Lecture Notes in Computer Science 4490

Commenced Publication in 1973
Founding and Former Series Editors:
Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Editorial Board

David Hutchison
Lancaster University, UK

Takeo Kanade
Carnegie Mellon University, Pittsburgh, PA, USA

Josef Kittler
University of Surrey, Guildford, UK

Jon M. Kleinberg
Cornell University, Ithaca, NY, USA

Friedemann Mattern
ETH Zurich, Switzerland

John C. Mitchell
Stanford University, CA, USA

Moni Naor
Weizmann Institute of Science, Rehovot, Israel

Oscar Nierstrasz
University of Bern, Switzerland

C. Pandu Rangan
Indian Institute of Technology, Madras, India

Bernhard Steffen
University of Dortmund, Germany

Madhu Sudan
Massachusetts Institute of Technology, MA, USA

Demetri Terzopoulos
University of California, Los Angeles, CA, USA

Doug Tygar
University of California, Berkeley, CA, USA

Moshe Y. Vardi
Rice University, Houston, TX, USA

Gerhard Weikum
Max-Planck Institute of Computer Science, Saarbruecken, Germany

Yong Shi Geert Dick van Albada
Jack Dongarra Peter M.A. Sloot (Eds.)

Computational Science – ICCS 2007

7th International Conference
Beijing, China, May 27 - 30, 2007
Proceedings, Part IV

Springer

Volume Editors

Yong Shi
Graduate University of the Chinese Academy of Sciences
Beijing 100080, China
E-mail: yshi@gucas.ac.cn

Geert Dick van Albada
Peter M.A. Sloot
University of Amsterdam, Section Computational Science
1098 SJ Amsterdam, The Netherlands
E-mail: {dick, sloot}@science.uva.nl

Jack Dongarra
University of Tennessee, Computer Science Department
Knoxville, TN 37996-3450, USA
E-mail: dongarra@cs.utk.edu

Library of Congress Control Number: 200792049

CR Subject Classification (1998): F, D, G, H, I.1, I.3, I.6, J, K.3, C.2-3

LNCS Sublibrary: SL 1 – Theoretical Computer Science and General Issues

ISSN 0302-9743
ISBN-10 3-540-72589-X Springer Berlin Heidelberg New York
ISBN-13 978-3-540-72589-3 Springer Berlin Heidelberg New York

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, re-use of illustrations, recitation, broadcasting, reproduction on microfilms 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.

Springer is a part of Springer Science+Business Media

springer.com

© Springer-Verlag Berlin Heidelberg 2007
Printed in Germany

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 12065783 06/3180 5 4 3 2 1 0

Preface

The Seventh International Conference on Computational Science (ICCS 2007) was held in Beijing, China, May 27-30, 2007. This was the continuation of previous conferences in the series: ICCS 2006 in Reading, UK; ICCS 2005 in Atlanta, Georgia, USA; ICCS 2004 in Krakow, Poland; ICCS 2003 held simultaneously at two locations in, Melbourne, Australia and St. Petersburg, Russia; ICCS 2002 in Amsterdam, The Netherlands; and ICCS 2001 in San Francisco, California, USA. Since the first conference in San Francisco, the ICCS series has become a major platform to promote the development of Computational Science. The theme of ICCS 2007 was "Advancing Science and Society through Computation." It aimed to bring together researchers and scientists from mathematics and computer science as basic computing disciplines, researchers from various application areas who are pioneering the advanced application of computational methods to sciences such as physics, chemistry, life sciences, and engineering, arts and humanitarian fields, along with software developers and vendors, to discuss problems and solutions in the area, to identify new issues, and to shape future directions for research, as well as to help industrial users apply various advanced computational techniques.

During the opening of ICCS 2007, Siwei Cheng (Vice-Chairman of the Standing Committee of the National People's Congress of the People's Republic of China and the Dean of the School of Management of the Graduate University of the Chinese Academy of Sciences) presented the welcome speech on behalf of the Local Organizing Committee, after which Hector Ruiz (President and CEO, AMD) made remarks on behalf of international computing industries in China. Seven keynote lectures were delivered by Vassil Alexandrov (Advanced Computing and Emerging Technologies, University of Reading, UK) - Efficient Scalable Algorithms for Large-Scale Computations; Hans Petter Langtangen (Simula Research Laboratory, Lysaker, Norway) - Computational Modelling of Huge Tsunamis from Asteroid Impacts; Jiawei Han (Department of Computer Science, University of Illinois at Urbana-Champaign, USA) - Research Frontiers in Advanced Data Mining Technologies and Applications; Ru-qian Lu (Institute of Mathematics, Chinese Academy of Sciences) - Knowledge Engineering and Knowledge Ware; Alessandro Vespignani (School of Informatics, Indiana University, USA) -Computational Epidemiology and Emergent Disease Forecast; David Keyes (Department of Applied Physics and Applied Mathematics, Columbia University) - Scalable Solver Infrastructure for Computational Science and Engineering; and Yves Robert (Ecole Normale Suprieure de Lyon , France) - Think Before Coding: Static Strategies (and Dynamic Execution) for Clusters and Grids. We would like to express our thanks to all of the invited and keynote speakers for their inspiring talks. In addition to the plenary sessions, the conference included 14 parallel oral sessions and 4 poster sessions. This year, we

received more than 2,400 submissions for all tracks combined, out of which 716 were accepted.

This includes 529 oral papers, 97 short papers, and 89 poster papers, spread over 35 workshops and a main track. For the main track we had 91 papers (80 oral papers and 11 short papers) in the proceedings, out of 360 submissions. We had some 930 people doing reviews for the conference, with 118 for the main track. Almost all papers received three reviews. The accepted papers are from more than 43 different countries and 48 different Internet top-level domains.

The papers cover a large volume of topics in computational science and related areas, from multiscale physics to wireless networks, and from graph theory to tools for program development.

We would like to thank all workshop organizers and the Program Committee for the excellent work in maintaining the conference's standing for high-quality papers. We would like to express our gratitude to staff and graduates of the Chinese Academy of Sciences Research Center on Data Technology and Knowledge Economy and the Institute of Policy and Management for their hard work in support of ICCS 2007. We would like to thank the Local Organizing Committee and Local Arrangements Committee for their persistent and enthusiastic work towards the success of ICCS 2007. We owe special thanks to our sponsors, AMD, Springer; University of Nebraska at Omaha, USA and Graduate University of Chinese Academy of Sciences, for their generous support.

ICCS 2007 was organized by the Chinese Academy of Sciences Research Center on Data Technology and Knowledge Economy, with support from the Section Computational Science at the Universiteit van Amsterdam and Innovative Computing Laboratory at the University of Tennessee, in cooperation with the Society for Industrial and Applied Mathematics (SIAM), the International Association for Mathematics and Computers in Simulation (IMACS), the Chinese Society for Management Modernization (CSMM), and the Chinese Society of Optimization, Overall Planning and Economical Mathematics (CSOOPEM).

May 2007 Yong Shi

Organization

ICCS 2007 was organized by the Chinese Academy of Sciences Research Center on Data Technology and Knowledge Economy, with support from the Section Computational Science at the Universiteit van Amsterdam and Innovative Computing Laboratory at the University of Tennessee, in cooperation with the Society for Industrial and Applied Mathematics (SIAM), the International Association for Mathematics and Computers in Simulation (IMACS), and the Chinese Society for Management Modernization (CSMM).

Conference Chairs

Conference Chair - Yong Shi (Chinese Academy of Sciences, China/University of Nebraska at Omaha USA)
Program Chair - Dick van Albada (Universiteit van Amsterdam, The Netherlands)
ICCS Series Overall Scientific Co-chair - Jack Dongarra (University of Tennessee, USA)
ICCS Series Overall Scientific Chair - Peter M.A. Sloot (Universiteit van Amsterdam, The Netherlands)

Local Organizing Committee

Weimin Zheng (Tsinghua University, Beijing, China) – Chair
Hesham Ali (University of Nebraska at Omaha, USA)
Chongfu Huang (Beijing Normal University, Beijing, China)
Masato Koda (University of Tsukuba, Japan)
Heeseok Lee (Korea Advanced Institute of Science and Technology, Korea)
Zengliang Liu (Beijing University of Science and Technology, Beijing, China)
Jen Tang (Purdue University, USA)
Shouyang Wang (Academy of Mathematics and System Science, Chinese Academy of Sciences, Beijing, China)
Weixuan Xu (Institute of Policy and Management, Chinese Academy of Sciences, Beijing, China)
Yong Xue (Institute of Remote Sensing Applications, Chinese Academy of Sciences, Beijing, China)
Ning Zhong (Maebashi Institute of Technology, USA)
Hai Zhuge (Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China)

Local Arrangements Committee

Weixuan Xu, Chair
Yong Shi, Co-chair of events
Benfu Lu, Co-chair of publicity
Hongjin Yang, Secretary
Jianping Li, Member
Ying Liu, Member
Jing He, Member
Siliang Chen, Member
Guanxiong Jiang, Member
Nan Xiao, Member
Zujin Deng, Member

Sponsoring Institutions

AMD
Springer
World Scientific Publlishing
University of Nebraska at Omaha, USA
Graduate University of Chinese Academy of Sciences
Institute of Policy and Management, Chinese Academy of Sciences Universiteit van Amsterdam

Program Committee

J.H. Abawajy, Deakin University, Australia
D. Abramson, Monash University, Australia
V. Alexandrov, University of Reading, UK
I. Altintas, San Diego Supercomputer Center, UCSD
M. Antolovich, Charles Sturt University, Australia
E. Araujo, Universidade Federal de Campina Grande, Brazil
M.A. Baker, University of Reading, UK
B. Balis, Krakow University of Science and Technology, Poland
A. Benoit, LIP, ENS Lyon, France
I. Bethke, University of Amsterdam, The Netherlands
J.A.R. Blais, University of Calgary, Canada
I. Brandic, University of Vienna, Austria
J. Broeckhove, Universiteit Antwerpen, Belgium
M. Bubak, AGH University of Science and Technology, Poland
K. Bubendorfer, Victoria University of Wellington, Australia
B. Cantalupo, DATAMAT S.P.A, Italy
J. Chen Swinburne, University of Technology, Australia
O. Corcho, University of Manchester, UK
J.C. Cunha, Univ. Nova de Lisboa, Portugal

S. Date, Osaka University, Japan
F. Desprez, INRIA, France
T. Dhaene, University of Antwerp, Belgium
I.T. Dimov, ACET, The University of Reading, UK
J. Dongarra, University of Tennessee, USA
F. Donno, CERN, Switzerland
C. Douglas, University of Kentucky, USA
G. Fox, Indiana University, USA
W. Funika, Krakow University of Science and Technology, Poland
H.J. Gardner, Australian National University, Australia
G. Geethakumari, University of Hyderabad, India
Y. Gorbachev, St. Petersburg State Polytechnical University, Russia
A.M. Goscinski, Deakin University, Australia
M. Govindaraju, Binghamton University, USA
G.A. Gravvanis, Democritus University of Thrace, Greece
D.J. Groen, University of Amsterdam, The Netherlands
T. Gubala, ACC CYFRONET AGH, Krakow, Poland
M. Hardt, FZK, Germany
T. Heinis, ETH Zurich, Switzerland
L. Hluchy, Institute of Informatics, Slovak Academy of Sciences, Slovakia
A.G. Hoekstra, University of Amsterdam, The Netherlands
W. Hoffmann, University of Amsterdam, The Netherlands
C. Huang, Beijing Normal University Beijing, China
M. Humphrey, University of Virginia, USA
A. Iglesias, University of Cantabria, Spain
H. Jin, Huazhong University of Science and Technology, China
D. Johnson, ACET Centre, University of Reading, UK
B.D. Kandhai, University of Amsterdam, The Netherlands
S. Kawata, Utsunomiya University, Japan
W.A. Kelly, Queensland University of Technology, Australia
J. Kitowski, Inst.Comp.Sci. AGH-UST, Cracow, Poland
M. Koda, University of Tsukuba Japan
D. Kranzlmüller, GUP, Joh. Kepler University Linz, Austria
B. Kryza, Academic Computer Centre CYFRONET-AGH, Cracow, Poland
M. Kunze, Forschungszentrum Karlsruhe (FZK), Germany
D. Kurzyniec, Emory University, Atlanta, USA
A. Lagana, University of Perugia, Italy
J. Lee, KISTI Supercomputing Center, Korea
C. Lee, Aerospace Corp., USA
L. Lefevre, INRIA, France
A. Lewis, Griffith University, Australia
H.W. Lim, Royal Holloway, University of London, UK
A. Lin, NCMIR/UCSD, USA
P. Lu, University of Alberta, Canada
M. Malawski, Institute of Computer Science AGH, Poland

M. Mascagni, Florida State University, USA
V. Maxville, Curtin Business School, Australia
A.S. McGough, London e-Science Centre, UK
E.D. Moreno, UEA-BENq, Manaus, Brazil
J.T. Moscicki, Cern, Switzerland
S. Naqvi, CoreGRID Network of Excellence, France
P.O.A. Navaux, Universidade Federal do Rio Grande do Sul, Brazil
Z. Nemeth, Computer and Automation Research Institute, Hungarian Academy of Science, Hungary
J. Ni, University of Iowa, USA
G. Norman, Joint Institute for High Temperatures of RAS, Russia
B. Ó Nualláin, University of Amsterdam, The Netherlands
C.W. Oosterlee, Centrum voor Wiskunde en Informatica, CWI, The Netherlands
S. Orlando, Università Ca' Foscari, Venice, Italy
M. Paprzycki, IBS PAN and SWPS, Poland
M. Parashar, Rutgers University, USA
L.M. Patnaik, Indian Institute of Science, India
C.P. Pautasso, ETH Zürich, Switzerland
R. Perrott, Queen's University, Belfast, UK
V. Prasanna, University of Southern California, USA
T. Priol, IRISA, France
M.R. Radecki, Krakow University of Science and Technology, Poland
M. Ram, C-DAC Bangalore Centre, India
A. Rendell, Australian National University, Australia
P. Rhodes, University of Mississippi, USA
M. Riedel, Research Centre Juelich, Germany
D. Rodríguez García, University of Alcalá, Spain
K. Rycerz, Krakow University of Science and Technology, Poland
R. Santinelli, CERN, Switzerland
J. Schneider, Technische Universität Berlin, Germany
B. Schulze, LNCC, Brazil
J. Seo, The University of Manchester, UK
Y. Shi, Chinese Academy of Sciences, Beijing, China
D. Shires, U.S. Army Research Laboratory, USA
A.E. Solomonides, University of the West of England, Bristol, UK
V. Stankovski, University of Ljubljana, Slovenia
H. Stockinger, Swiss Institute of Bioinformatics, Switzerland
A. Streit, Forschungszentrum Jülich, Germany
H. Sun, Beihang University, China
R. Tadeusiewicz, AGH University of Science and Technology, Poland
J. Tang, Purdue University USA
M. Taufer, University of Texas El Paso, USA
C. Tedeschi, LIP-ENS Lyon, France
A. Thandavan, ACET Center, University of Reading, UK
A. Tirado-Ramos, University of Amsterdam, The Netherlands

P. Tvrdik, Czech Technical University Prague, Czech Republic
G.D. van Albada, Universiteit van Amsterdam, The Netherlands
F. van Lingen, California Institute of Technology, USA
J. Vigo-Aguiar, University of Salamanca, Spain
D.W. Walker, Cardiff University, UK
C.L. Wang, University of Hong Kong, China
A.L. Wendelborn, University of Adelaide, Australia
Y. Xue, Chinese Academy of Sciences, China
L.T. Yang, St. Francis Xavier University, Canada
C.T. Yang, Tunghai University, Taichung, Taiwan
J. Yu, The University of Melbourne, Australia
Y. Zheng, Zhejiang University, China
W. Zheng, Tsinghua University, Beijing, China
L. Zhu, University of Florida, USA
A. Zomaya, The University of Sydney, Australia
E.V. Zudilova-Seinstra, University of Amsterdam, The Netherlands

Reviewers

J.H. Abawajy
D. Abramson
A. Abran
P. Adriaans
W. Ahn
R. Akbani
K. Akkaya
R. Albert
M. Aldinucci
V.N. Alexandrov
B. Alidaee
I. Altintas
K. Altmanninger
S. Aluru
S. Ambroszkiewicz
L. Anido
K. Anjyo
C. Anthes
M. Antolovich
S. Antoniotti
G. Antoniu
H. Arabnia
E. Araujo
E. Ardeleanu
J. Aroba
J. Astalos
B. Autin
M. Babik
G. Bai
E. Baker
M.A. Baker
S. Balfe
B. Balis
W. Banzhaf
D. Bastola
S. Battiato
M. Baumgarten
M. Baumgartner
P. Beckaert
A. Belloum
O. Belmonte
A. Belyaev
A. Benoit
G. Bergantz
J. Bernsdorf
J. Berthold
I. Bethke
I. Bhana
R. Bhowmik
M. Bickelhaupt
J. Bin Shyan
J. Birkett
J.A.R. Blais
A. Bode
B. Boghosian
S. Bolboaca
C. Bothorel
A. Bouteiller
I. Brandic
S. Branford
S.J. Branford
R. Braungarten
R. Briggs
J. Broeckhove
W. Bronsvoort
A. Bruce
C. Brugha
Y. Bu
K. Bubendorfer
I. Budinska
G. Buemi
B. Bui
H.J. Bungartz
A. Byrski
M. Cai
Y. Cai
Y.Q. Cai
Z.Y. Cai

B. Cantalupo
K. Cao
M. Cao
F. Capkovic
A. Cepulkauskas
K. Cetnarowicz
Y. Chai
P. Chan
G.-L. Chang
S.C. Chang
W.A. Chaovalitwongse
P.K. Chattaraj
C.-K. Chen
E. Chen
G.Q. Chen
G.X. Chen
J. Chen
J. Chen
J.J. Chen
K. Chen
Q.S. Chen
W. Chen
Y. Chen
Y.Y. Chen
Z. Chen
G. Cheng
X.Z. Cheng
S. Chiu
K.E. Cho
Y.-Y. Cho
B. Choi
J.K. Choi
D. Choinski
D.P. Chong
B. Chopard
M. Chover
I. Chung
M. Ciglan
B. Cogan
G. Cong
J. Corander
J.C. Corchado
O. Corcho
J. Cornil
H. Cota de Freitas
E. Coutinho
J.J. Cuadrado-Gallego
Y.F. Cui
J.C. Cunha
V. Curcin
A. Curioni
R. da Rosa Righi
S. Dalai
M. Daneva
S. Date
P. Dazzi
S. de Marchi
V. Debelov
E. Deelman
J. Della Dora
Y. Demazeau
Y. Demchenko
H. Deng
X.T. Deng
Y. Deng
M. Mat Deris
F. Desprez
M. Dewar
T. Dhaene
Z.R. Di
G. di Biasi
A. Diaz Guilera
P. Didier
I.T. Dimov
L. Ding
G.D. Dobrowolski
T. Dokken
J.J. Dolado
W. Dong
Y.-L. Dong
J. Dongarra
F. Donno
C. Douglas
G.J. Garcke
R.P. Mundani
R. Drezewski
D. Du
B. Duan
J.F. Dufourd
H. Dun
C. Earley
P. Edmond
T. Eitrich
A. El Rhalibi
T. Ernst
V. Ervin
D. Estrin
L. Eyraud-Dubois
J. Falcou
H. Fang
Y. Fang
X. Fei
Y. Fei
R. Feng
M. Fernandez
K. Fisher
C. Fittschen
G. Fox
F. Freitas
T. Friesz
K. Fuerlinger
M. Fujimoto
T. Fujinami
W. Funika
T. Furumura
A. Galvez
L.J. Gao
X.S. Gao
J.E. Garcia
H.J. Gardner
M. Garre
G. Garsva
F. Gava
G. Geethakumari
M. Geimer
J. Geiser
J.-P. Gelas
A. Gerbessiotis
M. Gerndt
S. Gimelshein
S.G. Girdzijauskas
S. Girtelschmid
Z. Gj
C. Glasner
A. Goderis

D. Godoy
J. Golebiowski
S. Gopalakrishnan
Y. Gorbachev
A.M. Goscinski
M. Govindaraju
E. Grabska
G.A. Gravvanis
C.H. Grelck
D.J. Groen
L. Gross
P. Gruer
A. Grzech
J.F. Gu
Y. Guang Xue
T. Gubala
V. Guevara-Masis
C.H. Guo
X. Guo
Z.Q. Guo
L. Guohui
C. Gupta
I. Gutman
A. Haffegee
K. Han
M. Hardt
A. Hasson
J. He
J. He
K. He
T. He
J. He
M.R. Head
P. Heinzlreiter
H. Chojnacki
J. Heo
S. Hirokawa
G. Hliniak
L. Hluchy
T.B. Ho
A. Hoekstra
W. Hoffmann
A. Hoheisel
J. Hong
Z. Hong
D. Horvath
F. Hu
L. Hu
X. Hu
X.H. Hu
Z. Hu
K. Hua
H.W. Huang
K.-Y. Huang
L. Huang
L. Huang
M.S. Huang
S. Huang
T. Huang
W. Huang
Y. Huang
Z. Huang
Z. Huang
B. Huber
E. Hubo
J. Hulliger
M. Hultell
M. Humphrey
P. Hurtado
J. Huysmans
T. Ida
A. Iglesias
K. Iqbal
D. Ireland
N. Ishizawa
I. Lukovits
R. Jamieson
J.K. Jan
P. Janderka
M. Jankowski
L. Jäntschi
S.J.K. Jensen
N.J. Jeon
T.H. Jeon
T. Jeong
H. Ji
X. Ji
D.Y. Jia
C. Jiang
H. Jiang
M.J. Jiang
P. Jiang
W. Jiang
Y. Jiang
H. Jin
J. Jin
L. Jingling
G.-S. Jo
D. Johnson
J. Johnstone
J.J. Jung
K. Juszczyszyn
J.A. Kaandorp
M. Kabelac
B. Kadlec
R. Kakkar
C. Kameyama
B.D. Kandhai
S. Kandl
K. Kang
S. Kato
S. Kawata
T. Kegl
W.A. Kelly
J. Kennedy
G. Khan
J.B. Kido
C.H. Kim
D.S. Kim
D.W. Kim
H. Kim
J.G. Kim
J.H. Kim
M. Kim
T.H. Kim
T.W. Kim
P. Kiprof
R. Kirner
M. Kisiel-Dorohinicki
J. Kitowski
C.R. Kleijn
M. Kluge
A. Knüpfer
I.S. Ko
Y. Ko

