR. *See* Line of response (LOR)
Lorentz factor, 141, 155
Lorentz force, 87, 140, 155
Low frequency, 62
LPM effect. *See* Landau–Pomeranchuk–Migdal (LPM) effect
LSO, 537, 539–543, 547, 548
Luminosity, 94, 97, 151–153, 275
–factories, 153
LYSO, 536–543, 547–549

M

Mach cone, 454
MACRO. *See* Monopole, Astrophysics, Cosmic Ray Observatory (MACRO)
MAF. *See* Multiple acquisition framing (MAF)
MAGIC, 609, 624
Magnet, 475–479, 484, 485, 487, 490, 491
Magnetic
–bremsstrahlung, 600
–field, 87, 88, 474–476, 484, 487, 490, 491, 494
–lines, 795, 796
–flux density, 292
–pair production, 600
–sector field (SF), 690
–spectrometers, 126, 128, 133, 475–478
Main quantum numbers, 836, 837
–fine structure of the electron shells, 836
Manhattan project, 1098
MAP. *See* Maximum a posteriori reconstruction algorithm (MAP)
Marginal likelihood, 118
Markov field theory, 1054
Mass
–absorption coefficient, 211, 212
–attenuation coefficient, 13, 16, 211, 212, 886, 902, 904
–bias, 688, 691, 693, 699
–fractionation, 688
–resolution, 686, 688, 690
–spectrometry, 686–699, 707
Mathematical morphology, 1052, 1054
Mathieu–Hill equation, 143
Maximum a posteriori reconstruction algorithm (MAP), 1001
Maximum detectable momentum, 478, 490, 492
Maximum energy transfer, 4
Maximum likelihood, expectation maximization (ML-EM), 963, 996
Maximum likelihood, extended, 110
Maxwellian energy distribution, 795, 805
MCPs. *See* Microchannel plates (MCPs)
Mean residence time (MRT), 1057, 1058
Mean transit time (MTT), 901
Mean value, 415, 420
Measured attenuation correction method, 956
Measurement equations, 288, 289
Measurement noise, 288, 289
Medical imaging, 1097–1122
Medipix2 ASIC, 258
Mercuric iodide (HgI_2), 443–444
Mersenne Twister, 1100
Metabolism, 1157, 1158
Metal detectors, 635, 639–640
Metals, 834, 842, 847–852
MicroBulk Micromegas, 249, 251, 254
Microcalorimeters, 446, 447, 449
Microchannel plates (MCPs), 299, 302–303, 437, 531, 575, 803
Micro-CT, 907–912
Micro-hole and strip plate (MHSP), 249, 250
Micro-mesh gaseous structure (Micromegas), 241, 246, 249–260
Micropattern gas detectors (MPGD), 241, 247–261, 269–270
Microporosity, 3D, 679
Micro-strip gas chamber (MSGC), 240, 247, 269
Microwave radiation, 624
Milagro, 618–619, 624
Millepede package, 279
Minimum-ionizing particle, 6
Missing energy, 612
MLEM. *See* Maximum Likelihood, Expectation Maximization (ML-EM)
MLP. *See* Most likely path (MLP)

Model fitting, 422–428, 430
Moderators, nuclear, 721, 722, 725, 730, 733–740, 742–744, 754
Molecular bremsstrahlung, 624
Molecular imaging, 1126, 1127, 1133, 1137
Molière radius, 19
Molière unit, 597, 599
Momentum
–direction, 281, 287
–measurement, 477, 480, 481, 489, 491
–resolution, 476–478, 480, 484
–absolute, 286
–transverse, 285, 286
Monocular observation, 611
Monopole, Astrophysics, Cosmic Ray Observatory (MACRO), 490–492, 494, 607
Monte Carlo simulation, 285, 958, 1097–1100, 1119
Moseley's law, 215, 835
Most likely path (MLP), 1166
Motion
–artifacts, 894, 896, 904, 909
–correction, 1008, 1014–1016, 1018–1040
–sensor, 1014, 1016–1019
–tracking, 1014, 1016, 1018–1020, 1022, 1025–1027, 1029, 1033–1034, 1038, 1041
–types, 1009, 1012, 1020
MRT. *See* Mean residence time (MRT)
MSCT. *See* Multi-slice CT (MSCT)
MTT. *See* Mean transit time (MTT)
Multi-collector arrays, 690–692
Multi-dynamic analytical routines, 692
Multigap resistive plate chambers (MRPC), 128
Multimodal, 1155, 1157, 1158, 1172
Multi-modality imaging, 965–966, 1141, 1143
Multiple acquisition framing (MAF), 1025–1028, 1034
Multiple coincidences, 948
Multiple Coulomb scattering, 281
Multiple ion counting (MIC), 692, 696
Multiple scattering (MS), 8–9, 13, 19, 21, 475–477, 479, 491, 493, 599
Multi-slice CT (MSCT), 898–899
Multivariate, 107, 120
Multi-vertex filter (MVF), 283, 291
Multiwire chambers, 1189, 1190
Multi-wire proportional chamber (MWPC), 240, 244–247, 268, 269, 275, 486, 487, 490
Muon, 126, 127, 133, 135
–radiography, 493–494
–spectrometers, 474–494
–tomography, 256
Mutual information, 1012, 1021, 1023, 1024
Mutual recognition arrangement (MRA), 198
MV sources, 1160