R. Kobler
B. Koblitz
G.A. Kochenberger
M. Koda
T. Koeckerbauer
M. Koehler
I. Kolingerova
V. Korkhov
T. Korkmaz
L. Kotulski
G. Kou
J. Kozlak
M. Krafczyk
D. Kranzlmüller
B. Kryza
V.V. Krzhizhanovskaya
M. Kunze
D. Kurzyniec
E. Kusmierek
S. Kwang
Y. Kwok
F. Kyriakopoulos
H. Labiod
A. Lagana
H. Lai
S. Lai
Z. Lan
G. Le Mahec
B.G. Lee
C. Lee
H.K. Lee
J. Lee
J. Lee
J.H. Lee
S. Lee
S.Y. Lee
V. Lee
Y.H. Lee
L. Lefevre
L. Lei
F. Lelj
A. Lesar
D. Lesthaeghe
Z. Levnajic
A. Lewis
A. Li
D. Li
D. Li
E. Li
J. Li
J. Li
J.P. Li
M. Li
P. Li
X. Li
X.M. Li
X.S. Li
Y. Li
Y. Li
J. Liang
L. Liang
W.K. Liao
X.F. Liao
G.G. Lim
H.W. Lim
S. Lim
A. Lin
I.C. Lin
I-C. Lin
Y. Lin
Z. Lin
P. Lingras
C.Y. Liu
D. Liu
D.S. Liu
E.L. Liu
F. Liu
G. Liu
H.L. Liu
J. Liu
J.C. Liu
R. Liu
S.Y. Liu
W.B. Liu
X. Liu
Y. Liu
Y. Liu
Y. Liu
Y. Liu
Y.J. Liu
Y.Z. Liu
Z.J. Liu
S.-C. Lo
R. Loogen
B. López
A. López García de Lomana
F. Loulergue
G. Lu
J. Lu
J.H. Lu
M. Lu
P. Lu
S. Lu
X. Lu
Y.C. Lu
C. Lursinsap
L. Ma
M. Ma
T. Ma
A. Macedo
N. Maillard
M. Malawski
S. Maniccam
S.S. Manna
Z.M. Mao
M. Mascagni
E. Mathias
R.C. Maurya
V. Maxville
A.S. McGough
R. Mckay
T.-G. MCKenzie
K. Meenal
R. Mehrotra
M. Meneghin
F. Meng
M.F.J. Meng
E. Merkevicius
M. Metzger
Z. Michalewicz
J. Michopoulos
J.-C. Mignot
R. mikusauskas
H.Y. Ming

G. Miranda Valladares
M. Mirua
G.P. Miscione
C. Miyaji
A. Miyoshi
J. Monterde
E.D. Moreno
G. Morra
J.T. Moscicki
H. Moshkovich
V.M. Moskaliova
G. Mounie
C. Mu
A. Muraru
H. Na
K. Nakajima
Y. Nakamori
S. Naqvi
S. Naqvi
R. Narayanan
A. Narjess
A. Nasri
P. Navaux
P.O.A. Navaux
M. Negoita
Z. Nemeth
L. Neumann
N.T. Nguyen
J. Ni
Q. Ni
K. Nie
G. Nikishkov
V. Nitica
W. Nocon
A. Noel
G. Norman
B. Ó Nualláin
N. O'Boyle
J.T. Oden
Y. Ohsawa
H. Okuda
D.L. Olson
C.W. Oosterlee
V. Oravec
S. Orlando
F.R. Ornellas
A. Ortiz
S. Ouyang
T. Owens
S. Oyama
B. Ozisikyilmaz
A. Padmanabhan
Z. Pan
Y. Papegay
M. Paprzycki
M. Parashar
K. Park
M. Park
S. Park
S.K. Pati
M. Pauley
C.P. Pautasso
B. Payne
T.C. Peachey
S. Pelagatti
F.L. Peng
Q. Peng
Y. Peng
N. Petford
A.D. Pimentel
W.A.P. Pinheiro
J. Pisharath
G. Pitel
D. Plemenos
S. Pllana
S. Ploux
A. Podoleanu
M. Polak
D. Prabu
B.B. Prahalada Rao
V. Prasanna
P. Praxmarer
V.B. Priezzhev
T. Priol
T. Prokosch
G. Pucciani
D. Puja
P. Puschner
L. Qi
D. Qin
H. Qin
K. Qin
R.X. Qin
X. Qin
G. Qiu
X. Qiu
J.Q. Quinqueton
M.R. Radecki
S. Radhakrishnan
S. Radharkrishnan
M. Ram
S. Ramakrishnan
P.R. Ramasami
P. Ramsamy
K.R. Rao
N. Ratnakar
T. Recio
K. Regenauer-Lieb
R. Rejas
F.Y. Ren
A. Rendell
P. Rhodes
J. Ribelles
M. Riedel
R. Rioboo
Y. Robert
G.J. Rodgers
A.S. Rodionov
D. Rodríguez García
C. Rodriguez Leon
F. Rogier
G. Rojek
L.L. Rong
H. Ronghuai
H. Rosmanith
F.-X. Roux
R.K. Roy
U. Rüde
M. Ruiz
T. Ruofeng
K. Rycerz
M. Ryoke
F. Safaei
T. Saito
V. Sakalauskas

L. Santillo
R. Santinelli
K. Sarac
H. Sarafian
M. Sarfraz
V.S. Savchenko
M. Sbert
R. Schaefer
D. Schmid
J. Schneider
M. Schoeberl
S.-B. Scholz
B. Schulze
S.R. Seelam
B. Seetharamanjaneyalu
J. Seo
K.D. Seo
Y. Seo
O.A. Serra
A. Sfarti
H. Shao
X.J. Shao
F.T. Sheldon
H.Z. Shen
S.L. Shen
Z.H. Sheng
H. Shi
Y. Shi
S. Shin
S.Y. Shin
B. Shirazi
D. Shires
E. Shook
Z.S. Shuai
M.A. Sicilia
M. Simeonidis
K. Singh
M. Siqueira
W. Sit
M. Skomorowski
A. Skowron
P.M.A. Sloot
M. Smolka
B.S. Sniezynski
H.Z. Sojka
A.E. Solomonides
C. Song
L.J. Song
S. Song
W. Song
J. Soto
A. Sourin
R. Srinivasan
V. Srovnal
V. Stankovski
P. Sterian
H. Stockinger
D. Stokic
A. Streit
B. Strug
P. Stuedi
A. Stümpel
S. Su
V. Subramanian
P. Suganthan
D.A. Sun
H. Sun
S. Sun
Y.H. Sun
Z.G. Sun
M. Suvakov
H. Suzuki
D. Szczerba
L. Szecsi
L. Szirmay-Kalos
R. Tadeusiewicz
B. Tadic
T. Takahashi
S. Takeda
J. Tan
H.J. Tang
J. Tang
S. Tang
T. Tang
X.J. Tang
J. Tao
M. Taufer
S.F. Tayyari
C. Tedeschi
J.C. Teixeira
F. Terpstra
C. Te-Yi
A.Y. Teymorian
D. Thalmann
A. Thandavan
L. Thompson
S. Thurner
F.Z. Tian
Y. Tian
Z. Tianshu
A. Tirado-Ramos
A. Tirumala
P. Tjeerd
W. Tong
A.S. Tosun
A. Tropsha
C. Troyer
K.C.K. Tsang
A.C. Tsipis
I. Tsutomu
A. Turan
P. Tvrdik
U. Ufuktepe
V. Uskov
B. Vaidya
E. Valakevicius
I.A. Valuev
S. Valverde
G.D. van Albada
R. van der Sman
F. van Lingen
A.J.C. Varandas
C. Varotsos
D. Vasyunin
R. Veloso
J. Vigo-Aguiar
J. Villà i Freixa
V. Vivacqua
E. Vumar
R. Walentkynski
D.W. Walker
B. Wang
C.L. Wang
D.F. Wang
D.H. Wang

F. Wang
F.L. Wang
H. Wang
H.G. Wang
H.W. Wang
J. Wang
J. Wang
J. Wang
J. Wang
J.H. Wang
K. Wang
L. Wang
M. Wang
M.Z. Wang
Q. Wang
Q.Q. Wang
S.P. Wang
T.K. Wang
W. Wang
W.D. Wang
X. Wang
X.J. Wang
Y. Wang
Y.Q. Wang
Z. Wang
Z.T. Wang
A. Wei
G.X. Wei
Y.-M. Wei
X. Weimin
D. Weiskopf
B. Wen
A.L. Wendelborn
I. Wenzel
A. Wibisono
A.P. Wierzbicki
R. Wismüller
F. Wolf
C. Wu
C. Wu
F. Wu
G. Wu
J.N. Wu
X. Wu
X.D. Wu
Y. Wu
Z. Wu
B. Wylie
M. Xavier Py
Y.M. Xi
H. Xia
H.X. Xia
Z.R. Xiao
C.F. Xie
J. Xie
Q.W. Xie
H. Xing
H.L. Xing
J. Xing
K. Xing
L. Xiong
M. Xiong
S. Xiong
Y.Q. Xiong
C. Xu
C.-H. Xu
J. Xu
M.W. Xu
Y. Xu
G. Xue
Y. Xue
Z. Xue
A. Yacizi
B. Yan
N. Yan
N. Yan
W. Yan
H. Yanami
C.T. Yang
F.P. Yang
J.M. Yang
K. Yang
L.T. Yang
L.T. Yang
P. Yang
X. Yang
Z. Yang
W. Yanwen
S. Yarasi
D.K.Y. Yau
P.-W. Yau
M.J. Ye
G. Yen
R. Yi
Z. Yi
J.G. Yim
L. Yin
W. Yin
Y. Ying
S. Yoo
T. Yoshino
W. Youmei
Y.K. Young-Kyu Han
J. Yu
J. Yu
L. Yu
Z. Yu
Z. Yu
W. Yu Lung
X.Y. Yuan
W. Yue
Z.Q. Yue
D. Yuen
T. Yuizono
J. Zambreno
P. Zarzycki
M.A. Zatevakhin
S. Zeng
A. Zhang
C. Zhang
D. Zhang
D.L. Zhang
D.Z. Zhang
G. Zhang
H. Zhang
H.R. Zhang
H.W. Zhang
J. Zhang
J.J. Zhang
L.L. Zhang
M. Zhang
N. Zhang
P. Zhang
P.Z. Zhang
Q. Zhang

S. Zhang
W. Zhang
W. Zhang
Y.G. Zhang
Y.X. Zhang
Z. Zhang
Z.W. Zhang
C. Zhao
H. Zhao
H.K. Zhao
H.P. Zhao
J. Zhao
M.H. Zhao
W. Zhao
Z. Zhao
L. Zhen
B. Zheng
G. Zheng
W. Zheng
Y. Zheng
W. Zhenghong
P. Zhigeng
W. Zhihai
Y. Zhixia
A. Zhmakin
C. Zhong
X. Zhong
K.J. Zhou
L.G. Zhou
X.J. Zhou
X.L. Zhou
Y.T. Zhou
H.H. Zhu
H.L. Zhu
L. Zhu
X.Z. Zhu
Z. Zhu
M. Zhu.
J. Zivkovic
A. Zomaya
E.V. Zudilova-Seinstra

Workshop Organizers

Sixth International Workshop on Computer Graphics and Geometric Modelling

A. Iglesias, University of Cantabria, Spain

Fifth International Workshop on Computer Algebra Systems and Applications

A. Iglesias, University of Cantabria, Spain,
A. Galvez, University of Cantabria, Spain

PAPP 2007 - Practical Aspects of High-Level Parallel Programming (4th International Workshop)

A. Benoit, ENS Lyon, France
F. Loulerge, LIFO, Orlans, France

International Workshop on Collective Intelligence for Semantic and Knowledge Grid (CISKGrid 2007)

N.T. Nguyen, Wroclaw University of Technology, Poland
J.J. Jung, INRIA Rhône-Alpes, France
K. Juszczyszyn, Wroclaw University of Technology, Poland

Simulation of Multiphysics Multiscale Systems, 4th International Workshop

V.V. Krzhizhanovskaya, Section Computational Science, University of Amsterdam, The Netherlands
A.G. Hoekstra, Section Computational Science, University of Amsterdam, The Netherlands

S. Sun, Clemson University, USA
J. Geiser, Humboldt University of Berlin, Germany

2nd Workshop on Computational Chemistry and Its Applications (2nd CCA)

P.R. Ramasami, University of Mauritius

Efficient Data Management for HPC Simulation Applications

R.-P. Mundani, Technische Universität München, Germany
J. Abawajy, Deakin University, Australia
M. Mat Deris, Tun Hussein Onn College University of Technology, Malaysia

Real Time Systems and Adaptive Applications (RTSAA-2007)

J. Hong, Soongsil University, South Korea
T. Kuo, National Taiwan University, Taiwan

The International Workshop on Teaching Computational Science (WTCS 2007)

L. Qi, Department of Information and Technology, Central China Normal University, China
W. Yanwen, Department of Information and Technology, Central China Normal University, China
W. Zhenghong, East China Normal University, School of Information Science and Technology, China

GeoComputation

Y. Xue, IRSA, China

Risk Analysis

C.F. Huang, Beijing Normal University, China

Advanced Computational Approaches and IT Techniques in Bioinformatics

M.A. Pauley, University of Nebraska at Omaha, USA
H.A. Ali, University of Nebraska at Omaha, USA

Workshop on Computational Finance and Business Intelligence

Y. Shi, Chinese Acedemy of Scienes, China
S.Y. Wang, Academy of Mathematical and System Sciences, Chinese Academy of Sciences, China
X.T. Deng, Department of Computer Science, City University of Hong Kong, China

Collaborative and Cooperative Environments

C. Anthes, Institute of Graphics and Parallel Processing, JKU, Austria
V.N. Alexandrov, ACET Centre, The University of Reading, UK
D. Kranzlmüller, Institute of Graphics and Parallel Processing, JKU, Austria
J. Volkert, Institute of Graphics and Parallel Processing, JKU, Austria

Tools for Program Development and Analysis in Computational Science

A. Knüpfer, ZIH, TU Dresden, Germany
A. Bode, TU Munich, Germany
D. Kranzlmüller, Institute of Graphics and Parallel Processing, JKU, Austria
J. Tao, CAPP, University of Karlsruhe, Germany
R. Wissmüller FB12, BSVS, University of Siegen, Germany
J. Volkert, Institute of Graphics and Parallel Processing, JKU, Austria

Workshop on Mining Text, Semi-structured, Web or Multimedia Data (WMTSWMD 2007)

G. Kou, Thomson Corporation, R&D, USA
Y. Peng, Omnium Worldwide, Inc., USA
J.P. Li, Institute of Policy and Management, Chinese Academy of Sciences, China

2007 International Workshop on Graph Theory, Algorithms and Its Applications in Computer Science (IWGA 2007)

M. Li, Dalian University of Technology, China

2nd International Workshop on Workflow Systems in e-Science (WSES 2007)

Z. Zhao, University of Amsterdam, The Netherlands
A. Belloum, University of Amsterdam, The Netherlands

2nd International Workshop on Internet Computing in Science and Engineering (ICSE 2007)

J. Ni, The University of Iowa, USA

Workshop on Evolutionary Algorithms and Evolvable Systems (EAES 2007)

B. Zheng, College of Computer Science, South-Central University for Nationalities, Wuhan, China
Y. Li, State Key Lab. of Software Engineering, Wuhan University, Wuhan, China
J. Wang, College of Computer Science, South-Central University for Nationalities, Wuhan, China
L. Ding, State Key Lab. of Software Engineering, Wuhan University, Wuhan, China

Wireless and Mobile Systems 2007 (WMS 2007)

H. Choo, Sungkyunkwan University, South Korea

WAFTS: WAvelets, FracTals, Short-Range Phenomena — Computational Aspects and Applications

C. Cattani, University of Salerno, Italy
C. Toma, Polythecnica, Bucharest, Romania

Dynamic Data-Driven Application Systems - DDDAS 2007

F. Darema, National Science Foundation, USA

The Seventh International Workshop on Meta-synthesis and Complex Systems (MCS 2007)

X.J. Tang, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, China
J.F. Gu, Institute of Systems Science, Chinese Academy of Sciences, China
Y. Nakamori, Japan Advanced Institute of Science and Technology, Japan
H.C. Wang, Shanghai Jiaotong University, China

The 1st International Workshop on Computational Methods in Energy Economics

L. Yu, City University of Hong Kong, China
J. Li, Chinese Academy of Sciences, China
D. Qin, Guangdong Provincial Development and Reform Commission, China

High-Performance Data Mining

Y. Liu, Data Technology and Knowledge Economy Research Center, Chinese Academy of Sciences, China
A. Choudhary, Electrical and Computer Engineering Department, Northwestern University, USA
S. Chiu, Department of Computer Science, College of Engineering, Idaho State University, USA

Computational Linguistics in Human–Computer Interaction

H. Ji, Sungkyunkwan University, South Korea
Y. Seo, Chungbuk National University, South Korea
H. Choo, Sungkyunkwan University, South Korea

Intelligent Agents in Computing Systems

K. Cetnarowicz, Department of Computer Science, AGH University of Science and Technology, Poland
R. Schaefer, Department of Computer Science, AGH University of Science and Technology, Poland

Networks: Theory and Applications

B. Tadic, Jozef Stefan Institute, Ljubljana, Slovenia
S. Thurner, COSY, Medical University Vienna, Austria

Workshop on Computational Science in Software Engineering

D. Rodrguez, University of Alcala, Spain
J.J. Cuadrado-Gallego, University of Alcala, Spain

International Workshop on Advances in Computational Geomechanics and Geophysics (IACGG 2007)

H.L. Xing, The University of Queensland and ACcESS Major National Research Facility, Australia
J.H. Wang, Shanghai Jiao Tong University, China

2nd International Workshop on Evolution Toward Next-Generation Internet (ENGI)

Y. Cui, Tsinghua University, China

Parallel Monte Carlo Algorithms for Diverse Applications in a Distributed Setting

V.N. Alexandrov, ACET Centre, The University of Reading, UK

The 2007 Workshop on Scientific Computing in Electronics Engineering (WSCEE 2007)

Y. Li, National Chiao Tung University, Taiwan

High-Performance Networked Media and Services 2007 (HiNMS 2007)

I.S. Ko, Dongguk University, South Korea
Y.J. Na, Honam University, South Korea

Table of Contents – Part IV

Roadmapping and *i*-Systems 1
Tieju Ma, Hongbin Yan, and Yoshiteru Nakamori

Exploring Computational Scheme of Complex Problem Solving Based on Meta-Synthesis Approach 9
Yijun Liu, Wenyuan Niu, and Jifa Gu

ICT and Special Educational Needs: Using Meta-synthesis for Bridging the Multifaceted Divide 18
Ying Liu, Annita Cornish, and June Clegg

Discovering Latent Structures: Experience with the CoIL Challenge 2000 Data Set 26
Nevin L. Zhang

Exploration of TCM Masters Knowledge Mining 35
Xijin Tang, Nan Zhang, and Zheng Wang

A Numerical Trip to Social Psychology: Long-Living States of Cognitive Dissonance 43
P. Gawroński and K. Kułakowski

A Hidden Pattern Discovery and Meta-synthesis of Preference Adjustment in Group Decision-Making 51
Huizhang Shen, Jidi Zhao, and Huanchen Wang

Discussion on the Spike Train Recognition Mechanisms in Neural Circuits 59
Yan Liu, Liujun Chen, Jiawei Chen, Fangfeng Zhang, and Fukang Fang

Extended Clustering Coefficients of Small-World Networks 67
Wenjun Xiao, Yong Qin, and Behrooz Parhami

Detecting Invisible Relevant Persons in a Homogeneous Social Network 74
Yoshiharu Maeno, Kiichi Ito, and Yukio Ohsawa

The Study of Size Distribution and Spatial Distribution of Urban Systems in Guangdong, China 82
Jianmei Yang, Dong Zhuang, and Minyi Kuang

Emergence of Social Rumor: Modeling, Analysis, and Simulations 90
ZhengYou Xia and LaiLei Huang

Emergence of Specialization from Global Optimizing Evolution in a Multi-agent System 98
Lei Chai, Jiawei Chen, Zhangang Han, Zengru Di, and Ying Fan

A Hybrid Econometric-AI Ensemble Learning Model for Chinese Foreign Trade Prediction 106
Lean Yu, Shouyang Wang, and Kin Keung Lai

The Origin of Volatility Cascade of the Financial Market 114
Chunxia Yang, Yingchao Zhang, Hongfa Wu, and Peiling Zhou

Tactical Battlefield Entities Simulation Model Based on Multi-agent Interactions 121
Xiong Li and Sheng Dang

Extensive Epidemic Spreading Model Based on Multi-agent System Framework 129
Chunhua Tian, Wei Ding, Rongzeng Cao, and Shun Jiang

Simulation of Employee Behavior Based on Cellular Automata Model ... 134
Yue Jiao, Shaorong Sun, and Xiaodong Sun

Modeling, Learning and Simulating Biological Cells with Entity Grammar 138
Yun Wang, Rao Zheng, and Yan-Jiang Qiao

Chance Discovery in Credit Risk Management: Estimation of Chain Reaction Bankruptcy Structure by Directed KeyGraph 142
Shinichi Goda and Yukio Ohsawa

Text Classification with Support Vector Machine and Back Propagation Neural Network 150
Wen Zhang, Xijin Tang, and Taketoshi Yoshida

Construction and Application of PSO-SVM Model for Personal Credit Scoring 158
Ming-hui Jiang and Xu-chuan Yuan

Feature Description Systems for Clusters by Using Logical Rule Generations Based on the Genetic Programming and Its Applications to Data Mining 162
Jianjun Lu, Yunling Liu, and Shozo Tokinaga

Artificial Immunity-Based Discovery for Popular Information in WEB Pages 166
Caiming Liu, Xiaojie Liu, Tao Li, Lingxi Peng, Jinquan Zeng, and Hui Zhao

Network Structure and Knowledge Transfer 170
Fangcheng Tang

Information Relationship Identification in Team Innovation 174
Xinmiao Li, Xinhui Li, and Pengzhu Zhang

Agile Knowledge Supply Chain for Emergency Decision-Making Support 178
Qingquan Wang and Lili Rong

Interactive Fuzzy Goal Programming Approach for Optimization of Extended Hub-and-Spoke Regional Port Transportation Networks 186
Chuanxu Wang and Liangkui Jiang

A Pseudo-Boolean Optimization for Multiple Criteria Decision Making in Complex Systems 194
Bahram Alidaee, Haibo Wang, and Yaquan Xu

The Study of Mission Reliability of QRMS Based on the Multistage Markov Process 202
Liang Liang and Bo Guo

Performance Analysis and Evaluation of Digital Connection Oriented Internet Service Systems 210
Shunfu Jin and Wuyi Yue

A Knowledge-Based Model Representation and On-Line Solution Method for Dynamic Vehicle Routing Problem 218
Lijun Sun, Xiangpei Hu, Zheng Wang, and Minfang Huang

Numerical Simulation of Static Noise Margin for a Six-Transistor Static Random Access Memory Cell with 32nm Fin-Typed Field Effect Transistors 227
Yiming Li, Chih-Hong Hwang, and Shao-Ming Yu

Numerical Solution to Maxwell's Equations in Singular Waveguides 235
Franck Assous and Patrick Ciarlet Jr.

Quantum-Inspired Genetic Algorithm Based Time-Frequency Atom Decomposition 243
Gexiang Zhang and Haina Rong

Latency Estimation of the Asynchronous Pipeline Using the Max-Plus Algebra 251
Jian Ruan, Zhiying Wang, Kui Dai, and Yong Li

A Simulation-Based Hybrid Optimization Technique for Low Noise Amplifier Design Automation 259
Yiming Li, Shao-Ming Yu, and Yih-Lang Li

Spectral Collocation Technique for Absorbing Boundary Conditions with Increasingly High Order Approximation 267
Zhenli Xu and Houde Han

Shockwave Detection for Electronic Vehicle Detectors 275
Hsung-Jung Cho and Ming-Te Tseng

Contour Extraction Algorithm Using a Robust Neural Network 283
Zhou Zhiheng, Li Zhengfang, and Zeng Delu

A Discrete Parameter-Driven Time Series Model for Traffic Flow in ITS 291
Yow-Jen Jou and Yan-Chu Huang

Peer-Based Efficient Content Distribution in Ad Hoc Networks 295
Seung-Seok Kang

Session Key Reuse Scheme to Improve Routing Efficiency in AnonDSR 303
Chunum Kong, Min Young Chung, and Hyunseung Choo

Clustering in Ad Hoc Personal Network Formation 312
Yanying Gu, Weidong Lu, R.V. Prasad, and Ignas Niemegeers

Message Complexity Analysis of MANET Address Autoconfiguration Algorithms in Group Merging Case 320
Sang-Chul Kim

A Robust Route Maintenance Scheme for Wireless Ad-Hoc Networks . . . 328
Kwan-Woong Kim, Mike Myung-Ok Lee, ChangKug Kim, and Yong-Kab Kim

Route Optimization with MAP-Based Enhancement in Mobile Networks 336
Jeonghoon Park, Tae-Jin Lee, and Hyunseung Choo

Performance Enhancement Schemes of OFDMA System for Broadband Wireless Access 344
Dong-Hyun Park, So-Young Yeo, Jee-Hoon Kim, Young-Hwan You, and Hyoung-Kyu Song

Performance Analysis of Digital Wireless Networks with ARQ Schemes 352
Wuyi Yue and Shunfu Jin

A Novel Frequency Offset Estimation Algorithm Using Differential Combining for OFDM-Based WLAN Systems 360
Sangho Ahn, Sanghun Kim, Hyoung-Kee Choi, Sun Yong Kim, and Seokho Yoon

Design and Performance Evaluation of High Efficient TCP for HBDP Networks 368
TaeJoon Park, ManKyu Park, JaeYong Lee, and ByungChul Kim

A Reliable Transmission Strategy in Unreliable Wireless Sensor Networks 376
Zhendong Wu and Shanping Li

Genetic Algorithmic Topology Control for Two-Tiered Wireless Sensor Networks 385
Donghwan Lee, Wonjun Lee, and Joongheon Kim

A Delay Sensitive Feedback Control Data Aggregation Approach in Wireless Sensor Network 393
Peng Shao-liang, Li Shan-shan, Peng Yu-xing, Zhu Pei-dong, and Xiao Nong

A Low Power Real-Time Scheduling Scheme for the Wireless Sensor Network 401
Mikyung Kang and Junghoon Lee

Analysis of an Adaptive Key Selection Scheme in Wireless Sensor Networks 409
Guorui Li, Jingsha He, and Yingfang Fu

Unusual Event Recognition for Mobile Alarm System 417
Sooyeong Kwak, Guntae Bae, Kilcheon Kim, and Hyeran Byun

Information Exchange for Controlling Internet Robots 425
Soon Hyuk Hong, Ji-Hwan Park, Key Ho Kwon, and Jae Wook Jeon

A Privacy-Aware Identity Design for Exploring Ubiquitous Collaborative Wisdom 433
Yuan-Chu Hwang and Soe-Tsyr Yuan

Performance Comparison of Sleep Mode Operations in IEEE 802.16e Terminals 441
Youn-Hee Han, Sung-Gi Min, and Dongwon Jeong

Performance Evaluation of the Optimal Hierarchy for Cellular Networks 449
So-Jeong Park, Gyung-Leen Park, In-Hye Shin, Junghoon Lee, Ho Young Kwak, Do-Hyeon Kim, Sang Joon Lee, and Min-Soo Kang

Channel Time Allocation and Routing Algorithm for Multi-hop Communications in IEEE 802.15.3 High-Rate WPAN Mesh Networks ... 457
Ssang-Bong Jung, Hyun-Ki Kim, Soon-Bin Yim, and Tae-Jin Lee

Nonlinear Optimization of IEEE 802.11 Mesh Networks 466
Enrique Costa-Montenegro, Francisco J. González-Castaño, Pedro S. Rodríguez-Hernández, and Juan C. Burguillo-Rial

Securely Deliver Data by Multi-path Routing Scheme in Wireless Mesh Networks 474
Cao Trong Hieu and Choong Seon Hong

Cross-Layer Enhancement of IEEE 802.11 MAC for Mobile Ad Hoc Networks 482
Taekon Kim, Hyungkeun Lee, Jang-Yeon Lee, and Jin-Woong Cho

An Incremental Topology Control Algorithm for Wireless Mesh Networks 490
Mani Malekesmaeili, Mehdi Soltan, and Mohsen Shiva

TCP Adaptation for Vertical Handoff Using Network Monitoring 498
Faraz Idris Khan and Eui Nam Huh

Optimization of Mobile IPv6 Handover Performance Using E-HCF Method 506
Guozhi.Wei, Anne.Wei, Ke Xu, and Gerard.Dupeyrat

HMIPv6 Applying User's Mobility Pattern in IP-Based Cellular Networks 514
Teail Shin, Hyungmo Kang, and Youngsong Mun

Performance Analysis and Comparison of the MIPv6 and mSCTP Based Vertical Handoff 522
Shi Yan, Chen Shanzhi, Ai Ming, and Hu Bo

Reliability of Wireless Sensor Network with Sleeping Nodes 530
Vladimir V. Shakhov and Hyunseung Choo

Energy Efficient Forwarding Scheme for Secure Wireless Ad Hoc Routing Protocols 534
Kwonseung Shin, Min Young Chung, and Hyunseung Choo

Sender-Based TCP Scheme for Improving Performance in Wireless Environment 538
Jahwan Koo, Sung-Gon Mun, and Hyunseung Choo

Design and Implementation of DLNA DMS Through IEEE1394 542
Gu Su Kim, Chul-Seung Kim, Hyun-Su Jang, Moon Seok Chang, and Young Ik Eom

Efficient Measurement of the Eye Blinking by Using Decision Function for Intelligent Vehicles 546
Ilkwon Park, Jung-Ho Ahn, and Hyeran Byun

Dynamic Bandwidth Allocation Algorithm Based on Two-Phase Cycle for Efficient Channel Utilization on Ethernet PON 550
Won Jin Yoon, Woo Jin Jung, Tae-Jin Lee, Hyunseung Choo, and Min Young Chung

Performance Evaluation of Binary Negative-Exponential Backoff Algorithm in Presence of a Channel Bit Error Rate 554
Bum-Gon Choi, Hyung Joo Ki, Min Young Chung, and Tae-Jin Lee

A Rough Set Based Anomaly Detection Scheme Considering the Age of User Profiles 558
Ihn-Han Bae

Space-Time Coded MB-OFDM UWB System with Multi-channel Estimation for Wireless Personal Area Networks 562
Bon-Wook Koo, Myung-Sun Baek, Jee-Hoon Kim, and Hyoung-Kyu Song

Performance Enhancement of Multimedia Data Transmission by Adjusting Compression Rate 566
Eung Ju Lee, Kyu Seol Lee, and Hee Yong Youn

A Feasible Approach to Assigning System Components to Hybrid Task Sets in Real-Time Sensor Networking Platforms 570
Kyunghoon Jung, Byounghoon Kim, Changsoo Kim, and Sungwoo Tak

Efficient Routing Scheme Using Pivot Node in Wireless Sensor Networks 574
Jung-Seok Lee, Jung-Pil Ryu, and Ki-Jun Han

Encoding-Based Tamper-Resistant Algorithm for Mobile Device Security 578
Seok Min Yoon, Seung Wook Lee, Hong Moon Wang, and Jong Tae Kim

Adaptive Vertical Handoff Management Architecture 582
Faraz Idris Khan and Eui Nam Huh

Performance Evaluation of the Route Optimization Scheme in Mobile IPv6 586
In-Hye Shin, Gyung-Leen Park, Junghoon Lee, Jun Hwang, and Taikyeong T. Jeong

An ID-Based Random Key Pre-distribution Scheme for Wireless Sensor Networks 590
Tran Thanh Dai and Choong Seon Hong

An Adaptive Mobile System to Solve Large-Scale Problems in Wireless Networks 594
Jehwan Oh and Eunseok Lee

Answer Extracting Based on Passage Retrieval in Chinese Question Answering System 598
Zhengtao Yu, Lu Han, Cunli Mao, Yunwei Li, Yanxia Qiu, and Xiangyan Meng

Performance Evaluation of Fully Adaptive Routing for the Torus Interconnect Networks 606
F. Safaei, A. Khonsari, M. Fathy, and M. Ould-Khaoua

A Study on Phonemic Analysis for the Recognition of Korean Speech ... 614
Jeong Young Song, Min Wook Kil, and Il Seok Ko

Media Synchronization Framework for SVC Video Transport over IP Networks 621
Kwang-deok Seo, Jin-won Lee, Soon-heung Jung, and Jae-gon Kim

Developing Value Framework of Ubiquitous Computing 629
Jungwoo Lee, Younghee Lee, and Jaesung Park

An Enhanced Positioning Scheme Based on Optimal Diversity for Mobile Nodes in Ubiquitous Networks 636
Seokyong Yang and Sekchin Chang

TOMOON: A Novel Approach for Topology-Aware Overlay Multicasting 644
Xiao Chen, Huagang Shao, and Weinong Wang

Fuzzy-Timing Petri Nets with Choice Probabilities for Response Time Analysis 652
Jaegeol Yim and Kye-Young Lee

A Telematics Service System Based on the Linux Cluster 660
Junghoon Lee, Gyung-Leen Park, Hanil Kim, Young-Kyu Yang, Pankoo Kim, and Sang-Wook Kim

Unequal Error Recovery Scheme for Multimedia Streaming in Application-Level Multicast 668
Joonhyoung Lee, Youngha Jung, and Yoonsik Choe

A Fast Handoff Scheme Between PDSNs in 3G Network 676
Jae-hong Ryu and Dong-Won Kim

Privacy Protection for a Secure u-City Life 685
Changjin Lee, Bong Gyou Lee, and Youngil Kong

Hybrid Tag Anti-collision Algorithms in RFID Systems 693
Jae-Dong Shin, Sang-Soo Yeo, Tai-Hoon Kim, and Sung Kwon Kim

Design and Implement Controllable Multicast Based Audio/Video Collaboration 701
Xuan Zhang, Dongtao Liu, and Xing Li

Solving a Problem in Grid Applications: Using Aspect Oriented Programming 705
Hyuck Han, Shingyu Kim, Hyungsoo Jung, and Heon Y. Yeom

Energy-Aware QoS Adjustment of Multimedia Tasks with Uncertain Execution Time 709
Wan Yeon Lee, Heejo Lee, and Hyogon Kim

SCA-Based Reconfigurable Access Terminal 713
Junsik Kim, Sangchul Oh, Eunseon Cho, Namhoon Park, and Nam Kim

Investigating Media Streaming in Multipath Multihop Wireless Network 717
Binod Vaidya, SangDuck Lee, Eung-Kon Kim, JongAn Park, and SeungJo Han

A Low-Power 512-Bit EEPROM Design for UHF RFID Tag Chips 721
Jae-Hyung Lee, Gyu-Ho Lim, Ji-Hong Kim, Mu-Hun Park, Kyo-Hong Jin, Jeong-won Cha, Pan-Bong Ha, Yung-Jin Gang, and Young-Hee Kim

VPDP: A Service Discovery Protocol for Ubiquitous Computing 725
Zhaomin Xu, Ming Cai, and Jinxiang Dong

A Study on the Aspects of Successful Business Intelligence System Development 729
Il Seok Ko and Sarvar R. Abdullaev

Robust Phase Tracking for High Capacity Wireless Multimedia Data Communication Networks 733
Taehyun Jeon

EVM's Java Dynamic Memory Manager and Garbage Collector 737
Sang-Yun Lee and Byung-Uk Choi

An Access Control Model in Lager-Scale P2P File Sharing Systems 741
Yue Guang-xue, Yu Fei, Chen Li-ping, and Chen Yi-jun

Sink-Independent Model in Wireless Sensor Networks 745
Sang-Sik Kim, Kwang-Ryul Jung, Ki-Il Kim, and Ae-Soon Park

An Update Propagation Algorithm for P2P File Sharing over Wireless Mobile Networks 753
Haengrae Cho

P2P Mobile Multimedia Group Conferencing: Combining SIP, SSM and Scalable Adaptive Coding for Heterogeneous Networks 761
Thomas C. Schmidt, Matthias Wählisch, Hans L. Cycon, and Mark Palkow

Federation Based Solution for Peer-to-Peer Network Management 765
Jilong Wang and Jing Zhang

A Feedback Based Adaptive Marking Algorithm for Assured Service 773
Fanjun Su, Chunxue Wu, and Guoqiang Sun

QoS-Aware MAP Selection Scheme Based on Average Handover Delay for Multimedia Services in Multi-level HMIPv6 Networks 777
Y.-X. Lei and Z.-M. Zeng

On Composite Service Optimization Across Distributed QoS Registries 785
Fei Li, Fangchun Yang, Kai Shuang, and Sen Su

Estimating Flow Length Distributions Using Least Square Method and Maximum Likelihood Estimation 793
Weijiang Liu

Local Link Protection Scheme in IP Networks 797
Hui-Kai Su and Cheng-Shong Wu

An Authentication Based Source Address Spoofing Prevention Method Deployed in IPv6 Edge Network 801
Lizhong Xie, Jun Bi, and Jianpin Wu

An Intrusion Plan Recognition Algorithm Based on Max-1-Connected Causal Networks 809
Zhuo Ning and Jian Gong

Impact of Buffer Map Cheating on the Streaming Quality in DONet 817
Yong Cui, Dan Li, and Jianping Wu

Architecture of STL Model of New Communication Network 825
Aibao Wang and Guangzhao Zhang

Experience with SPM in IPv6 833
Mingjiang Ye, Jianping Wu, and Miao Zhang

Dongting Lake Floodwater Diversion and Storage Modeling and Control Architecture Based on the Next Generation Network 841
Lianqing Xue, Zhenchun Hao, Dan Li, and Xiaoqun Liu

Query Processing to Efficient Search in Ubiquitous Computing..... 849
Byung-Ryong Kim and Ki-Chang Kim

Service and Management for Multicast Based Audio/Video Collaboration System on CERNET 853
Xuan Zhang, Xing Li, and Qingguo Zhao

A Double-Sampling and Hold Based Approach for Accurate and Efficient Network Flow Monitoring 857
Guang Cheng, Yongning Tang, and Wei Ding

A Power Saving Scheme for Heterogeneous Wireless Access Networks . . . 865
SuKyoung Lee, LaeYoung Kim, and Hojin Kim

Efficient GTS Allocation Algorithm for IEEE 802.15.4 869
Youngmin Ji, Woojin Park, Sungjun Kim, and Sunshin An

Hybrid Search Algorithms for P2P Media Streaming Distribution in Ad Hoc Networks 873
Dong-hong Zuo, Xu Du, and Zong-kai Yang

Improving Search on Gnutella-Like P2P Systems 877
Qi Zhao, Jiaoyao Liu, and Jingdong Xu

Non-preemptive Fixed Priority Scheduling of Hard Real-Time Periodic Tasks 881
Moonju Park

A New Mobile Payment Method for Embedded Systems Using Light Signal 889
Hoyoung Hwang, Moonhaeng Huh, Siwoo Byun, and Sungsoo Lim

Bounding Demand Paging Costs in Fixed Priority Real-Time Systems 897
Young-Ho Lee, Hoyoung Hwang, Kanghee Kim, and Sung-Soo Lim

OTL: On-Demand Thread Stack Allocation Scheme for Real-Time Sensor Operating Systems 905
Sangho Yi, Seungwoo Lee, Yookun Cho, and Jiman Hong

EF-Greedy: A Novel Garbage Collection Policy for Flash Memory Based Embedded Systems 913
Ohhoon Kwon, Jaewoo Lee, and Kern Koh

Power-Directed Software Prefetching Algorithm with Dynamic Voltage Scaling 921
Juan Chen, Yong Dong, Huizhan Yi, and Xuejun Yang

An Efficient Bandwidth Reclaim Scheme for the Integrated Transmission of Real-Time and Non-Real-Time Messages on the WLAN 925
Junghoon Lee, In-Hye Shin, Gyung-Leen Park, Wang-Cheol Song, Jinhwan Kim, Pankoo Kim, and Jiman Hong

A Fast Real Time Link Adaptation Scheme for Wireless Communication Systems 933
Hyukjun Oh, Jiman Hong, and Yongseok Kim

EAR: An Energy-Aware Block Reallocation Framework for Energy Efficiency 941
Woo Hyun Ahn

Virtual Development Environment Based on SystemC for Embedded Systems 949
Sang-Young Cho, Yoojin Chung, and Jung-Bae Lee

Embedded Fault Diagnosis Expert System Based on CLIPS and ANN 957
Tan Dapeng, Li Peiyu, and Pan Xiaohong

A Fault-Tolerant Real-Time Scheduling Algorithm in Software Fault-Tolerant Module 961
Dong Liu, Weiyan Xing, Rui Li, Chunyuan Zhang, and Haiyan Li

An Energy-Efficient Scheduling Algorithm for Real-Time Tasks 965
Youlin Ruan, Gan Liu, Jianjun Han, and Qinghua Li

An EDF Interrupt Handling Scheme for Real-Time Kernel: Design and Task Simulation 969
Peng Liu, Ming Cai, Tingting Fu, and Jinxiang Dong

Real-Time Controlled Multi-objective Scheduling Through ANNs and Fuzzy Inference Systems: The Case of DRC Manufacturing 973
Ozlem Uzun Araz

Recursive Priority Inheritance Protocol for Solving Priority Inversion Problems in Real-Time Operating Systems 977
Kwangsun Ko, Seong-Goo Kang, Gyehyeon Gyeong, and Young Ik Eom

An Improved Simplex-Genetic Method to Solve Hard Linear Programming Problems 981
Juan Frausto-Solís and Alma Nieto-Yáñez

Real-Observation Quantum-Inspired Evolutionary Algorithm for a Class of Numerical Optimization Problems 989
Gexiang Zhang and Haina Rong

A Steep Thermodynamical Selection Rule for Evolutionary Algorithms 997
Weiqin Ying, Yuanxiang Li, Shujuan Peng, and Weiwu Wang

A New Hybrid Optimization Algorithm Framework to Solve Constrained Optimization Problem 1005
Huang Zhangcan and Cheng Hao

In Search of Proper Pareto-optimal Solutions Using Multi-objective Evolutionary Algorithms 1013
Pradyumn Kumar Shukla

Cultural Particle Swarm Algorithms for Constrained Multi-objective Optimization 1021
Fang Gao, Qiang Zhao, Hongwei Liu, and Gang Cui

A Novel Multi-objective Evolutionary Algorithm 1029
Bojin Zheng and Ting Hu

New Model for Multi-objective Evolutionary Algorithms 1037
Bojin Zheng and Yuanxiang Li

The Study on a New Immune Optimization Routing Model 1045
Jun Qin, Jiang-qing Wang, and Zi-mao Li