N

NaI(Tl), 419, 434–436, 441, 442, 536, 537, 563, 569
–detector, 188, 193, 197
NaI, 805
NAND gate, 66–68, 70
Nano-CT, 884, 885, 907–911
Natural calorimeters
–extensive air showers, 514
–extra-galactic, 513
–IceCube, 513–515
Natural diamond, 806, 807
NbTi, 578
NE226, 805
Necrotic areas, 1201
Net area, 422, 423, 428
Net rate of influx, 1079
Network switches, 85
Neutral-current interaction, 603
Neutrino-induced showers, 603, 623
Neutrinos, 498, 500, 512–514, 516
–astronomy, 344
–beams, 214, 215
–cross section, 603
–detection, 314, 319, 321, 334, 341, 342
–detector, 314–345
–factories, 215
–physics, 316
–source, 216–217
Neutron-deficient C isotopes, 1184
Neutrons, 704, 705
–activation analysis, 761, 781
–detection
–with gas-filled detectors, 768–770
–mixed, 776, 777, 783
–passive methods, 771–773, 776
–principles, 768
–pulsed, 776–777
–recoil spectroscopy, 780–781
–with scintillators, 770–771
–with semiconductors, 770
–time-of-flight, 779–780
–detectors, 407–408, 716, 761, 772, 773, 776, 778, 779, 782–784, 817, 826
–diffraction, 814, 815, 822–824, 828, 829
–dosimetry, 223
–energy ranges, 762, 763
–fission neutrons, 216
–fission products, 216
–fluence-to-dose-equivalent conversion factors, 767, 785
–flying personnel, 223
–fusion, 216
–generation, 765, 779
–generator, 216

–grazing-incidence diffraction, 820–822
–high energies, 763, 765, 774
–humidity measurement, 784
–imaging, 784, 814, 817, 829
–interactions (coherent neutron scattering), 762–765, 770, 771, 784, 818–827
–lifetime, 217
–moderation, 764–766
–peak flux, 730, 736, 737
–plasma generator, 765
–powder diffractometry, 744–745, 748
–radiotherapy, 781
–rate, 798, 806, 807
–reference radiation fields, 784–785
–reflectometry, 820–822
–scattering, 761, 765, 780, 781, 785, 818–827
–shielding, 763, 766
–sources, 815–817, 825
–spallation neutron source, 216
–spectra, 780
–spectrometry, 779, 783
–spectroscopy, 806–807
–thermal neutron, 762, 763, 766, 769, 773, 778, 779, 782, 784
–tomography, 814
–track-etch detector, 223
–track-etch dosimeter, 223
–transmutation, 216
Nishimura–Kamata–Greisen function (NKG), 600
NKG function. *See* Nishimura–Kamata–Greisen function (NKG)
NMOS, 67–69
Noise
–analysis, 60–64
–cross coupling, 51
–dark, 693
–*vs.* dynamic range, 42–43
–electronic, 26, 37–43, 54, 61, 71, 79, 88
–measurement, 288, 289
–Poisson, 1085, 1090
–process, 288, 289
–readout, 536, 539, 553
–tomography, 1044
Noise equivalent counts (NEC), 948–949
Non-collinearity, 1131, 1132
Non-imaging atmospheric Cherenkov detectors, 608
Nonlinear models, 430
Nonlinear regression, 1071, 1075, 1078, 1079
Nonrigid motion, 1009, 1014, 1019, 1022, 1025
Non-transit dosimetry, 1167
NOR gate, 66
Normal distribution, 1103, 1104, 1116
Normalization, 952–953, 963, 1010, 1034–1036
Nuclear
–binding energy, 794
–decay, 1106–1108
–emulsion, 273, 287
–energy, 748–755
–forensics, 686–699
–fragmentation reactions, 1169, 1172
–fusion, 793–809
–interaction length, 20
–power plant, 225, 227
–track detectors, 1189, 1191–1192
Nuclide, 413, 414, 418, 420–422, 428–431, 449
Nuisance parameter, 120
NuStar, 566
Nyquist criterion, 78

O

Okayama muon telescope, 489, 490
Omega detector, 286, 293
Online blood sampler, 1077
Online monitoring, 1166
Online PET system, 1185
Operating voltage, 693
Operational amplifier, 691
Operational quantities, 767, 773, 774
Optical
–bleaching, 537, 542, 553
–flow, 1014, 1022–1025, 1029
–imaging, 1126
–systems, 1160
Optimal therapy ions, 1199
Optimum distance, 605, 606
Orbiting Wide-angle Light Collectors (OWL), 590
Ordered-subset algorithm for transmission tomography (OSTR), 999
Ordered-Subsets EM algorithm (OSEM), 998–1000
OR gate, exclusive, 66, 67
Organ at risk (OAR), 1156
Organic mass spectrometry, 686, 687
Organ motion, 1155, 1156, 1169
Orlov's condition, 988
OSEM. *See* Ordered-subsets EM algorithm (OSEM)
OSTR. *See* Ordered-subset algorithm for transmission tomography (OSTR)
Outliers, test criteria, 283
Overall treatment quality, 1182–1183
Overkill effect, 1192, 1197
Overlapping peaks, 413, 423–425, 428, 429
Oxide, origin of the radiation damage, 538
Oxygen-15, 1068
Oxygenation, 1155
Oxygen contamination, 540, 549–551, 553
Oxygen vacancies, 545, 551–554