Pheromone Based Dynamic Vaccination for Immune Algorithms 1053
Yutao Qi, Fang Liu, and Licheng Jiao

Towards a Less Destructive Crossover Operator Using Immunity Theory 1061
Yingzhou Bi, Lixin Ding, and Weiqin Ying

Studying the Performance of Quantum Evolutionary Algorithm Based on Immune Theory 1068
Xiaoming You, Sheng Liu, and Dianxun Shuai

Design of Fuzzy Set-Based Polynomial Neural Networks with the Aid of Symbolic Encoding and Information Granulation 1076
Sung-Kwun Oh, In-Tae Lee, and Hyun-Ki Kim

An Heuristic Method for GPS Surveying Problem 1084
Stefka Fidanova

Real-Time DOP Ellipsoid in Polarization Mode Dispersion Monitoring System by Using PSO Algorithm 1091
Xiaoguang Zhang, Gaoyan Duan, and Lixia Xi

Fast Drug Scheduling Optimization Approach for Cancer Chemotherapy 1099
Yong Liang, Kwong-Sak Leung, and Tony Shu Kam Mok

Optimization of IG-Based Fuzzy System with the Aid of GAs and Its Application to Software Process 1108
Sung-Kwun Oh, Keon-Jun Park, and Witold Pedrycz

Evolvable Face Recognition Based on Evolutionary Algorithm and Gabor Wavelet Networks 1116
Chuansheng Wu, Yong Ding, and Lishan Kang

Automated Design Approach for Analog Circuit Using Genetic Algorithm 1124
Xuewen Xia, Yuanxiang Li, Weiqin Ying, and Lei Chen

Evolutionary Algorithm for Identifying Discontinuous Parameters of Inverse Problems 1131
Zhijian Wu, Dazhi Jiang, and Lishan Kang

A WSN Coalition Formation Algorithm Based on Ant Colony with Dual-Negative Feedback 1139
Na Xia, Jianguo Jiang, Meibin Qi, Chunhua Yu, Yue Huang, and Qi Zhang

An Improved Evolutionary Algorithm for Dynamic Vehicle Routing Problem with Time Windows 1147
Jiang-qing Wang, Xiao-nian Tong, and Zi-mao Li

The Geometry Optimization of Argon Atom Clusters Using Differential Evolution Algorithm 1155
Yongxiang Zhao, Shengwu Xiong, and Ning Xu

A Genetic Algorithm for Solving a Special Class of Nonlinear Bilevel Programming Problems 1159
Hecheng Li and Yuping Wang

Evolutionary Strategy for Political Districting Problem Using Genetic Algorithm 1163
Chung-I Chou, You-ling Chu, and Sai-Ping Li

An ACO Algorithm with Adaptive Volatility Rate of Pheromone Trail 1167
Zhifeng Hao, Han Huang, Yong Qin, and Ruichu Cai

A Distributed Coordination Framework for Adaptive Sensor Uncertainty Handling 1171
Zhifeng Dai, Yuanxiang Li, Bojin Zheng, and Xianjun Shen

A Heuristic Particle Swarm Optimization for Cutting Stock Problem Based on Cutting Pattern 1175
Xianjun Shen, Yuanxiang Li, Jincai Yang, and Li Yu

Theory of Evolutionary Algorithm: A View from Thermodynamics 1179
Yuanxiang Li, Weiwu Wang, Xianjun Shen, Weiqin Ying, and Bojin Zheng

Simulated Annealing Parallel Genetic Algorithm Based on Building Blocks Migration 1183
Zhiyong Li and Xilu Zhu

Comparison of Different Integral Performance Criteria for Optimal Hydro Generator Governor Tuning with a Particle Swarm Optimization Algorithm 1186
Hongqing Fang, Long Chen, and Zuyi Shen

Author Index 1191

Roadmapping and *i*-Systems

Tieju Ma[1,2,3], Hongbin Yan[1], and Yoshiteru Nakamori[1]

[1] School of Knowledge Science, Japan Advanced Institute of Science and Technology,
1-1 Asahidai, Tatsunokuchi, Ishikawa 923-1292, Japan
[2] Research Center on Data Technology and Knowledge Economy,
CAS Beijing, China
[3] International Institute for Applied Systems Analysis
A-2361 Laxenburg, Austria
{tieju,hongbinyan,nakamori}@jaist.ac.jp

Abstract. The abstract should summarize the contents of the paper and should contain at least 70 and at most 150 words. It should be set in 9-point font size and should be inset 1.0 cm from the right and left margins. There should be two blank (10-point) lines before and after the abstract. This document is in the required format. Roadmapping, as a strategic planning tool, is attracting increasing applications in industries. By applying the principles of Interactive Planning (IP), this paper puts forward a new solution for making personal academic research roadmaps. Then this paper introduces an ongoing project that is to develop a roadmapping support system based on the solution, and gives some considerations about applying the *i*-System Methodology for enhancing the knowledge creation in a roadmapping process.

Keywords: We would like to encourage you to list your keywords in this section.

1 Introduction

Motorola Inc. firstly introduced the concept of "roadmap" as a strategic planning tool in the 1970s. Perhaps the most widely accepted definition of a roadmap was given by Bob Galvin, CEO of Motorola: "A roadmap is an extended look at the future of a chosen field of inquiry composed from the collective knowledge and imagination of the brightest drivers of change in that field". "Roadmaps" can mean different things to different people. What all those different roadmaps have in common, however, is their goal, to help their owners clarify the following three problems:

- Where are we now?
- Where do we want to go?
- How can we get there?

There are many existing solutions for roadmapping, developed for the purpose of industry, with strong commercial background. Roadmapping for supporting scientific research should be different from those solutions for industry, since academic labs have different features from commercial organizations. The main target of academic labs should be "emerging technology" and "creative invention", and academic labs

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 1–8, 2007.
© Springer-Verlag Berlin Heidelberg 2007

should also have the function for the accumulation and expansion of scientific knowledge and function for inspiring researchers. Based on the Interactive Planning (IP) Methodology, we developed a solution for the roadmapping for supporting scientific research. To improve the efficiency and effectiveness of the roadmapping process, we developed a web-based roadmapping support system. The purpose of making personal academic research roadmaps is not only to make plans, it should also be a knowledge creation process. The *i*-Systems methodology is a systems methodology that uses approaches in social and natural sciences complementarily [14-16], which is very useful for enhancing the knowledge creation in a roadmapping process.

The rest of this paper is organized as follows. Section 2 introduces the Interactive Planning Methodology and explains why we applied it for roadmapping. Section 3 gives the new solution for making personal academic research roadmap. Section 4 introduces the roadmapping support system. Section 5 gives some considerations about applying the i-system methodology for enhancing the knowledge creation in the roadmapping process.

2 Interactive Planning

IP was put forward by Ackoff R.L [1-4]. It is regarded as a basic methodology for solving creative problems by researchers in both the field of management science and the field of systems science. IP has the following three important principles, which we think are also very important in the roadmapping process.

- Participative principle. Ackoff believed that the process of planning is more important than the actual plan produced.
- Continuity principle. This principle points out that planning is a never-ending process, since the values of the organization's stakeholders will change over time and unexpected events will occur [6].
- Holistic principle. This principle insists that people should make plans both simultaneously and interdependently.

IP is composed five interrelated phases, namely, formulating the issue, ends planning, means planning, resource planning, and design of the implementation and controls. These phases should be "regarded as constituting a systemic process, so the phases may be started in any order and none of the phases, let alone the whole process, should ever be regarded as completed [4]".

In the description of IP, the objects are organizations, or systems from the viewpoint of systems science. A personal academic research plan can also been seen as a system inside human brain. In this sense, the five phases of IP can be clearly mapped to the three important questions that roadmapping aims to answer. The first phase of IP, namely "formulating the issue", in fact tries to answer the question "where are we now"; the second phase of IP, "ends planning", corresponds to the problem "where do we want to go"; and the remaining three phases of IP, "means planning", "Resource Planning" and "Design of Implementation and Control" -- are for answering the question "how can we get there". Fig.1 shows the relationship between IP and the three important problems which roadmapping aims to solve.

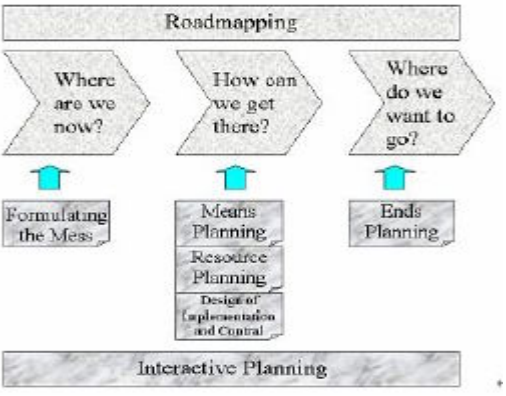


Fig. 1. Relationship between IP and Roadmapping

3 A New Solution for Making Personal Academic Research RoadMaps

Applying the ideas of IP to the process of making personal academic research roadmaps can enhance communication among researchers from different fields, since IP pays much attention to the participation of stakeholders. In addition, an atmosphere can be created in which research on "emerging technology" and "creative invention" are encouraged by what R.L. Ackoff has called "idealized design" [1-4].

Idealized design is a very important feature of IP. It is meant to generate maximum creativity among all the stakeholders involved. "To ensure this, only two types of constraint upon the design are admissible. First, it must be technologically feasible, not a work of science fiction; it must be possible with known technology or likely technological developments; but it should not for example, assume telepathy. Second, it must be operationally viable; it should be capable of working and surviving if it is implemented. Financial, political, or similar constraints are not allowed to restrict the creativity of the design [1-4, 6]." Our solution is composed of six interrelated phases, as shown in Fig.2.

Phase 1: Forming groups. We believe that roadmapping should be a group activity and a consensus building process. A group should contain two kinds of members in addition to the regular members. The first is experienced researchers, for example, professors, associate professors and so on. The second is knowledge coordinators. Knowledge coordinators are those people who can manage creative research activities based on the theory of knowledge creation [14].

Phase 2: Explanation from knowledge coordinators. In this phase, knowledge coordinators explain the following things (mainly) to all group members.

- Concept of roadmaps and the benefits of making roadmaps
- The role of every member
- The schedule of the group

Phase 3: Description of present situation. In this phase, the experienced researchers give a description of the present situation that mainly includes:

- Background knowledge in this research field
- The leading groups/labs, famous papers, journals and conferences over the world in the research field
- The common equipments and skills needed in this field
- Hot topics at current time in this research field

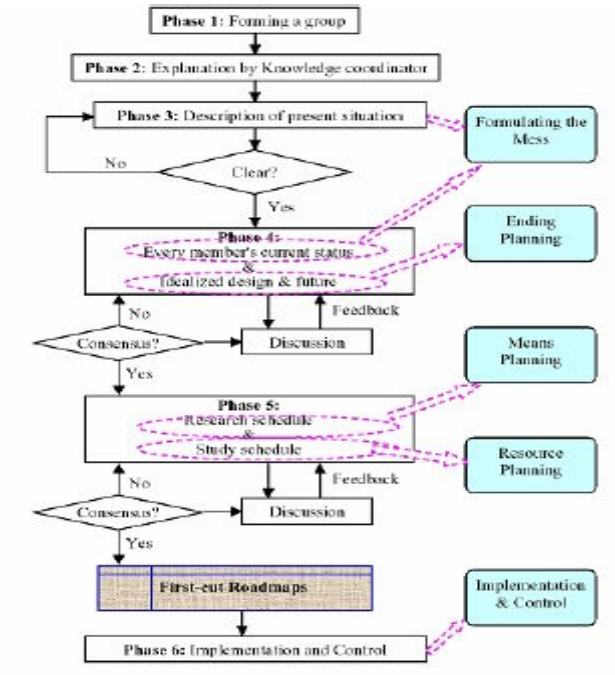


Fig. 2. Solution for making personal academic research roadmaps

Phase 4: Members' current status and idealized design. In this phase, every member firstly describes the experience (the skills and knowledge) he/she already has. Then, by using IP's idealized design, every member describes his/her research goals. The ideas generated by idealized design are discussed by the whole group, and each individual can refine, modify his/her idealized design with the benefit of the whole group's knowledge.

Phase 5: Research schedule and study schedule. In this phase, members not only make their own research schedules, but also make their own study schedules for reaching the goals. Those schedules should also be subjected to group discussion, and members need to modify those schedules according to the result of group discussion. After consensus is reached, group members can start to make their first-cut roadmaps. Road-mapping is a never ending process, people need to go back to some previous phases again and again, modify and improve their research roadmaps continuously.

Phase 6: Implementation and Control. This is mainly done by regular seminars, workshops and reports. By Phase 5, each researcher's personal research roadmap is ready. Although much effort has gone into making a reasonable research roadmap, it is still a first cut. The roadmap should be continuously refined in practice, which accords with the continuity principle of IP. The knowledge coordinator(s) should arrange regular seminars and workshops to monitor and control the implementation of the personal research roadmaps.

4 A RoadMapping Support System

As a project supported by the JAIST COE Program titled "technology creation based on knowledge science", a roadmapping support system is under developing. The benefits of using the system include: Helping researchers to managing their personal roadmaps; Helping the supervisor to managing his/her group/lab's research; Promoting the knowledge sharing among researchers, especially promoting the dispute among researchers; Building roadmap archives that can be used as the source of data mining (knowledge discovery).

The system is a web-based system. Basically, users only need a web browser, such as Internet Explorer or Netscape, and an Internet connection to access it. The following is the main techniques used for developing the system:

- Java [8] and Java Applet [9]. For running the system, client users need to download some Java plug-in. But users do not have to worry about this, the system will atomically check if there is the right plug-in in client computers, if there is no, then it will automatically download it.
- JSP (JavaServer Pages) [10].
- Java Servlet [11].
- Tomcat [7]. We use Tomcat 5.1 as the web server.
- SQL Server 2000 [12]. We use MSSQL Server 2000 as the background database server.

The user can see and modify his research roadmap which has been stored in the system. Besides the function of viewing and editing his/her own research roadmap, there are several functions that users can use.

Users can view other group members' research roadmap by clicking their names or research topics. The system provides two formats of a research roadmap, like an article, or like a table (ATRM model [17]). Users can make comments on other members' research roadmaps. The system allows users to make comments anonymously. As mentioned by Wierzbicki: Far Eastern societies are better than Western at Socialization and achieving consensus but (perhaps therefore) worse in Dispute [18]. Allowing making comments anonymously will promote the dispute among researchers, which is very important for knowledge creation. Sometimes users, especially the leader or supervisor of the group would like to have a general structure or a general view of his/her group's research. The system also provides tools, with which, it will be easy to see what the group is doing now, what it plans to do and when it will do them.

5 The I -Systems Methodology for Enhancing Knowledge Creation in RoadMapping Process

Developed by Nakamori, the *i*-systems methodology uses approaches in social and natural sciences complementarily [14-16]. According to Nakamori in [14], *i*-systems are composed of five subsystems/dimensions, at the subsystem Intervention, knowledge is a problem; at the subsystem Intelligence, knowledge is a model; at the subsystem Imagination, knowledge is scenarios; at the subsystem Involvement, knowledge is opinions; and at the subsystem Integration, knowledge is solutions.

Although the *i*-systems methodology does not give a sequential process or interrelated phases for practice, it identified the important dimensions and gave a clear description of the relationship among the five dimensions. The understandings of the knowledge creation process, which we can learn from the *i*-systems methodology, can help us to design a better knowledge creation space [18].

Interactive Planning Methodology is developed in industry, and it does not pay much attention to the "emerging technology" and "creative invention", which are the main targets of academic labs. So it is necessary to apply some methodologies, such as the *i*-systems methodology, which addresses much the knowledge creation process, to enhance the knowledge creation in the solution introduced in Section 3. Instead of giving some concrete examples of applying i-systems methodology in roadmapping process, this paper only gives some considerations of applying it for making personal research roadmaps.

As shown in Fig.3, we start from the intervention dimension. In this dimension, one researcher should answer questions such as "what do you want to achieve" or "what's your purpose and motivation".

When the researcher finds the answer to those questions, he/she would refer to the social dimension, the scientific dimension and creative dimension (referred to as the three dimensions in the following). In the scientific dimension, the researcher gathers knowledge of the existing research models related to his/her research purpose. This is mainly done by reading literatures. In the social dimension, the researcher gathers the opinions from industry and government, and of course also from other researchers, especially those experienced researchers who work in the same fields. Sometimes, communications with researchers from different field can bring surprising wonderful ideas. And in the creative dimension, the researcher generates his individual ideas, makes his purpose clearer and writes rough research proposals. We would not like to make a sequential process for the actions in the three dimensions because a researcher maybe refers to these three dimensions at the same time, or the researcher will frequently refer to one or more than one dimensions for many times. For example, we could not say clearly that the work of gathering existing research models should be finished in one or two months, and in this one or two months, a researcher does not refer to other dimensions. Researchers should consider the work referring to the three dimensions according to their own schedules. It is obviously that before a researcher answers the questions asked in the intervention process, he/she already referred to the three dimensions, and his/her answer in fact is based on the three dimensions. After answering the questions in intervention dimension, a researcher need deliberately refer to the three dimensions for improving his/her answers. After the answer is determined, the researcher need refer to the three dimensions again for making his/her

research roadmap. During those processes, discussions, brainstorming, seminars, workshops, and other methods of communication, with or without IT support, should be used for the knowledge sharing (learning from each other), and thus to enhance the knowledge creation.

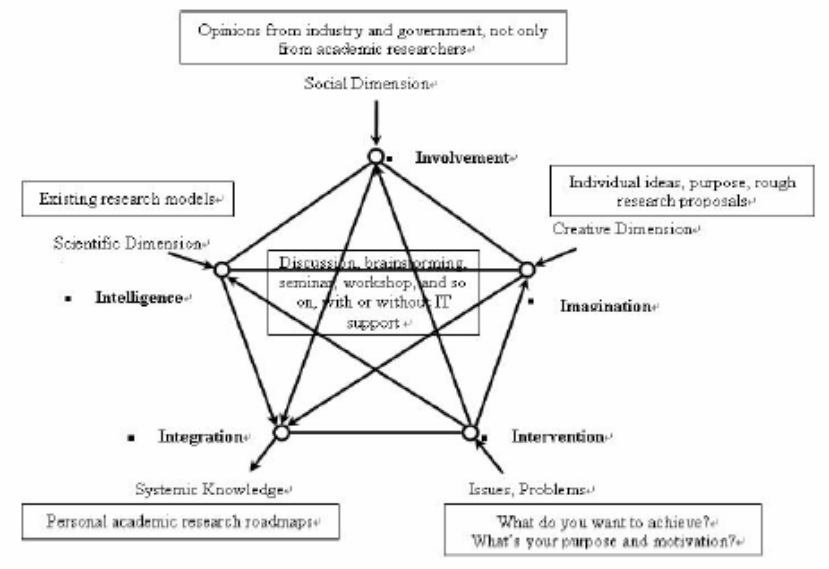


Fig. 3. The i-systems methodology for enhancing the knowledge creation in roadmapping process

In the integration dimension, researchers work out their personal academic research roadmaps, which can be viewed as the solutions for those questions asked in the intervention dimension. Since roadmapping is a never ending process, a researcher should continuously refer to all the five dimensions again and again for improving his/her research roadmap.

6 Concluding Remarks

This paper put forward a solution for making personal academic research roadmaps based on the IP Methodology, introduced a web-based roadmapping support system, and gave some considerations of applying the *i*-systems methodology for enhancing the knowledge creation in roadmapping process. In the practice of roadmapping [13], we found that roadmapping can be an unwieldy and time consuming process, which can discourage participation, while competent knowledge coordinators and proper IT (information technology) support can reduce this negative factor. In practice, we also found that roadmapping is more welcomed by junior researchers than senior researchers. It seems the benefits of roadmapping for junior researchers are obvious than those for senior researchers. Senior researchers are more likely to believe that they can

arrange their research by themselves, and will be reluctant to spend time on roadmapping, but most of them would like to help making juniors researchers' roadmaps. The junior researchers are more likely to find that they can get useful information, knowledge, and good suggestions and ideas through the roadmapping process.

References

1. R. L. Ackoff, Brief Guide to Interactive Planning and Idealized Design, available at:http://www.sociate.com/texts/AckoffGuidetoIdealizedRedesign.pdf, 2001.
2. R. L. Ackoff, Creating the Corporate Future, Wiley, New York, 1981.
3. R. L. Ackoff, Redesigning the Future, New York: Wiley, 1974.
4. R. L. Ackoff, The Art of Problem Solving, New York: Wiley, 1978.
5. B. Bergeron, Essentials of Knowledge Management, Wiley, 2003.
6. R. L. Flood and M. C. Jackson, Interactive Planning, Creative Problem Solving: total systems intervention, Wiley, Chapter 7, pp. 143-165, 1991.
7. http://jakarta.apache.org/tomcat/
8. http://java.sun.com/
9. http://java.sun.com/applets/
10. http://java.sun.com/products/jsp/
11. http://java.sun.com/products/servlet/index.jsp
12. http://www.microsoft.com/sql/default.asp
13. T. Ma, S. Liu, and Y. Nakamori, Roadmapping for supporting Scientific Research, The 17th International Conference on Multiple Criteria Decision Analysis, Canada, August, 2004.
14. Y. Nakamori and M. Takagi, Technology Creation Based on Knowledge Science, Proceedings of the First International Symposium on Knowledge Management for Strategic Creation of Technology, Ishikawa High-Tech Exchange Center, Japan, pp. 1-10, 2004.
15. Y. Nakamori, Knowledge Management System Toward Sustainable Society, Proceedings of the First International Symposium on Knowledge and System Sciences. E. Shimemura, Y. Nakamori, J. Gu and T. Yoshida (eds), 57-65. Ishikawa, Japan: Japan Advance Institute of Science and Technology, 2000.
16. Y. Nakamori, Towards Supporting Technology Creation Based on Knowledge Science, Systems Science and Systems Engineering, ICSSSE'03, Global-Link Publisher, pp. 33-38, 2003.
17. S. Okuzu, A Technology Management Methodology for Academic Science & Engineering Laboratories by Fusion of Soft System Methodology and Technology Road Mapping, masters thesis, Tokyo Institute of Technology, 2002.
18. A. P. Wierzbicki, Knowledge Integration: Creative Space and Creative Environments, working paper and a draft for a book, 2004.