P

Paintings, 834, 842, 844–847
Pair generation. *See* Pair production

Pair production, 15–18, 418, 705, 1110, 1112, 1113, 1115, 1118
PAMELA, 573, 577
PANDA experiment, 466
Parallax error, 1130, 1132, 1138
Parameter estimation
–estimate, 108–110, 120
–least squares, 108, 110–112, 116
–likelihood (ratio), 108, 110, 120
–linear problem, 111
–maximum likelihood, 108–110, 120
–mean-squared error, 108
Parametric imaging, 1058–1060, 1078
Partial-body dose, 206
Partial volume (effect), 891–892, 900, 959–961, 1047–1050, 1053, 1055, 1060
Particle
–charged, 86–88, 90, 91, 96
–decays, 84, 85, 88, 90, 91
–flow, 91
–interaction, 84, 86, 90, 92
–lighter, 1197
–meson, 90
–momentum, 84, 86–88, 90, 91, 95
–neutral, 88, 90, 91
–neutralino, 91
–neutrino, 91
–neutron, 88
–parton, 90
–photons, 88
–physics, 84–86, 88, 96, 98, 100, 316, 320, 326
–track, 86, 87, 90, 91
Particle flow analysis (PFA), 284, 287
Particle identification (PID), 90, 457, 463, 464, 469
Particle induced X-ray emission (PIXE), 551
Particle-tracking software, 1120
Passive
–beam forming techniques, 1186
–degraders, 1189
–detectors, 1191
–mm-wave systems, 637
Passive passenger portals (PPP), 635, 637–638
Patient positioning, 1155, 1159, 1161–1163, 1166, 1168, 1169, 1180
Patient set-up, 1155, 1159, 1160
Patlak analysis, 1058, 1079
Pattern recognition (PR), 1052–1054
–inwards/outwards strategy, 280
PDF. *See* Probability density function (PDF)
Peak centroid, 416, 423
Peak detector, 73, 74
Pearson's chi-square statistic, 116
PEM. *See* Positron emission mammography (PEM)
Penelope, 1120
Perfusion, 1066–1068, 1076, 1077
Perfusion CT (PCT), 884, 900–902
Peripheral QCT (PQCT), 902, 903
Persian Mummy, 662–665
PET. *See* Positron emission tomography (PET)
PET-CT. *See* Positron Emission Tomography–Computed Tomography (PET-CT)
PET-MRI. *See* Positron Emission Tomography–Magnetic Resonance Imaging (PET-MRI)
Phase stability, momentum compaction, 148
Phase velocity, 454, 457
Phononic excitation, 523
Phosphorescence
–afterglow, 536, 537
–bulk effect, 538
Phoswich detectors, 563
Photocathode, 436–438, 440, 943, 944
Photodetectors, 942, 943, 945, 947, 949, 965, 1128–1130, 1137–1139, 1142, 1144, 1147
Photodiodes, 306–308, 435, 437–438
Photoelectric absorption, 418, 1112–1115
Photoelectric effect, 13–14, 17, 705, 710
Photoelectric interaction, 940–942, 957
Photoelectron, 562
Photofraction (PhF), 1134, 1135, 1138, 1139
Photo-lithography, 247, 259, 261
Photoluminescence, 538–540
Photoluminescence weighted longitudinal transmittance (EWLT), 540
Photomultipliers, 29, 35, 36, 42, 563, 582, 708–710, 803, 807
Photomultiplier tube (PMT), 298–303, 310, 432, 434–438, 440, 943–945, 965
Photon-induced showers, 599–601, 609, 618
Photons, 704, 705, 710, 712, 713
–annihilation, 215
–characteristic X rays, 215
–decay-level diagram, 209
–detection efficiency, 222
–feedback, 250
–interaction, 1113, 1116
–in lead, 212
–pair production, 212, 220
–photoelectric effect, 211, 212, 220, 221
–quenching, 243
–sources, 215–216
Photon-tracking simulation, 1105, 1107, 1112, 1115, 1116, 1119, 1121
Photon-tracking software, 1120, 1121
Physical and biological decay, 1169
Pierre Auger
–array, 589
–observatory, 605, 613, 614, 621–623
Piezoelectric transducer, 49