Exploring Computational Scheme of Complex Problem Solving Based on Meta-Synthesis Approach

Yijun Liu[1], Wenyuan Niu[1], and Jifa Gu[2]

[1] Institute of Policy and Management, Chinese Academy of Sciences, Beijing 100080 P.R. China
yijunliu@casipm.ac.cn
[2] Institute of Systems Science, Academy of Mathematics and Systems Science
Chinese Academy of Sciences, Beijing 100080 P.R. China

Abstract. There are many approaches to complex problem solving, depending on the nature of the problem and the people involved in the problem. In this paper, a computational scheme based on meta-synthesis approach is applied to a social problem on exploring some issues about taxis drivers. Firstly, group argumentation environment (GAE) as a computerized tool is used to facilitate expert thinking process (meeting) about the concerned topics and form scenario or hypotheses based on qualitative meta-synthesis approach. Next, six strategies of quantitative meta-synthetic system modeling are introduced and one of that is used to construct a multi-agent model which aims to analyze the relations between taxis drivers' income and their behavior response intensity. Finally, some concluding remarks and future works are given.

Keywords: meta-synthesis, group argumentation, multi-agent system.

1 Introduction

Meta-synthesis approach (MSA) is proposed to tackle with complex, open and giant systems by Chinese scientists Qian, X.S. and his colleagues around the start of 1990s, which expects "to unite organically the expert group, data, all sorts of information, and the computer technology, and to unite scientific theory of various disciplines and human experience and knowledge" for proposing hypothesis and quantitative validating [1]. The essential idea of MSA can be simplified as "confident hypothesizing, rigorous validating", i.e. quantitative knowledge arises from qualitative understanding, which reflects the process of knowing and doing in epistemology. Later the concept of Hall of Workshop for Meta-Synthetic Engineering (HWMSE) is proposed as MSA practicing platform which is expected to utilize breaking advances in information technologies while the active roles of human beings are greatly emphasized during human-machine collaboration [2,3]. There are three kinds of meta-synthesis, 1) qualitative meta-synthesis; 2) qualitative - quantitative meta-synthesis; 3) meta-synthesis from qualitative hypothesis to quantitative validation. Each kind of meta-synthesis can be supported by various tools or methods.

In this paper, a computational scheme based on meta-synthesis approach is applied to a social problem on exploring some issues about taxis drivers. Firstly, versatile

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 9–17, 2007.
© Springer-Verlag Berlin Heidelberg 2007

computerized aids, such as visualization of expert opinion structure, clustering of contributed opinions for concept formation and automatic affinity diagramming, etc. have been developed and integrated into group argumentation environment (GAE), to facilitate expert thinking process (meeting) about the concerned topics and form scenario or hypotheses based on qualitative meta-synthesis approach. Next, six strategies of quantitative meta-synthetic system modeling are introduced and one of that is used to construct a multi-agent model based on the platform of StarLogo to analyze the relations between taxis drivers' income and their behavior response intensity. Finally, some concluding remarks and future works are given.

2 Qualitative Analysis for Scenario Forming

Qualitative meta-synthesis produces scenarios or hypotheses about the complex problems, i.e. to expose some qualitative relations or structures of the concerned problems. A variety of computerized tools, such as group support systems (GSS), creativity support systems (CSS), could be used to support idea generation which is the origin for qualitative meta-synthesis. In recent years, creative group activities have come to the fore with the growing complexity of society. One of aims of divergent thinking is for creative ideas toward unknown issues, hence demand for computerized tools to provide more help during such a thinking process. Here, a computer-based tool, group argumentation environment (GAE), which is an interactive task-oriented cooperation environment to support loosely coupled group's activities, is presented. Absorbing some ideas from AIDE [4], AA1 [5], GAE has two principal modules, i) divergence of visualizing, mainly supporting electronic brainstorming, ii) convergence of computing for the processes and results of group argumentation.

2.1 Visualized Shared Memory for Group Argumentation

Brainstorming is a technique, which involves generating a list of ideas in a creative, unstructured manner. The goal of brainstorming is to generate as many ideas as possible in a short period of time. During the brainstorming process, all ideas are recorded, and no idea is disregarded or criticized. After a long list of ideas is generated, one can go back and review the ideas to critique their value or merit. Brainstorming can encourage expert to use their imagination, tap the creativity of the group to capture unique ideas, which may be a little less conventional [4].

Electronic brainstorming is an idea-generating tool in GAE that allows participants to share ideas simultaneously on a specific topic posted to the group. Nowadays, most web-enabled forums can be regarded as simple electronic brainstorming sessions. However, only plain texts are given to all participants. By contrast, in our developed electronic brainstorming module, all collective information, mainly, utterances and keywords by each participant are processed by dual scaling method (a kind of multi-variant statistical method) and results are displayed at a 2-dimension space [6].

Such a visualized map can be understood more easily than plain texts by participants and helpful to expose some relations or structures of the concerning topic, which can also stimulate participants' further emergence of more creative ideas and hold active participation. There are three viewers in GAE.

- Common Viewer: This discussion space can be regarded as a joint thought space for the participants. All users can participant in argumentation and understand the global structure and relationships by viewing the shared discussion space.
- Personal Viewer: It is an anonymous idea-gathering space in which the relationships between utterances and keywords can be visualized.
- Retrospective Viewer: It applies same mechanism as common viewer and provides participants to "drill down" the discussing process for visualized partial perspectives.

Different from the topological graph, the graph of viewer is an interpretable graph, which reflects the data's nature in the database. But the topological structures have been designed and the forms are structured. In our research, we want to cluster the utterances and keywords of the experts and then externalize the mental process of the human thinking. Here, we think, the interpretable diagram is more suitable to embody the thinking activities than the topological graph [7].

2.2 Automatic Affinity Diagramming

The affinity diagramming (sometimes called the KJ diagram after its creator, Kawkita Jiro [8]) was developed to discover meaningful groups of ideas within a raw list. Usually it is used to refine a brainstorming into something that makes sense and can be dealt with more easily. In GAE, an affinity diagramming about the concerned topic discussed in the electronic brainstorming module is automatically produced according to 2-dimension map in the personal viewer. Not only plain text files are saved after a brainstorming session, but a document including a list of affinity sets, which reflects some structure of the concerned topic, is acquired [9].

Automatic affinity diagramming is synchronously working with the participants' discussion. Both visual map and affinity document may be helpful for participants to find some structures of the concerned issues, because visualization can offer users capabilities for self-guided exploration and visual analysis of large amounts of information. Both functions support confident hypothesis formulation during complex problem solving.

2.3 Concept Formation Based on Centroidal Clustering Algorithm

Concept formation means automatic summarizing and clustering of experts' utterances and detecting the typical keywords which represent meaningful groups or sub-communities of ideas based on visualized maps. Centroid is the center of each set produced with cluster and given by $C_m = \frac{1}{n}\sum_{i=1}^{n} t_{mi}$. Combining K-means clustering method [10], which equation is $m_i = \frac{1}{m}\sum_{j=1}^{m} t_{ij}$, k centroids, where k is an assumed number of clusters, can be detected. The closest keyword to the centroid could be regarded as cluster label.

The detailed introduction of other functions of GAE can see the reference [7,11,12].

3 Quantitative Computing Based on Meta-Synthetic Modeling

Given the scenario or hypotheses based on qualitative expert argumentation, modeling is one of the essential and effective computing methods for the complex or un-structured problem. The meta-synthesis approach emphasized during complex system modeling needs to have many kinds of views (perspectives). According to some basic principles of modeling proposed by Ackoff [13], Gu & Tang summarized the following 6 strategies of modeling which has been used to study the macro-economic problems [3, 14].

- Modeling by knowing mechanism: such as input-output models, econometric models for economic forecasting, etc.
- Modeling by analogy: also be called case reasoning or case study. When constructing some new objects or solving the complex problems, we usually consider and absorb the experiments from the existed cases.
- Modeling by rule: typical refers to multi-agent simulation (MAS), which analyzes the group's behavior with that of individuals guided by complex adaptive system theory. At present, many MAS supporting tools, such as SWARM, ASCAPE, REPAST, TNG-lab, AgentSheets and StarLogo, are already here.
- Modeling by data: such as various statistical models, reconstructability analysis (RA) model, etc.
- Modeling by evolution: such as chaos and fractal methods, etc., which helpful to investigate and explore the complexity.
- Modeling by learning: such as models based on data mining and knowledge discovery. Essentially, these models are modeling by data, but modeling by learning emphasizes the hidden structure in the massive data and knowledge, absorbs expert's experience, and improves the models based on data.

Above modeling activities may depend on hypotheses and assumptions given by qualitative meta-synthesis. Next, a whole example for some topics about taxi drivers will be demonstrated to show the computational scheme of social issues solving based on meta-synthesis approach.

4 A Whole Example of Computational Scheme for Complex Problem Solving

Taxi has been considered to be one of the most important transportation tools and takes great revenue for taxi companies, but recently, the taxi companies in China have to pay more cost because price of oil has got a continuous rising in global market. A hot discussion on rising in unit price of taxi is accordingly produced. To minimize the effect of cost increasing on public transportation system, governments of many cities, such as Beijing and Shanghai, launched a new solution which controls price of taxi directly depending on the price of oil. At present, the balance between government, taxi companies, and taxi drivers becomes social scale issue which belongs to the complex, open and giant systems.

4.1 Qualitative Scenario Forming by View of Group Argumentation

In this example, the topic for discussion is about the different strategies to deal with the taxi problems under the current situation. All opinions are summarized into 48 utterances contributed by nine persons who participate in the discussion as a community. Fig. 1 shows basic analysis taken in this test. Fig. 1(a) is a whole perspective of all concerned participants' contributions. It shows participants who share more common keywords locate closer in the 2-dimension space. Fig. 1(b) is the opinion structures of partial users as a subset community formed in retrospective viewer. Fig. 1(c) shows the affinity diagramming based on personal viewer, which divides the whole utterance set into several cells according to their locations. It could be seen that the utterances in one cell are related to each other. Automatic affinity diagramming could be regarded as a rough classification about participants' opinions during the brainstorming session. Further processing could be taken to acquire a more reasonable classification. Fig. 1(d) and (e) shows 5 clusters by centroidal clustering algorithm, by which keywords '水流', '访谈', '信息量', '定量' and '机理' are acquired as the label (centroid) of each cluster.

Combining the result of expert argumentation with the expert's background, some methods and strategies to analyze and model the taxi problems are the following:

- Using game theory to balance the benefits of the government, taxi company, taxi drivers and customs;
- Using neural network to forecast the oil price;
- Using questionnaire to acquire the taxi drivers' psychological change;
- Using GIS to grasp the distribution of the taxis in the city;
- Using MAS to simulate the taxi drivers' behavior response intensity according to the fluctuations of the taxi drivers' income.

Dynamic visualized structures of the concerned topic may reinforce the stimulation and facilitate further thinking during complex problem solving process. The evolving diagrams may also help to find some hidden structures to aid qualitative scenario or hypotheses forming which is the foundational work of quantitative modeling.

4.2 Multi-agent System for Quantitative Simulation

Here, only MAS method in StarLogo platform developed by MIT (downloaded at http://education.mit.edu/starlogo/) is used to simulate the taxi drivers' behavior response intensity according to the fluctuations of the taxi drivers' income. The only agent (taxi drivers) is categorized into 6 groups by level of incomes.

The behaviors of each agent and the conditions of the system are defined as following.

- Initial condition: each agent owns energy to afford his daily life. Here energy of each agent is defined as its income.
- Terminate condition: after a step, here meaning one day, the energy of agent will be decreased according to a const number of energy. As energy is less than the minimum cost for a daily life, agent stops movement.

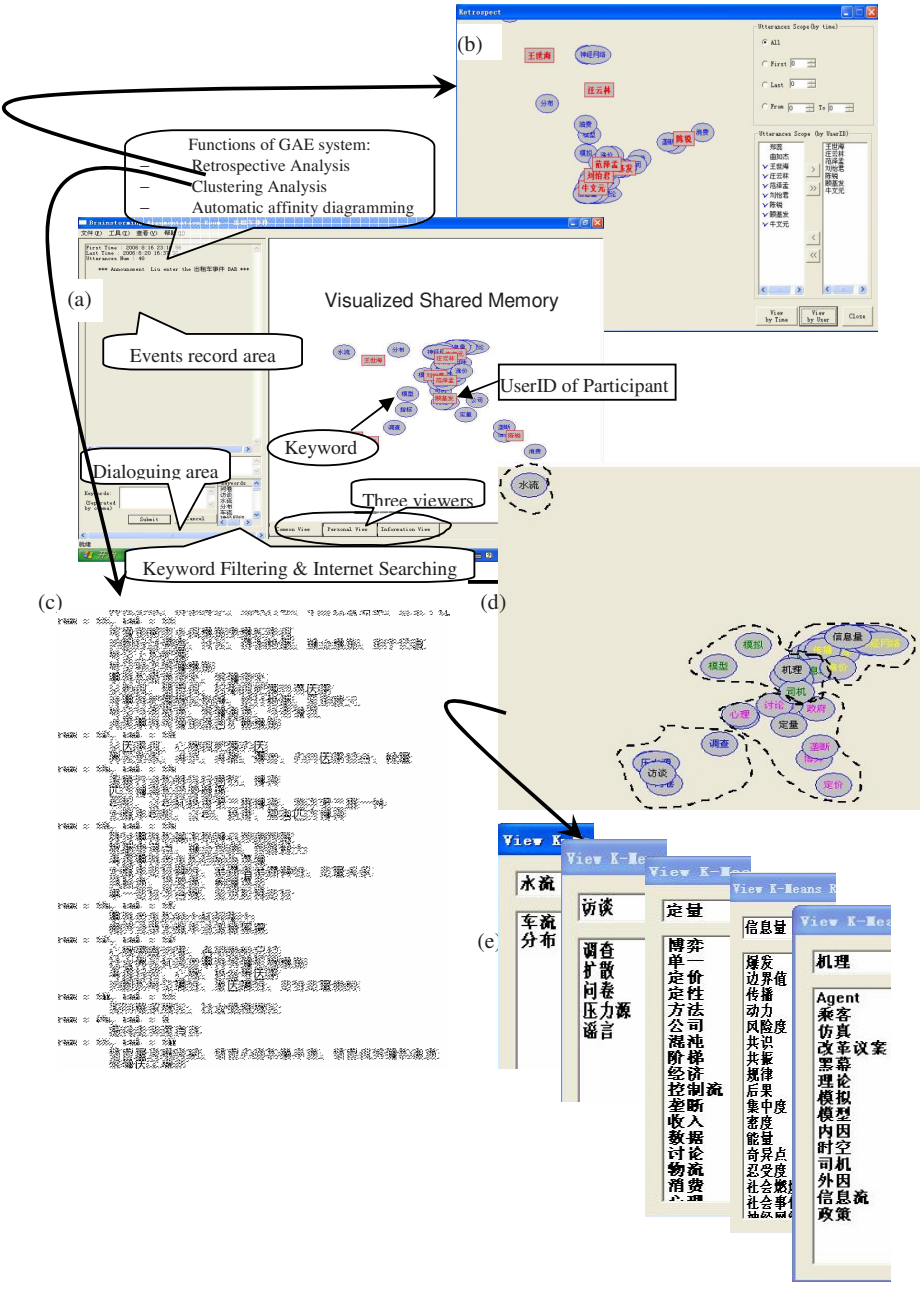


Fig. 1. Client Window of BAR. (a) Main client window. (b) Main client window. (c) Automatic affinity diagramming.(d) Clustering analysis (K=5).(e) List of a cluster of keywords by (d).

- Hypothetic condition: it is the process of modeling.
  - assuming the initial oil price equal to 5 RMB Yuan, which shows the tendency of gradually rising;
  - adjusting the unit price 1.6 RMB Yuan per kilometer to 2 RMB Yuan per kilometer, which causes the passengers reducing and the rate of unload driving increasing. We assumed that when the unit price is 1.6 RMB Yuan per kilometer, the rate of unload driving equals to 20%, which will go up to 40% when the unit price is 2 RMB Yuan per kilometer;
  - canceling the subsidy of oil price, which means the taxi drivers' income will be reduced by 670 RMB Yuan per month;
  - embodying the taxi drivers' psychological change via reducing of their income. Different income level of taxi drivers was given the different degree of energy reducing.
- Simulation experiment analysis: Agents who own the lower energy (income) will stop movement in 2 days, as shown in Fig 2(a), (b), and the middle level of agents will stop in 3 days, as Fig 2(c), (d) showed, while the high energy agents last 4 days, as Fig 2(e), (f) showed. More strictly the conditions assuming, sooner the agents stop movement.

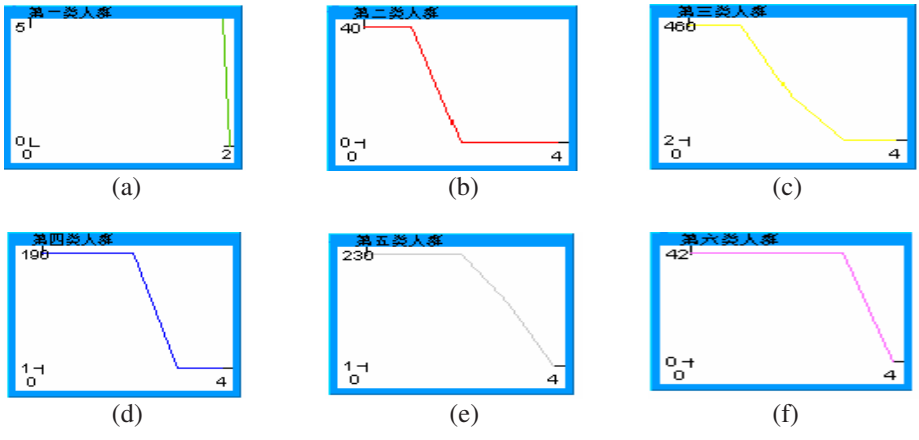


Fig. 2. Taxi Drivers' Behavior Response Intensity by MAS Simulation

Till now, the practice of meta-synthesis approach to the concerned problems is addressed. We wish the above analysis could propose an effective perspective and way for decision makers to aid them to make practical and feasible policies, such as suitable taxi price adjustments, to satisfy both the taxi companies and the taxi drivers.

5 Concluding Remarks

In this paper, we are concerned to explore a computational scheme and provide one kind of concept and some demonstrations for complex problem solving by meta-synthesis

approach. Such kind of work aims to provide different perspectives for some systemic solutions toward social issues (topics about taxi driver) unlike traditional way. Qualitative scenario or hypotheses forming for those problems through GAE is useful, advanced and unique. GAE acts as a platform promoting experts to exchange ideas, stimulating their creativity and enhancing argumentation effects. Simultaneously, meta-synthetic modeling is also a core activity used in describing and solving problems based on assumptions.

Lots of further works are under exploration, such as better human-machine interaction, evolving process of keyword network to detect the pathway of the concerned issues which can be absorbing some ideas from Chance Discovery [15], and applying more modeling methods and skills to tackle with the complex or unstructured problems, etc. More experiments and models will also be undertaken for verification and validation of the capacities of these computational scheme based on the meta-synthesis approach in practice.

Acknowledgments. This work was granted financial support from China Postdoctoral Science Foundation.