Pileup, 1117, 1118
Pinhole, 1133–1135, 1147
pin-junction devices, 389–390
Pion
–neutral, 88
–tumor treatment, 214
Pitch, 897–899, 906
PIXE. *See* Proton induced X-ray emission (PIXE)
Pixelated volume, 1106
Pixel readout, 257–260
Plane wave Born approximation (PWBA), 836
Planning CT, 1156, 1157, 1159, 1167, 1171
Planning optimization, 1197
Planning target volume (PTV), 1156
Plasma facing components, 798–800, 802
Plasma parameters, 793, 795, 796, 798, 799, 802, 806
Plastic scintillator, 563
PMF. *See* Probability mass function (PMF)
PMOS, 67–69
PMT. *See* Photomultiplier tube (PMT)
pn junction, 380, 386–390, 404, 405, 407
Point of closest approach (PCA), 293
Point spread function (PSF), 1009–1013, 1048, 1050
Poisson
–data, 104, 122–123
–distribution, 1103–1105
–independence, 110
–noise, 1085, 1090
–statistics, 1191
Polar angle, 286
Polarization, 443, 444
–of Cherenkov radiation, 614
Polyatomic interferences, 688, 689, 696–698
Portal imaging, 255
Position-sensitive ionization chamber, 1190
Position-sensitive photomultiplier tubes (PS-PMT), 1128–1130, 1134, 1146
Position-sensitive wire chamber, 1190
Positron, annihilations, 418, 419
Positron decay, 1108, 1115
Positron emission mammography (PEM), 1146
Positron emission tomography (PET), 868–875, 1008–1011, 1013, 1015, 1016, 1019, 1021, 1022, 1024, 1025, 1028, 1029, 1032, 1038, 1040, 1041, 1044, 1046–1048, 1050, 1052–1060, 1066, 1067, 1069, 1070, 1080, 1126–1133, 1137–1146, 1157, 1158, 1169–1172
–annihilation, 937–939
–anti-particle, 938
–of beam-stopping distribution, 1185
–coincidence detection, 937, 938
–image reconstruction, 868
–line of response (LOR), 938
–molecular imaging, 937
–new detectors and technologies for medical imaging systems, 868
–positron, 938
–positron decay, 937, 938
–positronium, 938
–tomographic image reconstruction, 937, 938
–verification of ion beam therapy, 1170
Positron Emission Tomography–Computed Tomography (PET-CT), 965
Positron Emission Tomography–Magnetic Resonance Imaging (PET-MRI), 965–966
Positron range, 1131, 1132
Power dissipation, 56, 63, 68–69, 72, 74, 79
Precision, 417, 421, 1045, 1052
Predictor formulae, 288
Preprocessing, 1049–1051, 1053
Pressure sensors, 1161
Primary ionization, 241, 242, 244, 249–251, 254, 260
Probability
–conditional, 104, 105
–distribution, 106, 107
–limiting frequency, 105
–posterior, 105, 116–118
–prior, 105, 113, 117, 119
–sample space, 104, 105
–significance level, 114
–subjective, 105
Probability density function (PDF), 415, 416, 423, 424, 1100–1105
–conditional, 106
–marginal, 106
–posterior, 109, 112, 113, 117
–prior, 109, 112, 113, 118
Probability mass function (PMF), 1104, 1105
Process noise, 288, 289
Profile likelihood, 120
Prompt coincidences, 948
Prompt gamma, 1169, 1172–1173
Propagation delay, 69–70
Proportional chambers, 486, 487, 490
Proportional counter, 218–220, 223, 561, 803, 804
–4π, 193
Proton, 206, 207, 209, 210, 214–217, 229, 230
–accelerators, 730, 751
–difference to heavy ions, 1184
Proton induced γ-ray emission (PIGE), 841–842
Proton induced X-ray emission (PIXE)
–cultural heritage, 834, 852
–nondestructive, 834, 845, 852
–noninvasive, 834
Pseudo-rapidity, 88, 96
PSF. *See* Point spread function (PSF)
PS-PMT. *See* Position-sensitive photomultiplier tubes (PS-PMT)
PTV. *See* Planning target volume (PTV)
Pulse, 54–62, 64–69, 71–80
–height, 413–420, 425, 449
–analysis, 800

–spectrum, 417–420, 449
–shape, scintillator, 31
Pulsed neutron sources, 726–741
Pulse-height analyzers (PHAs), 946
PWBA. *See* Plane Wave Born Approximation (PWBA)
PWO, equilibrium, 544, 545

Q

QA procedure, 1195
QFEB. *See* Quasi free electron bremsstrahlung (QFEB)
Quadrant sharing, 945
Quadrupole mass analyzer, 690
Quality assurance, 1157, 1159, 1160, 1166
Quality control, 1180, 1189, 1192, 1201, 1204
Quantification, 1066, 1069, 1070
Quantitative analysis, 422–431, 449
Quantitative CT (QCT), 884, 902–903, 911
Quantization, 79
–noise, 79
Quantum chromodynamics (QCD), 96
Quantum efficiency (QE), 546
Quark
–flavor, 94
–gluon, 90
Quasi free electron bremsstrahlung (QFEB), 840

R

Rad (Radiation absorbed dose), 204, 205
Radar reflection, 624
Radial direction, 949
Radiated power, 798, 799
Radiation
–absorbed dose, 204, 205
–accident, 229
–cancer, 210, 233
–contaminations, 218, 220, 225, 228, 229
–cosmic, 203, 206, 231
–damage, 536–554
–damage mechanism, 549
–decay constant, 204
–decontamination, 228
–delayed, 233
–detectors, 218–227
–dose-rate constant, 206–208
–early, 232
–effects, 205, 225, 232–233
–environment, 536
–exposure, 884, 894, 899, 900, 905–907, 912
–fields, 766, 767, 771, 775–777, 784–785
–genetic, 233
–gray, 204, 205
–half-life, 204, 225, 228
–hardness, 521, 523, 525
–hormesis, 233
–incorporations, 218, 228, 229
–length, 9, 11–12, 16, 20, 583, 599
–losses, 11–12
–natural radiation, 218, 230, 231, 233, 234
–officer, 227, 228
–particle, 213–215
–point-like, 206
–protection supervisor, 227
–quality, 1201
–radioactive waste, 218, 228
–radioisotope, 213, 225, 229
–risk, 206, 227, 229, 233
–shielding, 229
–sources, 161–163, 165, 166, 170, 176, 177, 181, 184, 213–218
–terrestrial radiation, 217, 229, 231
–weighting factor, 205, 206
Radio Cherenkov radiation, 614
–lateral distribution, 614
–signal detection, 614–616
Radioactive waste, 669
Radiocarbon, 654, 657, 660–665
Radiofrequency (RF)
–acceleration, 140, 145–146, 154, 157
–cyclotron, 146
–superconducting, 157
–transponders, 1161
Radiogenic isotope, 687
Radioisotopes, 1131, 1133, 1137, 1143
–carbon dating, 220
–photopeak, 213
–primordial isotopes, 230
Radioresistant tumors, 1198
Radiotherapy, 1154–1174
Radon
–inhalation, 227, 230
–plastic detector, 225
Radon transform, 887–889, 891, 893, 908
Raether limit, 248
RAMLA. *See* Row-action maximum-likelihood algorithm (RAMLA)
Ramo's theorem, 33, 37, 524
Random
–coincidences, 1109, 1117, 1118
–error, 1159
–number generator, 1099–1100
–variable, 105–107, 115
–independence, 107
–moment, 106
Range
–dilution, 1163, 1165
–positron, 1131, 1132
–of protons, 837, 839
–straggling, 1182

–telescope, 1163–1165
–uncertainties, 1159, 1163
Rao–Cramér–Frechet bound, 108
Raster-like pattern, 1187
Rate constants, 1068–1071, 1078
Rathgen, F., 834
RBE. *See* Relative biological effectiveness (RBE)
Reactor instrumentation, 783
Readout noise, 536, 539, 553
Real time verification, 1172
Rebinning, 897–899
Receiver operating characteristic (ROC) analysis, 1088–1090, 1092
Receptor
–density, 1066, 1071
–ligand, 1067, 1070, 1077, 1078
Recombination radiation, 800
Reconstruction
–from cone-beam projections, 899, 908
–filter, 889, 893
–physics objects, 85–92, 100
–resolution, 88, 91, 97
–resonance, 87, 91, 95
–shower, 88
Recovery coefficient, 959, 960
Recurrent tumors, 1201
Reference surface, 281, 282, 288
Reference tissue, 1066, 1073–1076, 1079, 1080
Refractive index, 127, 132–134, 454, 457–459, 468, 469, 607, 608
Region growing, 1052, 1053
Regularization, 984, 985, 1000
Relative abundances, 422
Relative biological effectiveness (RBE), 1186, 1195–1204
Relative quantitation, 1046, 1055, 1056
Relativistic rise, 130
Rem counter, 761, 773–775, 777, 782
Repeatability, 1045, 1046, 1054
Reproducibility, 1045–1046, 1054, 1057
Rescanning techniques, 1190
Residuals
–filtered, 289, 290
–smoothed, 289, 290
Resistive charge division, 49–52
Resistive plate chambers (RPC), 484–486
Resistive thick GEM (RETGEM), 241, 249, 252, 254, 256
Resistivity
–depletion region, 406
–depletion width, 405
Resolution
–energy, 26, 28, 30, 31, 36
–oversampling, 79
–time, 65, 66
Respiration belt, 1015, 1016
Respiratory gating, 1028, 1039, 1041
Response functions, 420, 428–430
Response time, 523
Response to dose equivalent, 767
Restricted energy loss, 523
RF. *See* Radio frequency (RF)
RICH detector, 462
Rigid motion, 1009, 1014, 1020, 1025, 1027
Ring artifacts, 893, 896, 910
Ring geometry, 1127, 1128, 1132, 1144
Ring imaging Cherenkov counter (RICH), 132–135
Rise time, 65, 69
Robinson detector, 710
Robustification, non-linear estimator, 282
ROC. *See* Receiver operating characteristic (ROC) analysis
Roentgen, 205, 206
Roentgen equivalent man, 205
Room-temperature-operated devices, 402
Rotating planar detector, 1128
Row-action maximum-likelihood algorithm (RAMLA), 998
RPC. *See* Resistive plate chambers (RPC)
Runge–Kutta–Nyström algorithm, 281
Rutherford, E., 203
Rydberg constant, 215