References

1. Qian, X. S., Yu J. Y., Dai R. W.: A new Discipline of Science - the Study of Open Complex Giant Systems and its Methodology, Chinese Journal of Systems Engineering & Electronics, Vol. 4, No. 2 (1993) 2-12
2. Gu, J. F., Tang,X. J.: Some Developments in the Studies of Meta-Synthesis System Approach, Journal of Systems Science and Systems Engineering, Vol.12, No.2 (2003) 171-189
3. Gu, J. F., Tang X. J.: Meta-synthesis approach to Complex System Modeling, European Journal of Operational Research, Vol.166, No.33 (2005) 597-614
4. Mase, K., Sumi, Y., Nishimoto, K.: Informal Conversation Environment for Collaborative Concept Formation. In: Ishida, T. (eds.): Community Computing: Collaboration over Global Information Networks, Wiley, New York (1998) 165-205
5. Hori, K.: A System for Aiding Creative Concept Formation. IEEE Transactions on Systems, Man and Cybernetcis, Vol.24, No. 6 (1994) 882-893
6. Nishisato S.: Analysis of categorical data Dual scaling and its applications. University of Toronto Press, Toronto (1980) 1-53
7. Liu, Y. J., Tang, X, J.: Computerized Collaborative Support for Enhancing human's Creativity for Networked Community. Proceedings of Workshop on Internet and Network Economics, Lecture Notes in Computer Science, Springer-Verlag Berlin Heidelberg, Hong Kong, (2005) 545-553
8. Ohiwa H, Takeda N, Kawai K, Shiomi A. KJeditor: a card-handing tool for creative work support. Knowledge-Based Systems, Vol.10, (1997) 43-50
9. Tang, X. J., Liu, Y. J.: A Prototype Environment for Group Argumentation. In the Proceedings of the 3rd International Symposium on Knowledge and Systems Sciences, Shanghai (2002) 252-256
10. Duda, R. O., Hart, P. E., Stork, D. G.: Pattern Classification. Wiley, New York (2001) 526-528

11. Tang, X. J., Liu, Y. J.: Exploring computerized support for group argumentation for idea generation. In: Nakamori, Y. et al. (eds.): Proceedings of the fifth International Symposium on Knowledge and Systems Sciences, Japan (2004) 296-302
12. Tang, X. J., Liu, Y. J.: Computerized Support for Qualitative Meta-synthesis as Perspective Development for Complex Problem Solving. In Adam, F. et al. (eds.): Creativity and Innovation in Decision Making and Decision Support (proceedings of IFIP WG 8.3 International Conference on Creativity and Innovation in Decision Making and Decision Support), Vol.1, Decision Support Press, London, (2006) 432-448
13. Ackoff,R L, Sasieni, M. W.: Fundamentals of Operations Research. John Wiley & Sons, 1968
14. Tang, X. J., Nie, K., Liu,Y. J.: Meta-synthesis Approach to Exploring Constructing Comprehensive Transportation System in China, Journal of Systems Science and Systems Engineering, Vol.14, No.4 (2005) 476-494
15. Ohsawa, Y.: Product Designed on Scenario Maps Using Pictorial KeyGraph. WSEAS Transaction on Information Science and Application, Vol.3 No.7 (2006) 1324-1331

ICT and Special Educational Needs: Using Meta-synthesis for Bridging the Multifaceted Divide

Ying Liu[1], Annita Cornish[1], and June Clegg[2]

[1] Homerton College, Cambridge University, Hills Road, Cambridge CB2 2PH UK
{yl317,ac550}@cam.ac.uk
[2] New Hall College, Cambridge University, Huntingdon Road, Cambridge CB3 0DF UK

Abstract. We elicit some critical views on how the applications of information communication technology (ICT) can be approached for people with special educational needs (SEN). The findings, based on a number of concrete case-studies, clearly confirm that, despite of the huge research and developmental effort having been made for advancing ICT applications in education, ICT as a whole is surprisingly lacking a systematic discourse with educational domains, even much less concerning those who have learning difficulties. Whilst well known technical tools such as word-processors, screen-reading software and problem-solving software packages still provide useful "snapshots" of ICT applications from particular points in time, we argue that these pictures now require updating. To enable the ICT development crossing multi-disciplinary boundaries, a meta-synthesis approach is innovated. The approach consists of four inter-operated components through ICT, i.e., assistive, sensory, communicative and interactive component.

Keywords: Meta-synthesis, Digital divide, e-Inclusion, Complex problem-solving, Education, Special educational needs.

1 Introduction

The term "special educational needs" (SEN) covers many kinds of significant learning difficulties that learners can have [1], relating to a range of problems from particular impairments [2], to the complex interaction between the learners and related school, ethnic, social, political, cultural factors [3][4].

Special education is therefore a field involved in a theoretical as well as empirical enquiry on all the factors that cause significant learning barriers experienced by some learners compared with other similar learners, and seeking active intervention and practical procedures that can be used to enable people with SEN better learning in a typical setting such as a school, home, social and/or virtual community [5].

To date, the results of national and international investments in ICT for education have been the major driving forces of the e-learning transformation; so have been the regional, national and international initiatives established for integrating ICT within special education. E.g., the British Educational and Communications Technology Agency (Becta) leads agencies in supporting the government's programme to develop National Grid for Learning (NGfL). It has a new and clear remit to ensure NGfL

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 18–25, 2007.
© Springer-Verlag Berlin Heidelberg 2007

developments take account for SEN and the teachers [6]. The Irish National Centre for Technology in Education published a report shows that the interest among teachers in using ICT with students with SEN is very high. The centre sets up sites for Suppliers, Section of SEN, SEN Technology, Software Central and Curriculum Resources [7].

There are varieties of commercial products for SEN as well. As easily found on the Web, the software developers claim to offer professional resources, special needs software and assistive technology, such as a) software and typing books for students with cerebral palsy, missing fingers, learning disabilities, dyslexia, visually impaired, etc.; b) learning solutions for classrooms and homes for kids, and c) accessibility software and text to speech software. However, how to evaluate such a toolkit is outside of the scope of this paper. Nevertheless this is an important research area. See, e.g. Woodward et al examined the research on software curriculum designed specifically for pupils with SEN [8].

Here the primary concern is that, ICT itself can create significant barriers for learners and teachers. Williams et al [9] conclude that the current ICT for education surprisingly lacks research into the usability of the various applications developed, even less concerning those with learning difficulties. Mooij et al conclude that the use of ICT in general merely showed characteristic of traditional approaches to learning, after interviewed 331 teachers in the highest grade of primary education. The methods employed by teachers to adapt education to the needs and abilities of individual pupils proved quite limited [10].

With the mixed views given above and more others to be discussed further, we argue, in section 2, that it is equivalently crucial to innovate a meta-synthesis approach that enables deeper understandings of multi-disciplinary uses of ICT for SEN. In section 3 we introduce the areas of need, i.e. “communication and interaction”, “sensory and/or physical”, “cognition and learning”, and “behaviour, emotional and social development”. We synthesis the uses of ICT in the first and second area in section 4 and give conclusions in section 5.

2 Bridging ICT and SEN - Converging or Accommodating the Needs?

As noted, ICT and SEN are complex research areas; Linking the two, however, can be even more difficult. A great deal of knowledge within the area of special education is mainly based on how the practitioners understand their own work. Bridging ICT and SEN should receive more attentions beyond the studies on a basis of either accepting or rejecting ICT. To do this, we use meta-synthesis as our analytical tool.

Meta-synthesis methodology has been widely employed in solving many kinds of problems: from the concern for classic engineering analysis, such as the flow of matter and/or energy, to the concern for modern-day controlling communications, such as information. Hence, for a research strategy, it forces one to look at a problem in its entirety. Meta-synthesis method is a powerful way of thinking to tackle problems of complex, open and giant systems. The method is formulated by Chinese scientists Qian, X.S. and his colleagues [12][13]; it emphasizes the synthesis of acquisitioning knowledge from human experts and integrating quantitative methods with qualitative knowledge. Later it is evolved into meta-synthetic engineering which emphasises the

discourse among research fields in studying human, society, information and communication technologies. At present, we do not seem having such a systematic discourse between ICT and SEN. Yet complex problems are emerging in SEN: a) new learning environments with different teaching styles, b) equity in access to ICT and c) an ongoing infusion of new ICT oriented opportunities.

ICT covers a large variety of technologies in any form of a product, system, network, tool, solution, method or technique that stores, retrieves, processes, manipulates, displays, and/or transmits data electronically, e.g. a PDA (personal digital assistant), handheld computer, wireless internet, mobile WWW, smart gadget, broadband, sensory device or robot. Evidently, the new digital technologies have transformed the way we live and work. This has become the reality of the information revolution. Many believe that ICT is now transforming the way we learn, along with our needs for leaning [14]. There have been earlier communication revolutions – print being a primary example. It took centuries for the printing press to impact on living, working and learning. The current digital changes are taking place only within a generation – even within a decade. When change takes place so quickly, it happens very unevenly.

Since here the major emphasis given on ICT is assisting the learners with SEN, it must be noted however, that not all impairments or disabilities create barriers to learning. However, there are conditions and impairments that are known to create the significant barriers, unless accommodations are made. Enabling special education through the uses of ICT becomes increasingly important. Such importance is clearly indicated by several existing journals in subject of "special educational technology" or even "educational technology" in general.

Up to date, an accommodation of persons with SEN is based on a geo-located institution such as a school. The accommodation relies on instruments of school systems, laws, policies, legislations, initiatives, frameworks, guidelines, trainings and teaching practice to adapt to the learners. The best instance for illustrating this approach for accommodating SEN is perhaps the UK government policy that has tended to favour models of inclusive education that promote a process of increasing participation and decreasing the exclusion of vulnerable students from the culture, curricula and communities of local schools. The Audit Commission examined how the 1981 and 1996 Special Educations were working [15]. It reported that schools were struggling. A considerable amount of difficulties still exist in enabling inclusive policies today. To this extent, inclusive policy presents an even more confused message and causes the closure of special schools and "forcing" some children into mainstream schools when it is not in their best interests to be there, resulting in distress for pupils and parents. Thus, alternative approaches, strategies or methods must be sought beyond such an accommodation. Providing opportunities to receive an education without being accountable for ensuring educational outcomes simply perpetuates inequity in a more subtle form.

Hence we argue, from a meta-synthesis point of view, that it is a *convergence* that brings multi-disciplinary resources together physically or virtually; It is such a convergence that provides *opportunities* where solutions can be found for SEN wherever the learners and teachers are placed. For instance, children with speech and language communication needs benefit from mainstream education with additional support mechanisms, especially in the early years, but not extending into secondary educations; There is evidence that the multi-method approach is promising. Researchers

report that a combination of strategies produces more powerful effects than a single strategy solution. For children with communication and interaction difficulties associated with severe and profound learning difficulties, intervention aims may vary from bringing the child's language skills up to an age equivalent level, social interaction with peers, using basic cognitive processes to develop information handling and management within the curriculum, removing obstacles to enable the child to *participate* in learning [16].

Moreover, our meta-synthesis approach views that the question is no longer how to meet individual needs at one institution, but in which environment is this person likely to learn and by what means is this person likely to get supports. In principle, ICT has the capacity to create virtual spaces and dissolve boundaries, whether between countries or between a subject, teacher and learner. But ICT itself may not make an accommodation for users. Others call this area of research as inclusive design, universal design, user adaptation, adaptive user interface, or e-inclusion. Simple examples are changing the size of the text, the colour, or adding speaking browser or special keyboards.

3 The Learning Contexts

Many researchers have also usefully reviewed ICT in SEN and provided the knowledge bridging ICT and SEN to some extent, demonstrating many useful insights of ICT applications for the teachers and the kind of leaning they want to foster [17]. However we found that it is necessary to take a more systematic approach to investigate the uses of ICT for the following three reasons.

First, we found that many ICT innovations have occurred where technology may assist people with SEN, but may not yet be explored. Product engineering is evolving from stand-alone devices and applications to distributed, connected, integrated, and multi-technology systems. Electronic products are becoming smart and software systems are becoming adaptive and personalized. The movement toward smaller, easier to use, micro-technologies, with larger-scale integration, increased performance, and reduced price not only benefits the general population, but also may have the potential to benefit those with SEN. Many exemplary ICT practices and case studies have been reported, but these developments may require to be usefully generalised.

Second, the subject of ICT in special education also refers to one of the most influential changes in our educational systems – a change process that is not only going to determine a classroom culture, a new teaching experience, a new training programme, a new policy, or a technological innovation, but also the nature of learning and the teaching process and hence the nature of coming generations of special educational systems.

Third, systematic views on current SEN in respect to learning and teaching are established. The "areas of need" as defined in the 2001 SEN Code of Practice are used as essential categories here to enable this investigation [18]; they are: Communication and Interaction, Sensory and/or Physical, Cognition and Learning, Behaviour, Emotional and Social Development. Within the each area, teaching strategies, types of the learners, as well as the ways and barriers where the teachers and learners communicate are identified. The sections below are within these contexts.

4 Converging the Needs Through ICT

So what is now needed is to follow to identify the greater ICT challenges.

4.1 For Assisting the Learners

Many utilize assistive technologies to enhance functioning in activities of daily living, control of the environment, positioning and seating, vision, hearing, recreation, mobility, reading, learning and studying math, motor aspects of writing, composition of written material, communication, and computer access. Technologies used range from low-tech devices, such as pictorial communication boards or adapted eating utensils, to high-tech devices including adapted software and voice output devices with speech synthesis.

The uses of ICT in SEN can be approached by assistive technology (AT). AT is commonly accepted as the technology designed to enable and improve access a range of services for disabled users. In contexts of SEN, AT is referred to a set of aids in a form of device, hardware, software and network system, to enable the learners to access classes, practical sessions, informal/optional study skills sessions, distance learning, libraries, learning centres, etc.

4.2 For Sensory and Physical Access

The Joint Information Systems Committee (JISC) of the UK's Higher and Further Education Funding Councils has been involved for some years in supporting those responsible for ensuring appropriate access to electronic information services for those with disabilities. Since 2001, JISC has funded the TechDis service. A list of the hardware, software and network systems that can be used within SEN contexts can be found with links in TechDis Accessibility Database [19].

To use hardware for instances, CCTV and magnifying systems can assist students to magnify library books and other documents. There are alternative in-put devices such as trackballs, touch-pads and specialist keyboards for mobility impaired students and those with dexterity difficulties who may need adaptations to a conventional mouse or keyboard. Workstations are adjustable or high enough for wheelchair users and a good selection of ergonomic items to support those with back problems, RSI and encourage safe computer use.

To use software for instances, a screen reader software enables blind and visually impaired students to work with computers, listen to spoken descriptions of text-based online content such as web sites, intranets and virtual learning environments, as long as the content has been designed to meet W3C WAI guidelines. Spell checkers, word prediction, text to speech and speech recognition can be particularly suitable for students with dyslexia and other learning difficulties. Speech recognition software can also assist those with those with mobility and dexterity difficulties. Accessible features of Windows and Office products have made many accessible options built in to Microsoft products. These can be accessed by clicking on the START button and then going to Programs > Accessories > Accessibility. These include magnifier, narrator and on-screen keyboard which are useful as temporary ways of providing access

(see, Microsoft). Validation and Repair Tools will help you to check whether your web sites and other web-based content is fully accessible for disabled students. A list of validation tools can be found on the TechDis web site [19]. Multimedia content should be captioned so that it is accessible for deaf students

In the case of provisioning electronic information service for visually impaired persons. personal computers, software and devices such as “Braille touch technology”, “sound technology”, “Window-Eyes” have been developed; The Vocal-Eyes product which provided speech output from a DOS-based screen [20], and “sight technology” are good examples [11].

4.3 For Communication and Interaction

Augmentative /Alternative Communication (AAC) includes the use of eye pointing, gesture, signing, symbols, word boards, or a speech output device. Everyone uses AAC techniques from time to time, but some people depend on them all the time [22].

Computer based AAC systems can be essential aids for SEN in contexts of communication and interaction [23]. Such a system exists to assist people to overcome communication-difficulties. The systems can contain stored communication material for people with impaired communication to retrieve and use during interaction. For example, a person who cannot speak functionally may still have some limited use of speech that allows him or her to communicate simple messages to familiar partners. This individual's speech may then be augmented by the use of gestures, a non electronic communication board, or an electronic voice-output device. The context and content of the message to be communicated would then determine which component(s) the individual used to communicate.

However, researchers also found that, as the size of the stored information corpus in the AAC system increases, the task of searching and retrieving items from that corpus will become more demanding, which will make it less easy for someone to successfully access and use such material, and thus make it more dependent on the individual's needs and skills. Thus, the full potentials of appropriate AAC systems that provide an individual with a multimodal means of communication are yet to be realised. How other children’s attitudes towards a peer who used AAC were influenced by type of AAC system/device accessed by the child for communication is also an emerging and important subject to investigate.

Personal support technologies, such as personal digital assistants (PDAs), may have the ability to aid learners with cognitive disabilities [24]. Researchers show that parents or caregivers can pre-program a PDA or desktop software with educational, vocational, or daily living tasks to prompt individuals with cognitive disabilities to perform defined vocational and independent living tasks. Reportedly, specialized PDA software is currently available for enabling individuals with developmental and other cognitive disabilities to manage personal schedules during their work tasks. PDAs may also interface with wireless communication protocols to track and monitor an individual's daily activities, and provide prompts to the individual as needed to complete educational or work tasks [25].

4.4 For Learning

The meta-synthesis enables us to specifically focus on using ICT in the learning processes.

First, ICT may assist learners with disabilities with empowerment to access to their environment and opportunities for personal development which are otherwise denied them. ICT can be a vital tool in supporting advocacy and self-advocacy for people with learning disabilities and can be the means of bringing marginalized people back into their communities.

Second, assistive technologies (see section 4.1) may include specialized training services, voice interfaces, picture-based email programs, and adapted Web browsers. Wearable intelligent devices may also assist learners. For example, a wearable data glove has been developed that translates Sign Language and transmits this information wirelessly to an electronic display.

Third, personal support technologies (see section 4.2) may benefit individuals in the classroom to remain on task, remind them of pending assignments, and provide access to information on the computer or the Internet. The effectiveness of computer-based learning techniques for students with cognitive disabilities has been documented [26].

5 Conclusion

Education should be able to reach the special educational needs of all learners. There is great ICT potential that can be explored to facilitate this challenging task. However, little has been done to develop a methodology bridging ICT and special educational needs. The lack of such a methodology can be one of the main factors that limit the current ICT in developing a systematic construction of a toolkit for a specific population. Many useful enquiries about special educational needs have been conducted by individual use-cases and specific experiments in practical educational settings. Whilst known technical tools such as word-processor, screen-reading, graphics, databases, the internet, and problem-solving software packages in a special need oriented "curriculum" still provide useful snapshots of ICT applications from particular points in time, these pictures now require updating. To do this, the meta-synthesis methodology is a powerful analytical tool for bridging the divide. In particular, the methodology enables ICT developers to think of a disability in the more functional contexts, not merely in a form of abstraction, i.e. the contexts of assistive, sensory, communicative and interactive functions in learning.

References

1. Lacey, P. and Ouvry, C. (ed.): People with Profound and Multiple Learning Disabilities – a collaborative approach to meeting complex needs. London: Fulton (2006)
2. Closs, A.:Education of Children with Medical Conditions. London: David Fulton (2000)
3. Ainscow, M.: Towards inclusive schooling. British Journal of Special Education (1997) 24 (1) 3-6
4. Davis, P., Florian, L.: Teaching Strategies and Approaches for Pupils with Special Educational Needs: A Scoping Study. DfES Research Report No.516 (2004)

5. Kauffman, J., Hallahan, D.: Special Education: What Its and Why We Need It. MA: Pearson Allyn and Bacon (2006)
6. Becta: http://www.becta.org.uk/
7. Phelan, A., Haughey, E. (ed.): Information & Advice - Special Educational Needs and Information and Communication Technology. National Centre for Technology in Education, Dublin City University, Ireland (2000)
8. Woodward, J., Gallagher, D., Rieth, H.: The Instructional Effectiveness of Technology for Students with Disabilities. In: Woodward, J. Cuban, L. (ed.): Technology, Curriculum and Professional Development: Adapting Schools to Meet the Needs of Students with Disabilities. Thousand Oaks, CA: Corwin Press (2001)
9. Williams P, Jamali H.R., Nicholas, D.: Using ICT with People with Special Education Needs: What the Literature Tells Us. In: Aslib Proceedings (2006) 58(4) 330-345
10. Mooij, T., Smeets, E.: Modelling and Supporting ICT Implementation in Secondary Schools. Computers & Education (2001) 36(3) 265-281
11. Jones, A., Tedd, L.: Provision of Electronic Information Services for the Visually Impaired: An Overview with Case Studies from Three Institutions within the University of Wales. Journal of Librarianship and Information Science, (2003) 35 (2) 105-113
12. Gu, J-F., Tan, X-J.: A Test on Meta-synthesis System Approach to Forecasting the GDP Growth Rate in China. In: Proceedings of the 47th Annual Meeting of the International Society for the Systems Sciences, at Hersonissos, Crete (2003) 6-11 July
13. Tang, X-J., Liu, Y.: Computerized Support for Qualitative Meta-synthesis as Perspective Development for Complex Problem Solving. In: International Conference on Creativity and Innovation in Decision Making and Decision Support (CIDMDS 2006) LSE London (2006)
14. Thomas, T.: Renaissance eLearning. Pfeiffer Wiley (2006)
15. Audit Commission (2002) www.audit-commission.gov.uk
16. Dee, L., Byers, R., Hayhoe, H., Maudslay, L.: Enhancing Quality of Life – Facilitating Transitions for People with Profound and Complex Learning Difficulties: a Literature Review. London: Skill/Cambridge: University of Cambridge (2002)
17. Florian, L., Hegarty, J. (ed.): ICT and Special Educational Needs – a tool for inclusion. Open University Press (2004)
18. DfES Special Educational Needs Code of Practice. London: DfES (2001)
19. TechDis (2006) http://www.niad.sussex.ac.uk/subcategory_descriptions.cfm
20. Bowman, V.: Reading between the Lines: an Evaluation of Window-Eyes Screen Reader as a Reference Tool for Teaching and Learning. Library Hi Tech. (2002) 20 (2)162-168
21. Wahl, L.: Assistive technology enhances learning for all (2006) http://www.edutopia.org/php/article.php?id=Art_1045
22. International Society for Augmentative & Alternative Communication (2006), http://www.isaac-online.org/en/home.shtml
23. Beck, A. R., Parette, P., Baley, R. L.: Multimedia Effectiveness in An AAC Pre-service Setting. Journal of Special Education Technology. (2006) 20(4) 39-49
24. Bergman, M. M.: The Benefits of a Cognitive Orthotic in Brain Injury Rehabilitation. Journal of Head Trauma Rehabilitation. (2002) 17(5) 431-445
25. McDonough, B.: Wearable tech helps disabled students. NewsFactor Network (2002) http://sci.newsfactor.com/perl/story/17419.html
26. Blischak, D. M., Schlosser, R. W. (2003) Use of Technology to Support Independent Spelling by Students with Autism. Topics in Language Disorders. (2003) 23(4) 293-304

Discovering Latent Structures: Experience with the CoIL Challenge 2000 Data Set

Nevin L. Zhang

Hong Kong University of Science and Technology, Hong Kong, China
lzhang@cse.ust.hk

Abstract. We present a case study to demonstrate the possibility of discovering complex and interesting latent structures using hierarchical latent class (HLC) models. A similar effort was made earlier [6], but that study involved only small applications with 4 or 5 observed variables. Due to recent progress in algorithm research, it is now possible to learn HLC models with dozens of observed variables. We have successfully analyzed a version the CoIL Challenge 2000 data set that consists of 42 observed variable. The model obtained consists of 22 latent variables, and its structure is intuitively appealing.

Keywords: Latent structure discovery, Bayesian networks, learning, case study.

1 Introduction

Hierarchical latent class (HLC) models [7] are tree-structured Bayesian networks where variables at leaf nodes are observed and are hence called *manifest variables*, while variables at internal nodes are hidden and hence are called *latent variables*. All variables are assumed discrete. HLC models generalize latent class (LC) models [3] and were first identified as a potentially useful class of Bayesian networks (BN) by Pearl [4].

HLC models can used for latent structure discovery. Often, observed variables are correlated because they are influenced by some common hidden causes. HLC models can be seen as hypotheses about how latent causes influence observed variables and how they are correlated among themselves. Finding an HLC model that fits data amounts to finding a latent structure that can explain data well.

The CoIL Challenge 2000 data set [5] contains information on customers of a Dutch insurance company. The data consists of 86 variables, around half of which are about ownership of various insurance products. Different product ownership variables are correlated. One who pays a high premium on one type of insurance is more likely, than those who do not, to also purchase other types of insurance. Intuitively, such correlations are due to people's (latent) attitudes toward risks. The more risk-aversion one is toward one category of risks, the more likely one is to purchase insurance products in that category. Therefore, the CoIL Challenge 2000 data set is a good testbed for latent structure discovery methods.

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 26–34, 2007.
© Springer-Verlag Berlin Heidelberg 2007

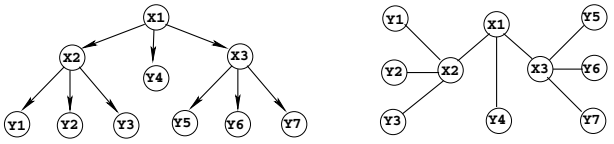


Fig. 1. An example HLC model and the corresponding unrooted HLC model. The X_i's are latent variables and the Y_j's are manifest variables.

We have analyzed the CoIL Challenge 2000 data set using HLC models. The structure of the model obtained is given in Section 4. There are 42 manifest variables and 22 latent variables, and the structure is intuitively very appealing. Latent structure discovery is very difficult. It is hence exciting to know that we are able to discover such a complex and meaningful structure.

HLC models can also be used simply for probabilistic modeling. They possess two nice properties for this purpose. First, they have low inferential complexity due to their tree structures. Second, they can model complex dependencies among the observed. In Section 5, the reader will see the implications of the second property on prediction and classification accuracy in the context of the CoIL Challenge 2000 data.

We begin with a review of HLC models in Section 2 and a description of the CoIL Challenge 2000 data set in Section 3.

2 Hirarchical Latent Class Models

Figure 1 shows an example HLC model (left diagram). A *latent class (LC) model* is an HLC model where there is only one latent node. We usually write an HLC model as a pair $M = (m, \theta)$, where θ is the collection of parameters. The first component m consists of the model structure and cardinalities of the variables. We will sometimes refer to m also as an HLC model. When it is necessary to distinguish between m and the pair (m, θ), we call m an *uninstantiated HLC model* and the pair (m, θ) an *instantiated HLC model*.

Two instantiated HLC models $M{=}(m, \theta)$ and $M'{=}(m', \theta')$ are *marginally equivalent* if they share the same manifest variables Y_1, Y_2, ..., Y_n and

$$P(Y_1, \ldots, Y_n|m, \theta) = P(Y_1, \ldots, Y_n|m', \theta'). \tag{1}$$

An uninstantiated HLC models m *includes* another m' if for any parameterization θ' of m', there exists parameterization θ of m such that (m, θ) and (m', θ') are marginally equivalent, i.e. if m can represent any distributions over the manifest variables that m' can. If m includes m' and vice versa, we say that m and m' are *marginally equivalent*. Marginally equivalent (instantiated or uninstantiated) models are *equivalent* if they have the same number of independent parameters. One cannot distinguish between equivalent models using penalized likelihood scores.

Let X_1 be the root of an HLC model m. Suppose X_2 is a child of X_1 and it is a latent node. Define another HLC model m' by reversing the arrow $X_1 \rightarrow X_2$. In m', X_2 is the root. The operation is hence called *root walking*; the root has walked from X_1 to X_2. Root walking leads to equivalent models [7]. This implies that it is impossible to determine edge orientation from data. We can learn only *unrooted HLC models*, which are HLC models with all directions on the edges dropped. Figure 1 also shows an example unrooted HLC model. An unrooted HLC model represents a class of HLC models. Members of the class are obtained by rooting the model at various nodes. From now on when we speak of HLC models we always mean unrooted HLC models unless it is explicitly stated otherwise.

Assume that there is a collection $\mathbf{D}$ of i.i.d samples on a given set of manifest variables that were generated by an unknown regular HLC model. The learning task is to reconstruct the unrooted HLC models that corresponds to the generative model.

The first principled algorithm for learning HLC models was developed by Zhang [7]. The algorithm consists of two search routines, one optimizes model structure while the other optimizes cardinalities of latent variables in a given model structure. It is hence called *double hill-climbing (DHC)*. It can deal with data sets with about one dozen manifest variables. Zhang and Kočka [8] recently proposed another algorithm called *heuristic single hill-climbing (HSHC)*. HSHC combines the two search routines of DHC into one and incorporates the idea of structural EM [2] to reduce the time spent in parameter optimization. HSHC can deal with data sets with dozens of manifest variables.

Results presented in this paper were obtained using the HSHC algorithm. The algorithm hill-climbs in the space of all unrooted regular HLC models for the given manifest variables. We assume that the BIC score is used to guide the search. The BIC score of a model m is:

$$BIC(m|\mathbf{D}) = logP(\mathbf{D}|m, \theta^*) - \frac{d(m)}{2} logN$$

where θ^* is the ML estimate of model parameters, $d(m)$ is the *standard dimension* of m, i.e. the number of independent parameters, and N is the sample size.

3 The Coil Challenge 2000 Data Set

The training set of the COIL Challenge 2000 data consists of 5,822 customer records. Each records consists of 86 attributes, containing sociodemographic information (Attributes 1-43) and insurance product ownerships (Attributes 44-86). The sociodemographic data is derived from zip codes. In previous analyses, these variables were found more or less useless. In our analysis, we include only three of them, namely Attributes 43 (purchasing power class), 5 (customer main type), and 4 (average age). All the product ownership attributes are included in the analysis.

The data was preprocessed as follows: First, similar attribute values were merged so that there are at least 30 cases for each value. Thereafter, the attributes have 2 to 9 values. In the resultant data set, there are fewer than 10

cases where Attributes 50, 60, 71 and 81 take "nonzero" values. Those attributes were therefore excluded from further analysis. This leaves us with 42 attributes.

We analyzed the data using a Java implementation HSHC algorithm. In each step of search, HSHC runs EM on only one model to optimize all its parameters. However, it may run local EM on several candidate models to optimize the parameters that are affected by search operators. The number of such candidate models is denoted by K, and K is a parameter for the algorithm. We tried four values for K, namely 1, 5, 10, and 20. The experiments were run on a Pentium 4 PC with a clock rate of 2.26 GHz. The running times and the BIC scores of the resulting models are shown in the following table. The best model was found in the case of K=10. We denote the model by M^*. The structure of the model is shown in Figure 3.[1]

K	1	5	10	20
Time (hrs)	51	99	121	169
BIC	-52,522	-51,625	-51,465	-51,592

4 Latent Structure Discovery

Did HSHC discover interesting latent structures? The answer is positive. We will explain this by examining different aspects of Model M^*. First of all, the data contains two variables for each type of insurance. For bicycle insurance, for instance, there are "contribution to bicycle insurance policies (v_{62})" and "number of bicycle insurance policies (v_{83})". HSHC introduced a latent variable for each such pair. The latent variable introduced for v_{62} and v_{83} is h_{11}, which can be interpreted as "aversion to bicycle risks". Similarly, h_{10} can be interpreted as "aversion to motorcycle risks", h_9 as "aversion to moped risks", and so on.

Consider the manifest variables below h_{12}. Besides "social security", all the other variables are related to heavy private vehicles. HSHC concluded that they are influenced by one common latent variable. This is clearly reasonable and h_{12} can be interpreted as "aversion to heavy private vehicle risks". Besides "social security", all the manifest variables below h_8 are related to private vehicles. HSHC concluded that they are influence by one common latent variable. This is reasonable and h_8 can be interpreted as "aversion to private vehicle risks".

All the manifest variables below h_{15}, except "disability", are agriculture-related; while the manifest variables below h_1 are firm-related. It is therefore reasonable for HSHC to conclude that those two groups of variables are respectively influenced by two latent variables h_1 and h_{15}, which can be interpreted as "aversion to firm risks" and "aversion to agriculture risks" respectively.

It is interesting to note that, although delivery vans and tractors are vehicles, HSHC did not conclude that they are influenced by h_8. HSHC reached the correct

[1] Note that what HSHC obtains is an unrooted HLC model. The structure of the model is visually shown as a rooted tree in Figure 3 partially for readability and partially due to the discussions of the following section.

conclusion that the decisions to buy insurance for tractors, for delivery vans, or for other private vehicles are influenced by different latent factors.

The manifest variables below h_3 intuitively belong to the same category; those below h_6 are also closely related to each other. It is therefore reasonable for HSHC to conclude that those two groups of variables are respectively influenced by latent variables h_3 and h_6.

The three sociodemographic variables (v_{04}, v_{05}, and v_{43}) are connected to latent variable h_{21}. Hence h_{21} can be viewed as a venue for summarizing information contained in those three variables. Latent variable h_0 can interpreted as "general attitude toward risks". Under this interpretation, the links between h_0 and its neighbors are all intuitively reasonable: One's general attitude toward risks should be related to one's sociodemographic status (h_{21}), and should influence one's attitudes toward specific risks (h_8, h_1, h_{15}, ..., etc).

There are also aspects of Model M^* that do not match our intuition well. For example, since there is a latent variable (h_{12}) for heavy private vehicles under h_8, we would naturally expect a latent variable for light private vehicles. But there is no such variable. Below h_3, we would expect a latent variable specifically for life insurance. Again, there is no such variable. The placement of the variables about social insurance and disability is also questionable. With an eye on improvements, we have considered a number of alterations to M^*. However, none resulted in models better than M^* in terms of BIC score.

Those mismatches are partially due to the limitations of HLC models. Disability is a concern in both agriculture and firms. We naturally would expect h_{17} (aversion to disability risks) be connected to both h_1 (aversion to firm risks) and h_{15} (aversion to agriculture risks). But that would create a loop, which is not allowed in HLC models. Hence, there is a need to study generalizations of HLC models in the future. As mentioned in Section 2, it would also be interesting to study the impact of standard model dimensions versus effective model dimensions.

5 Probabilistic Modeling

We have so far mentioned two probabilistic models for the CoIL Challenge 2000 data, namely the HLC model M^* and the latent class model produced during latent class analysis. In this section, we will denote M^* as M_{HLC} and the latent class model as M_{LC}. For the sake of comparison, we have also used the greedy equivalence search algorithm [1] to obtain a Bayesian network model that do not contain latent variables. This model will be denoted as M_{GES}. This structure of M_{GES} is shown in Figure 2. In general, we refer to Bayesian networks that do not contain latent variables *observed BN models*.

The structure of M_{HLC} is clearly more meaningful than that of M_{LC} and M_{GES}. The structure of M_{LC} is too simplistic to be informative. The relationships encoded in M_{GES} are not as interpretable as those encoded in M_{HLC}.

How do the models fit the data? Before answering this question, we note that HLC models and observed BN models both have their pros and cons when it

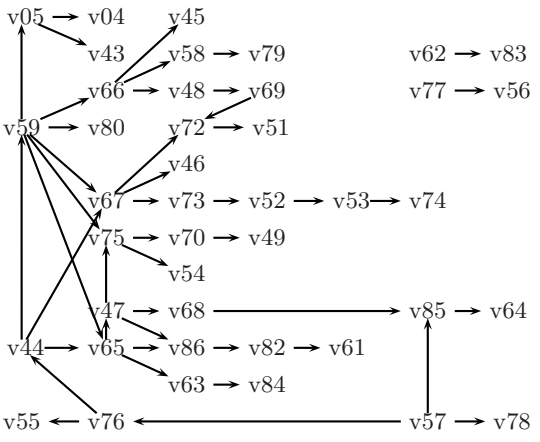


Fig. 2. Bayesian network model without latent variables

comes to represent interactions among manifest variables. The advantage of HLC models over observed BN models is that they can model high-order interactions. In M_{HLC}, latent variable h_{12} models some of the interactions among the heavy private vehicle variables; h_8 models some of the interactions among the private vehicle variables; while h_0 models some of the interactions among all manifest variables. On the other hand, observed BN models are better than HLC models in modeling details of variable interactions. In M_{GES}, the conditional probability distributions $P(v_{59}|v_{44})$ and $P(v_{67}|v_{59}, v_{44})$ contain all information about the interactions among the three variables v_{44}, v_{59}, and v_{67}.

As can be seen from the table below, the logscore of M_{HLC} on training data is slightly higher than that of M_{GES}. On the other hand, M_{GES} is less complex than M_{HLC}, and its BIC score is higher than that of M_{HLC}. In COIL Challenge 2000, there is a test set of 4,000 records. The logscore of M_{HLC} on test data is higher than that of M_{GES} and the difference is larger than that on the training data. In other words, M_{HLC} is better than M_{GES} when it comes to predicting the test data.

Model	Logscore	Complexity	BIC	Logscore (test data)
M_{LC}	-62328	739	-65532	-43248
M_{GES}	-49792	284	-51023	-34627
M_{HLC}	-49688	410	-51465	-34282

Because HLC models represent high-order variable interactions, M_{HLC} should perform better than M_{GES} in classification tasks. Out of the 4,000 customers in the COIL Challenge 2000 test data, 238 own mobile home policies (v_{86}). The

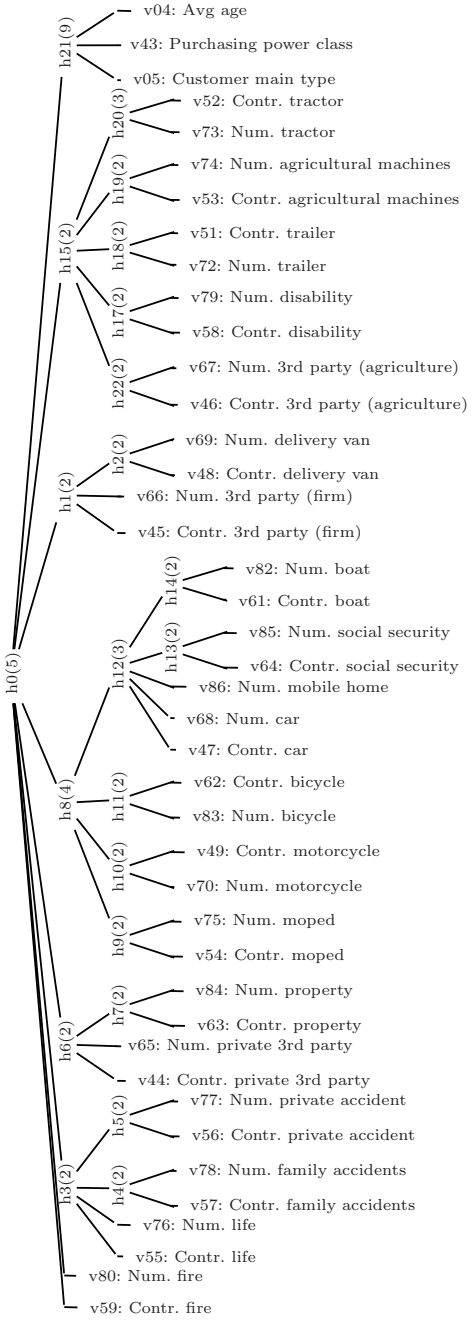


Fig. 3. Structure of Model M^*. The number next to a latent variable is the cardinality of that variable.

classification task is to identify a subset of 800 that contains as many mobile home policy owners as possible. As can be seen from the following table, M_{HLC} does perform significantly better than M_{GES}.

Model/Method	# of Mobile Home Policy Holders Identified	Hit Ratio
Random	42	17.6%
G_{GES}	83	34.9%
G_{LC}	105	44.1%
G_{HLC}	110	46.2%
CoIL 2000 Best	121	50.8%

The classification performance of M_{HLC} ranks at Number 5 among the 43 entries to the CoIL Challenge 2000 contest [5], and it is not far from the performance of the best entry. This is impressive considering that no attempt was made to minimize classification error when learning M_{HLC}. In terms of model interpretability, M_{HLC} would rank Number 1 because all the 43 entries focus on classification accuracy rather than data modeling.

6 Conclusions

Through the analysis of the CoIL Challenge 2000 data set, we have demonstrated that it is possible to infer complex and meaningful latent structures from data using HLC models.

Acknowledgements

Research on this work was supported by Hong Kong Grants Council Grant #622105. We thank Tao Chen, Yi Wang and Kin Man Poon for valuable discussions.

References

1. Chickering, D. M. (2002). Learning equivalence classes of Bayesian-network structures. *Journal of Machine Learning Research*, 2: 445-498.
2. Friedman, N. (1997). Learning belief networks in the presence of missing values and hidden variables. In *Proc. of 14th Int. Conf. on Machine Learning (ICML-97)*, 125-133.
3. Lazarsfeld, P. F., and Henry, N.W. (1968). *Latent structure analysis*. Boston: Houghton Mifflin.
4. Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference* Morgan Kaufmann Publishers, Palo Alto.
5. van der Putten, P. and van Someren, M. (2004). A Bias-Variance Analysis of a Real World Learning Problem: The CoIL Challenge 2000. *Machine Learning*, Kluwer Academic Publishers, 57, 177-195.

6. Zhang, N. L. (2002). Hierarchical Latent Class models for Cluster Analysis, AAAI-02.
7. Zhang, N. L. (2004). Hierarchical latent class models for cluster analysis. *Journal of Machine Learning Research*, 5:697–723.
8. Zhang, N. L. and Kocka, T. K. (2004). Efficient Learning of Hierarchical Latent Class Models. In *Proc. of the 16th IEEE International Conference on Tools with Artificial Intelligence (ICTAI-2004)*.

Exploration of TCM Masters Knowledge Mining

Xijin Tang, Nan Zhang, and Zheng Wang

Academy of Mathematics and Systems Science, Chinese Academy of Sciences
Beijing 100080 P.R.China
{xjtang,zhangnan,wzheng}@amss.ac.cn

Abstract. With a very long history, traditional Chinese medicine (TCM) has a rich knowledge about human health and disease by its special way. To avoid losing lots of precious knowledge of TCM masters, endeavors have been engaged to keep down those knowledge of TCM masters, such as their growth experiences, effective practical cases toward sickness and typical treating methods and principles. In this paper, some computerized methods are applied toward those collected materials about some alive TCM masters in China mainland to show a different way of exposing essential ideas of those TCM masters which aims to help people understand the correspondence of TCM views toward disease and body, and facilitate tacit knowledge transfer. This work is one kind of qualitative meta-synthesis of TCM masters' knowledge.

Keywords: Traditional Chinese medicine, knowledge mining, idea map, meta-synthesis approach.

1 Introduction

Analysis is one of salient features of all modern science. The analytical approach is the very foundation of modern medicine. Allied to the notion of analysis are the techniques of quantification and the idea of causality. Analysis is far less important to traditional Chinese medicine (TCM), which views human health and disease in terms of functional entities and disease-causing influence that are observed with the naked senses. "Its sophistication lies in its observation of correspondence between gross phenomena, and its organization of these observations through holistic systems of *yin-yang* and five *phases*"[1]. Qualitativity and holistic correspondence are two principal features in TCM whose basic concepts seem very simple while on the other hand create difficulties in applying them to practical situation. TCM diagnosis requires the identification of subtle variations of the working body and assessment of their significance in relation of each other. This is usually done by synthesis, rather than analytical reasoning. Ability to synthesize a host of subtle clues into a clear image and thus actually visualize a patient's condition is the mark of an experienced TCM physician. Due to complicated reasons, TCM is confronting difficulties in its own development in comparison to that of modern medicine. The knowledge transmission of TCM meets problems, even lots of precious knowledge of TCM masters are losing. Endeavors have been taken to save those masters' tacit experiences by systematic organizing to keep down the knowledge of those alive TCM masters, such as their growth experiences, effective practical cases toward disease, typical treating methods, principles and prescriptions. On the other

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 35–42, 2007.
© Springer-Verlag Berlin Heidelberg 2007

hand, a variety of information technologies have been applied to different facets of TCM research to find some patterns or laws. Among those, data mining, text mining and ontology are widely studied [2-9]. Those kind of research requires lots of datasets or prerequisites for mining. In this paper, instead of concerning IT applications to TCM research, the delivery of TCM masters' knowledge is firstly addressed as a knowledge conversion process where new insights may be acquired by TCM followers. To facilitate knowledge conversion, a computerized tool, TCM Master Miner, is designed to help find basic concepts or constructs of TCM masters' thoughts and applied to the meta-synthetic engineering of TCM masters' knowledge conversion. Next TCM knowledge conversion is addressed.

2 TCM Knowledge Transfer and Computerized Support

Due to the holistic correspondence feature, new TCM college graduates still require apprentice training after several year of institutionalized learning. Usually junior physicians write down prescriptions (explicit knowledge) for their mentors during daily practice for rather a period of time to gain the ability of holistic observation. That is one of biggest differences in education between TCM and modern medicine. Through learning and practice under guidance of the experienced TCM physicians, less experienced physicians may gradually sense the insights of their mentors' know-how by careful observation and practice, an indication of masters' tacit knowledge transfer to students' own knowledge, which could be regarded as a normal SECI (socialization, externalization, combination and internalization) process of knowledge conversion proposed by Nonaka and his colleague [10]. While the process of TCM knowledge conversion lasts longer. Mass production of modern medical doctors is impractical to train genuine TCM physicians.