S

Safeguards monitoring, 698
Safety margins, 1155, 1156, 1159
Safety standards
–effective dose limit, 227
–equivalent dose, 227
Sample preparation (AMS)
–archaeological samples, 658–659
–bones, 658
–sediment, 657–658
Sampling, 76–79
–distance of CT projections, 891
–effect, 1048
–fluctuations, 505, 511
–fraction, 505, 506, 511
Satellites, 560
Saturation effects (quenching), 1192
Saturation value, 1197
Scarab, 850–851
Scatter, 864, 868, 886, 896, 905, 926, 947, 994, 1047, 1113–1114, 1181
–correction, 927, 957–958
Scattered coincidences, 947, 948
Schottky devices, 390–391

Schubweg, 524
Schwarz criterion, 1078
Scintillating fiber detector, 274, 287
Scintillation, 350–373
–counter
–light yield, 221
–photomultiplier, 221
–detectors, 414–416, 419, 434–437, 440, 444, 449, 939, 942–944, 946, 947, 949, 957, 958, 964
–mechanism, 536–539, 547, 553
–photon, 1113, 1115, 1116, 1121
–spectrometers, 432–439
–type detector, 694
Scintillators, 474, 481, 482, 490, 709–711, 803, 807, 809, 1128, 1129, 1131–1136, 1138, 1139, 1144, 1146
–crystal, 357, 364–369, 373
–detector, 28
–pulse shape, 31
Scintimammography, 1146, 1147
SDDs. *See* Silicon drift detectors (SDDs)
SEB. *See* Secondary electron bremsstrahlung (SEB)
Secondary electron bremsstrahlung (SEB), 840
Secondary electron multiplier (SEM), 686, 692–696, 698, 699
Secondary electrons, 692, 694
Secondary ionization mass spectroscopy (SIMS), 550, 553
SELEX experiment, 465
SELEX RICH, 134
Self-absorption, 541
Self-powered detectors, 783
SEM. *See* Secondary electron multiplier (SEM)
Semiconductors, 474
–counter, 222
–detectors, 1134, 1138
–radiation detectors, 380, 407
–spectrometers, 439–445
SENECA, 597
Setup errors, 1156, 1158, 1159
Shales, 677–680
Shannon–Nyquist theorem, 891
Shepp–Logan filter, 893
Shower
–age, 599
–curvature, 604, 605
–leakage, 500, 501, 503, 507, 510
Shower–detector plane, 611
SIBYLL, 606, 613
Sideband, 95
Si detectors, 440, 442–443
Si(Li) detectors, 398, 399, 408, 442–444, 710–712
Sievert, 205
Signal
–charge measurement, 43–48
–formation, 31–37
–processing, digital, 56, 57, 74, 76–80
–speed, 502
Signal-to-noise ratio (SNR), 686, 691, 695, 697, 1088, 1089
Significance test, *p*-value, 115, 116
Silicon detector, 270–273
Silicon drift detectors (SDDs), 710–712
Silicon photomultipliers (SiPM), 308, 943, 1138, 1142, 1147
Silicon-strip detectors (SSD), 271, 272, 570, 580
Silver coins, 848–850
Simplified reference tissue model (SRTM), 1074, 1075, 1079
SIMS. *See* Secondary ionization mass spectroscopy (SIMS)
Simulation, 1097–1122
–fast, 285, 286
–full, 285, 287
–Monte Carlo, 285, 958, 1097–1100, 1119
–software, 1120–1121
Single-mask GEM, 256, 257
Single-photon emission computed tomography (SPECT), 861–867, 918–932, 1044, 1046–1048, 1050, 1052, 1053, 1055–1060, 1126, 1127, 1133–1139, 1141–1147, 1157, 1158
–/computed tomography (SPECT-CT), 863–864
Single photon emission mammography (SPEM), 1146
Single-scatter model, 958
Sinogram, 887–889, 893, 897, 898
SiPM. *See* Silicon photomultipliers (SiPM)
SLD experiment, 464
Slow-scan charge-coupled device, 711
Small animals, 1126–1143, 1146
Smoother, 282, 283, 288
–formulae, 289, 290
SNO experiment, 467
SNR. *See* Signal-to-noise ratio (SNR)
Software framework in large experiments, 98–100
Soft X-rays, 801, 803
Solar energetic particles (SEP), 573
Solar modulation, 573
Solenoid, 476–478, 484, 491
Solid-state annular detectors, 710
Solid-state detectors, 523, 575
SOPB. *See* Spread out Bragg peak (SOPB)
Sounding rockets, 560
Source strength, 417, 422, 428, 429, 449
Space-based instruments, 560
Space charge, 247, 251
Space point, 292
Spallation, 761, 762, 765, 774, 781
Spatial distribution between electrons, 1197
Spatial resolution, 1046–1050, 1052, 1127–1129, 1131–1139, 1141, 1143, 1146, 1190, 1191, 1202, 1204

Specific energy loss dE/dx, 5–8, 11–13, 126–130
SPECT. *See* Single-photon emission computed tomography (SPECT)
SPECT-CT. *See* Single-photon emission computed tomography / Computed Tomography (SPECT-CT)
Spectra, 414, 417–420, 422, 424–426, 429, 430, 446, 448, 449
Spectral CT, 884, 903–905, 911
Spectral interferences, 688, 690, 691
Spectroscopy, 414, 415, 418, 422, 423, 431–449
Spectrum, 90, 96–99
–stripping, 422, 428–429
SPEM. *See* Single photon emission mammography (SPEM)
Spherical GEM, 253, 254
Spiral CT, 884, 897–898, 902, 906, 911
Spiral interpolation method, 898
Spirometers, 1161
Spontaneous fission, 210
Spreading kernel, 415
Spread out Bragg peak (SOPB), 1186, 1199
SPS. *See* Super proton synchrotron (SPS)
SRTM. *See* Simplified reference tissue model (SRTM)
SSD. *See* Silicon-strip detectors (SSD)
Stable isotope, 687, 688, 691
Standard deviation, 106, 107, 109, 110, 112, 113, 115, 118, 120, 413, 415, 416, 423–425, 427
Standardized uptake value (SUV), 1055
Standard model, 84, 91
State vector, 289
Statistical test
–goodness-of-fit, 116
–hypothesis, 114–115
–significance, 114–116
Stellarator, 796, 797
Stereoscopic Cherenkov telescope, 466, 468
Stereoscopic observation, 612, 620
Stoichiometry, 551, 553, 554
Storage rings, 154
Straßmann, F., 203
Stratified sampling, 1119
Stretcher, 72, 73
Strip detectors, 271, 272
–charge division, 49–52
–current pulses, 76
–induced charge, 32–33
Student's t distribution, 107
Successive approximation, 72, 73
Successive-approximation ADC, 72
SUGAR, 607
SuperB experiment, 466
Superheated emulsion detectors, 761, 771, 772
Super-Kamiokande experiment (Super-K), 466
Superposition model, 602, 603, 606
Super proton synchrotron (SPS), 146–148, 150, 156
Suppression of neutron background, 1172
Surface
–contamination, 225
–cylinder, 292
–detector arrays, 604–607, 614, 623
–imaging, 1160
–plane, 292
Survival
–curve, 1196
–experiments, 1196
SUV. *See* Standardized uptake value (SUV)
Symbolic Monte Carlo (SMC), 422, 431
Synchronous, 70, 77, 78
Synchrotron, 144, 146, 154–157, 1189
–radiation, 154–155, 161–184, 214, 215
–micro-CT, 907–911
Systematic, 1159
Système international de référence (SIR), key comparison, 198
System equations, 288, 289

T

Tag-and-Probe approach, 95
Tamm, I., 454
Tangential direction, 949
Targets, 733–735, 752
TDC. *See* Time-to-digital converter (TDC)
Telescope array (TA), 604, 614, 624
TEM. *See* Transmission electron microscopy (TEM)
TEM/EDS, 551, 553, 554
Temperature, 536–538, 540, 542, 546, 553
TEPC. *See* Tissue equivalent proportional counter (TEPC)
Test-retest, 1045
Test statistic, 114
Tetrakis dimethyl-amine ethylene (TMAE), 305
Tevatron, 96, 150
Texture analysis, 1055
Theory of radio radiation, 614
Therapy modalities, 1180
Thermal annealing, 537, 542, 551, 553
Thermal ionization, 688
Thermal neutrons, 721, 722, 729, 733, 736–738, 742, 753, 754
Thermistor, 49
Thermography, 800
Thermoluminescence dosimeter (TLD), 224, 531, 771–772
Thick-GEM (THGEM), 241, 249–255
3D reprojection algorithm (3DRP), 989, 1003
Threshold Cherenkov counter, 133, 464
Through-silicon-vias, 260
THz systems, 635, 637–639
Tibet AS-γ, 604, 607

TIGER, 585, 586
Time, 54–56, 58–67, 69–75, 77–80
Time and frequency domain, 54, 55
Time of flight (TOF), 90, 581, 690, 724, 742, 744, 1117, 1118, 1133, 1139, 1140
Time-of-propagation (TOP) counter, 466
Timepix ASIC, 258, 259
Time projection chamber (TPC), magnetic field, 270
Time structure of calorimeter signals, 499
Time-to-digital converter (TDC), 74–75
Timing, 54, 64–67, 70, 80
  –measurements, 54, 64–66
Tissue-equivalent proportional counter (TEPC), 761, 775–777
Tissue-fraction effect, 1048
Tissue weighting factor, 206, 207
Titanium
  –gold, 526
  –silver, 526
  –undermetal, 526, 528
TLD. *See* Thermoluminescence dosimeter (TLD)
TOF. *See* Time of flight (TOF)
Tokamaks, 797–800, 805, 808
Tomography
  –limited contrast, 1044
  –noise, 1044
  –quantitation, 1044
  –spatial resolution, 1044
Toroid, 476–478, 480, 484
Total x-ray production cross sections, 838
Townsend coefficient, 244, 251
TRACER, 586
Tracer kinetic modeling, 1066, 1067
Track(ing), 1161, 1164, 1166, 1169
  –bundling, 282
  –curvature, 87
  –dip angle, 285
  –finding, 86–87, 280, 282
  –fitting, 281–283, 287–289
  –helix, 87
  –hit, 86, 87
  –impact parameter, 91
  –length integral, 599
  –model, 280–282, 288–290, 292
  –parameters, 280–283, 285, 287–289, 293
  –performance, analytical calculation, 285
  –reconstruction, 281–282, 287
    –bremsstrahlung, 281, 282
    –energy loss, 281
    –material budget, 279, 281
    –particle trajectory, 281
    –reference track, 282
    –stepwise numerical integration, 281
  –regions, over-instrumentation, 286
  –search, 279, 280
  –sensitive detector, 274
  –system, 86
Tracker
  –gaseous, 278, 286
  –layout, 284, 292
  –silicon, 276, 284, 286
Transit dose verification, 1167
Transition, 147, 148
  –radiation, 475, 491, 705
  –region, 285
Transition radiation detector (TRD), 586
Translation–rotation CT, 895–896
Transmission electron microscopy (TEM), 551, 552
Transmittance, 540–543, 546–548
Transverse map of beam intensities, 1188
Trapping, 402, 403, 444
TRD. *See* Transition radiation detector (TRD)
Treatment planning (system), 1167, 1168, 1184, 1186–1189, 1192–1193, 1195, 1201–1203
Tri-ethyl-amine (TEA), 305
Triple-to-double coincidence ratio (TDCR), 196
Tritium, 794, 795, 797, 806, 808
Tube current modulation (TCM), 907
Tumor response to carbon ions, 1201
Tumor delineation, 1157
Tune, 143, 144, 148
Tunka, 609, 620

U

UHECR. *See* Ultrahigh-energy cosmic ray (UHECR)
ULEIS, 575
Ultrahigh-energy cosmic ray (UHECR), 572, 588–590
Ultrasound imaging, 1161
ULYSSES, 576
Uncertainty
  –statistical, 93, 95, 98
  –systematic, 93, 95, 96
Underground CO_2 storage, 669
Underground neutrino detector, 466–467
Undersampling, 78
Unfolding, 98
Uniform distribution, 107, 115
Unwanted side reactions, 1186

V

Vacuum ultraviolet, 803, 808
Valence, 432, 433, 439, 440, 444
  –band, 522
Variance, 426, 438, 444
  –reduction, 1119
VERITAS, 609

Vertex
–detector, 284, 286, 287
–finding, 280, 282, 283
–fitting
–geometric, 282, 283
–kinematic, 280, 283
–robust, 291–292
–locator, 520
–parameters, 282
–reconstruction
–beam interaction profile, 283
–beam tube, 282, 286, 287
–beauty factory, 284
–hard assignment, 283
–kinematic constraints, 282
–perigee parameters, 282
–RAVE toolkit, 283
–soft assignment, 283
–virtual measurements, 282
–ZvTop algorithm, 283
–secondary, 91, 92
Volcano Ranch, 607, 612
Voltage divider, 35–37, 46
Voltage-to-frequency converter, 691
Volume of distribution, 1068, 1072, 1079

W

Walk, 65, 66
Waste transmutation, 721, 755, 756
Water Cherenkov telescope, 457, 469
Water phantom, 1189, 1193
Wavelength-dispersive spectroscopy (WDS), 432, 446–449
Wavelength shifters (WLS), 502
Wavelet transform, 1051
Weak interaction, 314, 317, 330, 331, 341
Weighting factor, 1075–1076
Weighting function, 33
Whipple, 609
Whole-body dose, 206, 227, 231–234
Wiener estimator, 1091, 1092
Wilkinson ADC, 73–75
Wire chamber, 268–269
Wolter, H., 564
Wolter type-I, 565

X

Xe, 562
Xe/CH_4, 561, 586
Xenon CT, 901
XPS. *See* X-ray Photoelectron Spectroscopy (XPS)
X-ray diffraction (XRD), 670
X-ray photoelectron spectroscopy (XPS), 551
X-rays, 161–165, 169, 176–184, 414, 431, 432, 434, 443, 444, 446, 447, 449
–applications, 230
–cargo examination systems, 643, 648–649
–characteristic, 215
–coronary angiography, 215
–diagnostics, 215, 231
–generation, 1108
–neutron examination systems, 647
–pallet examination systems, 635, 643–649
–passenger baggage systems, 635, 643–647
–photon examination systems, 643
–reflectometry, 820
–shipping container examination systems, 635, 643, 647, 649
–spectrum, 215
–therapy, 231
–total per capita dose, 231
–tube, 215, 1141
–yield, 839

Y

Yakutsk, 607, 609
Yield drift, 693, 694

Z

Zerodur, 565
Zircon, 695–698