To enable effective knowledge conversion, ideas of computerized support are naturally adopted to TCM knowledge conversion process to help less experienced physicians or even nonprofessionals to understand those TCM masters' thoughts easier, i.e. to acquire the essential framework or structure of their thoughts, especially the mechanism of qualitative correspondence in diagnosis and treatment. Such kind of supporting tools is expected to bring new threads for association and expand human's thinking space. If the disease recognition is an unstructured problem, the particular diagnosing way of TCM is a problem structuring process. Those computerized aids are expected to visualize the perspectives or structures of those TCM masters' diagnosis. It is actually one kind of qualitative meta-synthesis, i.e. to find assumptions or hypotheses about problems (syndromes) for further actions (treatment).

Among various developed supporting tools, group argumentation environment (GAE) is specifically designed to support divergent group thinking and qualitative meta-synthesis by versatile ways, such as visualization of expert opinion structure, clustering of contributed opinions, various analysis about participation, etc. [11, 12]. GAE has been applied to various conference mining, on-line group discussions of social issues, etc. However, few group activities such as conference exist in TCM practice. It is inappropriate to apply GAE directly to exploring TCM masters' thoughts. A TCM Master Miner is designed with improvements of analytical technologies in GAE.

3 TCM Master Miner for Thoughts Structuring

Current explorations by TCM Master Miner are mainly based on those materials contributed by TCM masters. One piece of thought can be expressed by a structure as <*master's name, text of thoughts, keywords set*>, which indicates that a *master* expresses his thoughts by a *text* (one sentence) with a set of *keywords*. The *keywords set* is manually selected by domain people according the related text. Based on simple representation of thoughts, a variety of explorations toward those masters are provided in TCM Master Miner, such as

- Visualization of correspondence between masters and their academic thoughts by exploratory analysis
- Clustering of masters' academic thoughts and concept extraction
- Visualization of idea structure by keyword network
- Comparisons between TCM masters, such as dominance, agreement and discrepancy, etc.

Next mechanisms of two feature functions are explained briefly.

3.1 Visualization of Correspondence Between Masters and Their Academic Thoughts by Exploratory Analysis

This is achieved by correspondence analysis using frequency matrix $F = (a_{ij})$, a_{ij} denotes the frequency of keyword j referred by the master i, $i = 1,2,\ldots,m$; $j = 1,2,\ldots,n$. The keywords are articulated as attributes of the masters.

Given frequency matrix, the mechanism of correspondence analysis is employed to explore the correspondence relations between masters and keywords; the principal components for given relations between keywords and masters could be acquired, then both masters and keywords can be mapped into 2-dimensional space. As a result, a pair of masters with more shared keywords may locate closer in the 2D space.

In TCM Master Miner, "exploratory analysis" fulfills above computing and displays the global structure of masters' collective thoughts. Such kind of analysis can be applied to any combination of available TCM masters, and may help to "drill down" into those masters' thoughts to detect some possible or emerging academic schools among those masters even they have never realized before.

As a matter of fact, the visualized map may be more useful in understanding masters' thought quickly, stimulating further thinking, such as finding interesting or strange ideas which are worth in-depth investigation, etc.

Moreover, a variety of clustering method such as *k*-means clustering can then be applied to ideas clustering and concept extraction for qualitative meta-synthesis based on the spatial relations.

3.2 Idea Viewer by Keyword Network

The clustering of the thoughts of the concerned masters by spatial correspondence provides perspectives of those masters, which is easier for novices to understand the major ideas of those masters. While above mentioned clustering is not the only way to detect structures of academic thoughts. Here is another way. Each text record of the

masters' thought has a group of keywords, which actually explain the basic constructs or ideas applied to the specific problem solving by the masters. Then a keyword graph $G_l = (K_l, E_l)$ of the lth record of the thoughts can be constructed where the vertex refers to a keyword $k_i \in K_l$ (K_l is the keyword set of the lth record), and if both keyword k_i and keyword k_j occur simultaneously in one record, then an edge exists between two vertexes $e_{ij} = (k_i, k_j), i \neq j$, $e_{ij} \in E_l$ (E_l is the edge set). Each vertex is connected with all others at one keyword graph for one piece of text. Then the aggregation of all complete keyword graphs of one master or a group of selected masters brings forward a topological keyword network, $G = (K, E)$, $K = \cup K_l$ = $\{ k_1^l, k_2^l, \cdots, k_n^l \}$, $E = \cup E_l = \cup \{e_{ij}\}$, $i, j = 1,2,\ldots,m; i \neq j$. This map is a weighted undirected network where the weight of edge refers to the frequency of co-occurrence of both keywords among all contributed texts of the master(s) and is referred as an idea map of the concerned master(s). Various network analyses can then be undertaken to detect different perspectives of the master's knowledge scope. The basic mechanism had already been discussed in Ref. [13].

Next some trials are given.

4 Practical Analysis of Some Masters' Thoughts Using TCM Master Miner

Before applying TCM Master Miner, data preprocessing is undertaken.

1) Select representative texts from each TCM master's contributing file;
2) Convert each selected text into the structure: <*master's name, text of thoughts, keywords set*> where each keyword denotes only one idea, syndrome, disease, diagnosis or treatment.
3) Put all structured records of the concerned master(s) together into a data base file;
4) Converge keywords by their synonyms based on a corpus of TCM masters' thoughts. The corpus is not a comprehensive one but growing with increasing TCM masters' materials. For example, TCM masters prefer to cite an ancient book; while sometimes they refer its author. For keyword convergence, if the keyword is the book name, then it is replaced by its author name.

With 8 TCM masters materials, some testing is undertaken here to show basic features of TCM Master Miner in exposing different perspectives about those masters' thoughts, and help to experience the holistic view in TCM thinking.

Fig.1 shows a global correspondence structure of the 8 TCM masters. It is easily to find that at the center of the map lies the keyword 内经(actually denotes the famous TCM book *Yellow Emperor's Inner Canon*) which is surrounded by names of some famous ancient TCM masters (keywords). This reveals the basic fact that those alive TCM masters mainly got basic ideas set forth in the Inner Canon written in the beginning of the first millennium and from other ancient masters. Moreover, the specialty of some masters could also be speculated, such as both 郑新 and 胡建华 are experts on stomach and spleen disease according to their surrounding keywords. Here 4 experts,



Fig. 1. Visualization of the 8 TCM masters' thought structure

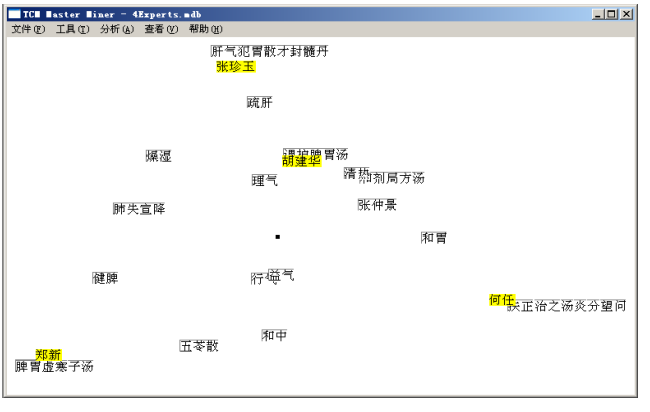


Fig. 2. Visualization of the selected 4 TCM masters' thought structure

郑新(in the center), 胡建华(below the center), 张珍玉(close to the bottom) and 何任(close to left border), are selected and their group structure is as shown in Fig.2.

The absolute location of each expert is changed in Fig.2 while the relative location of each expert still maintains, which may infer a somewhat stable joint knowledge structure of those 4 experts. Further observation indicates those 4 experts all treat stomach and spleen disease. Moreover, it could be noticed that those keywords at the center of Fig. 2 are 行气, 益气 and 理气, all related with 气(*qi*, the dynamic product of the orchestration of muscle action, or the invisible but observable force that carries food downward or upward in the digestive tract), which also exposes the treatment principles applied by

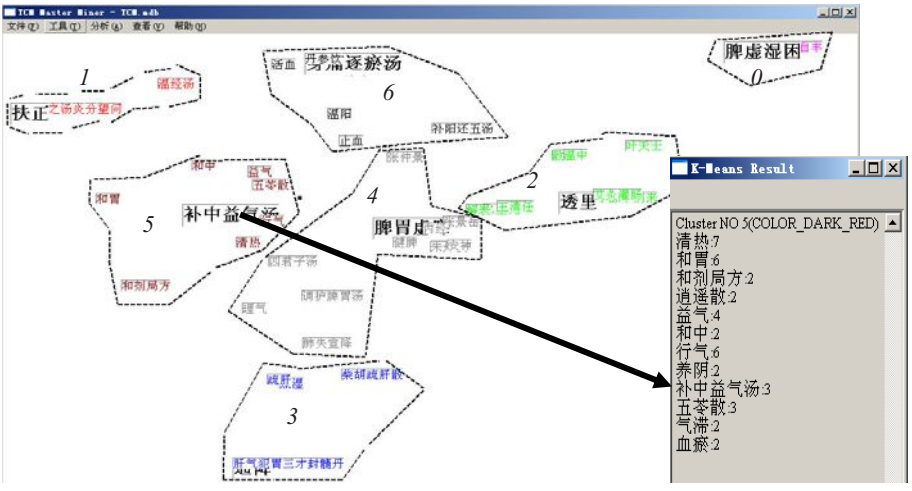


Fig. 3. Clustering of keywords of the 8 TCM masters' thought

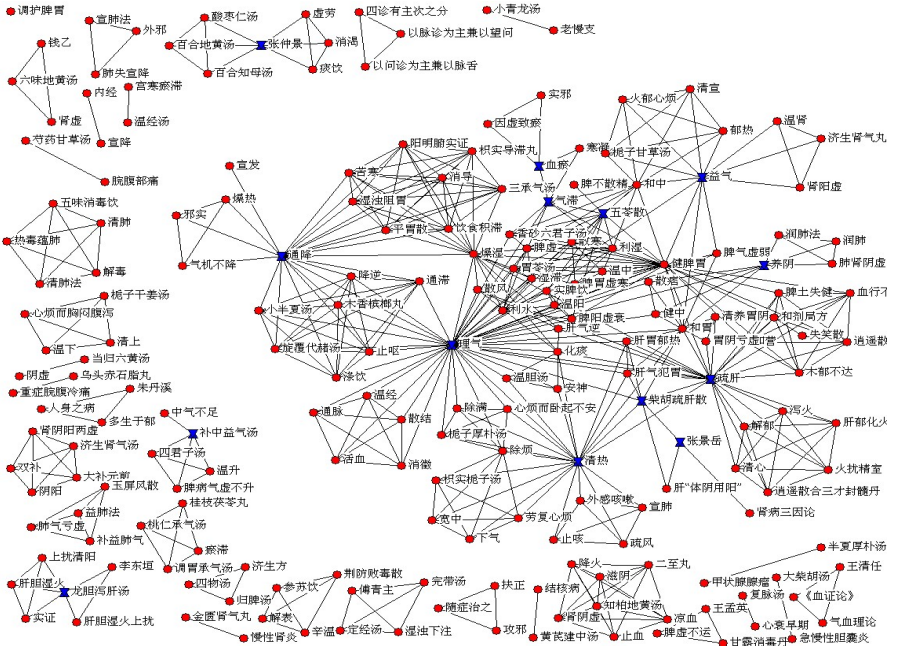


Fig. 4. Four TCM masters' thought map via keyword network (cutpoint: non circle node)

those TCM masters to stomach disease. With simple materials, basic principles of those TCM experts' thoughts are easily acquired.

With the spatial relations as shown in Fig.1, a centroid-based *k*-means clustering of keywords is undertaken. Here as $k = 7$, seven clusters are acquired as shown in Fig. 3.

The keyword (whose label is of bigger size of fonts) which is closest to the centroid of the affiliated cluster could be regarded as label of the cluster. For example, Cluster No.5 includes 12 keywords and the keyword "补中益气汤" is denoted as the representative of that cluster, which actually reflects one kind of concept extraction. Observers can check details of that cluster and define a more appropriate label.

Fig. 4 is the keyword network of those 4 selected TCM masters whose knowledge correspondence is as shown in Fig.2. Given such a network, more senses may be acquired by a variety of network analysis in detecting some features of the idea map, such as cutpoints, keyword structure of the network, etc. For example, 行气 and 益气 are two cutpoints, which may reflect their principal roles among those 4 masters. Together with Fig.2, more senses could be acquired about the major treating principles applied by those 4 experts.

5 Concluding Remarks

With long history, traditional Chinese medicine still confronts a lot of difficulties, such as the dissemination of its thoughts. This paper initially focus on modeling the TCM knowledge conversion and then proposes a computerized tool, TCM Master Miner, which aims to facilitate TCM knowledge conversion and qualitative meta-synthesis during structuring process of masters' thought.

By adopting various methods, such as correspondence analysis, graph theory and complex networks analysis, TCM Master Miner provides

- perspective analysis of TCM masters' thoughts, which help people to acquire TCM basic scenario about the working body of human beings easier;
- exploratory detection of possible academic schools of current TCM masters;
- extraction of essential TCM masters' ideas;
- awareness of unknown correspondence between different masters, between syndromes, diagnosis and treatment, etc.

TCM masters belong to TCM expert system and a variety of IT supports is regarded as machine system for quantitative computing and analysis. Both systems contribute TCM knowledge to the increasing and validating TCM knowledge system. Then those three systems construct a meta-synthetic system of TCM masters for TCM knowledge production. TCM Master Miner belongs to machine system and undertakes somewhat knowledge mining by exposing hidden structure of those TCM masters' thought and characteristics of basic TCM thinking, which may even reflect basic situations of current TCM diagnosis and treatment, and then help understand the situation of TCM in a right way.

Our current work is still at a very initial stage at both research and practice. Here shows very basic analysis provided by TCM Master Miner in exploring alive TCM famous experts' thoughts due to space limits. Lots of further work will be under exploration, such as expert group detection by considering the working location of TCM masters, consideration of more semantic meanings of keywords, such as origins or background of the thoughts, concerned patterns or syndromes, treating methods and principles or prescriptions, in correspondence analysis, etc. Besides, with more TCM masters' materials provided, more analysis will be undertaken for verification and validation of TCM Master Miner in practice.

Acknowledgments. The authors are grateful to Professor Weiliang WENG and Mr. Fei WU who provide the basic datasets and help in this study, which is supported by Natural Sciences Foundation of China under Grant No. 70571078 and the National Key Technologies R&D Program (No. 2004BA721A01H05-02).

References

1. Wiseman, N., Boss, K.: Introduction to Glossary of Chinese Medical Terms and Acupuncture Points. Library of Congress number 89-2982, Paradigm, Brookline (1995)
2. Xiang, Z.-G.: A 3-Stage Voting Algorithm for Mining Optimal Ingredient Pattern of Traditional Chinese Medicine. Journal of Software. 14(11) (2003)1882-1890
3. Sun, Y.N., et al. OLAP and Data Mining Technology in Decision Supporting System for Chinese Traditional Medical Diagnosis. Computer Engineering. 32(9) (2006) 251-252, 255 (in Chinese)
4. Liu, H.Y., Cao, Y.F., Qin, L.N.: Knowledge Acquisition Method of Traditional Chinese Medical Expert by Case Base on Ontology. Computers System and Applications. 3 (2005) 80-83 (in Chinese)
5. Yan, J.-F, Zhu, W.-F.: Apply Rough Sets Theory in TCM Syndrome Factor Diagnosis Research. Chinese Journal of Basic Medicine In Traditional Chinese Medicine.12(2) (2006) 90-93 (in Chinese)
6. Zhu, Y.-H., Zhu, W.-F.: Syndrome Differentiation System of Traditional Chinese Medicine Based on Bayesian Network. Journal of Hunan University (Natural Sciences). 33(4) (2006) 123-125 (in Chinese)
7. Cao, C.-G.: Extracting and Sharing Medical Knowledge. Journal of Computer Science and Technology. 17(3) (2002) 295-303
8. Li, C., et al.: TCMiner: A High Performance Data Mining System for Multi-dimensional Data Analysis of Traditional Chinese Medicine Prescriptions. In: Wang, S., et al. eds.: ER Workshops 2004, LNCS 3289. Springer-Verlag, Berlin Heidelberg (2004) 246-257
9. Wu, Z. H., et al.: Text Mining for Finding Functional Community of Related Genes Using TCM Knowledge. In: Boulicaut, J.-F., et al. (eds.): PKDD 2004, LNAI 3202. Springer-Verlag, Berlin Heidelberg (2004) 459-470
10. Nonaka, I., Takeuchi, Y.: Knowledge Creating Company. Oxford University Press, New York (1995)
11. Tang, X. J., Liu, Y. J.: Computerized Support for Idea Generation during Knowledge Creating Process. In: Cao, C.G., Sui, Y.F. (eds.): Knowledge Economy Meets Science and Technology (KEST'2004). Tsinghua University Press, Beijing, (2004) 81-88
12. Tang, X. J., Liu, Y. J.: Exploring computerized support for group argumentation for idea generation. In: Nakamori, Y. et al. (eds.): Proceedings of the 5th International Symposium on Knowledge and Systems Sciences (KSS'2004), Japan (2004) 296-302
13. Tang, X. J., Liu, Y. J.: Computerized Support for Qualitative Meta-synthesis as Perspective Development for Complex Problem Solving. In: Adam, F. et al. (eds.): Creativity and Innovation in Decision Making and Decision Support (proceedings of IFIP WG 8.3 International Conference, CIDMDS'2006), Vol.1. Decision Support Press, London (2006) 432-448

A Numerical Trip to Social Psychology: Long-Living States of Cognitive Dissonance

P. Gawroński and K. Kułakowski

Faculty of Physics and Applied Computer Science, AGH University of Science and Technology, al. Mickiewicza 30, 30-059 Kraków, Poland
kulakowski@novell.ftj.agh.edu.pl

Abstract. The Heider theory of cognitive dissonance in social groups, formulated recently in terms of differential equations, is generalized here for the case of asymmetric interpersonal ties. The space of initial states is penetrated by starting the time evolution several times with random initial conditions. Numerical results show the fat-tailed distribution of the time when the dissonance is removed. For small groups (N=3) we found some characteristic patterns of the long-living states. There, mutual relations of one of the pairs differ in sign.

PACS numbers: 89.65.-s, 02.50.-r.
Keywords: opinion dynamics, Heider balance, numerical calculations.

1 Introduction

The Heider theory of cognitive dissonance in social groups was formulated in 1944 [1,2] in terms of relations between triad members. A state is defined as balanced when the following four conditions are met: a friend of my friend is my friend, an enemy of my friend is my enemy, a friend of my enemy is my enemy, an enemy of my enemy is my friend. In an unbalanced state, group members suffer from the cognitive dissonance and they try to remove it. It was proven in terms of the graph theory [3] that a fully connected network is balanced if and only if it is divided into two antagonistic groups, with all relations between the groups positive and all intergroup relations negative. The question if the state of balance is ever attained remains open. It is likely that the answer does depend on the assumed dynamics. Most authors consider the model when the set of possible states is limited to a positive (friendly) and negative (hostile) one, sometimes including zero (neutral or lack of contact) [4,5,6,7]. In the time evolution, these states are changed sharply. In fact, the balanced state is attained in all investigated cases, if only the ties were present between all group members and they were properly informed on the relations in the group. The case of incomplete information was discussed in [4].

Recently, the model was reformulated in terms of a continuous change of ties, governed by differential equations [8,9,10]

$$\frac{dx_{i,j}}{dt} = g(x_{i,j}) \sum_k x(i,k)x(k,j) \tag{1}$$

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 43–50, 2007.
© Springer-Verlag Berlin Heidelberg 2007

where $x_{i,j}(t)$ is the time-dependent relation of i to j, and $g(x)$ is a function to bound the relation x within some prescribed range. Advantages of this modification are that *i)* there is no ambiguity due to the order of modified ties, and *ii)* the condescription is more realistic from the psychological point of view. Also, when this approach was applied to some commonly discussed examples ("the women of Natchez" [11] and "the Zachary karate club" [12]), the results were the same as the best obtained in literature [10]. In particular, the division of the group of 34 club members obtained theoretically was the same as observed by Zachary [12,6]. Numerical realizations of the system dynamics indicated, that there are two stages of the system. During the first – yet unbalanced – stage the relations vary slowly in an apparently incoherent way. At the end of this stage they appear to be at the edge of balance, with only some ties to be changed. Then the time evolution is accelerated in the sense that the number of unbalanced triads abruptly decreases. Once the balance is attained, the time derivative of each relation is of the same sign as this relation. Then, the absolute values of x_{ij} increase till their limits, which depend on the function $g(x)$. However, the shape of this function seems not to influence directly the first stage of the balancing process.

Here we generalize the approach to include a possible asymmetry of ties, when the relation of i to j is not necessarily the same as the relation of j to i. This asymmetry reflects the fact that the social relations are never perfectly reciprocated [13] ; therefore, our generalization reflects the actual human behaviour. We ask, if the introduced asymmetry generates any new pattern of the system behaviour. In particul1ar we are interested, if imbalanced states can persist.

2 The Calculations and the Results

Looking for generic solutions, we average the results over a set of initial values of x_{ij} selected randomly from a narrow range $(-\delta, \delta)$. The same method was applied previously [8] for the case of symmetric ties. These results suggested that in the as-yet imbalanced stage, the values x_{ij} remain small. On the other hand, the essential mechanism of removing the cognitive dissonance is captured by Eq. 1 even if $g(x_{ij})$=1 for all ties. Trying to keep things as simple as possible, we use either $g(x) = 1$, or $g(x) = 1 - (x/R)^2$ with R of the order of 10^4, whereas

Table 1. The values of the exponents α and the percentages of trajectories when $\tau > T$ against the number of nodes N

N	α	the percent of unbalanced triads
3	+2.07	0.08
4	+1.83	4.21
5	+4.4	0.43
6	+3.7	5.43
7	$\approx$ +5	19.51

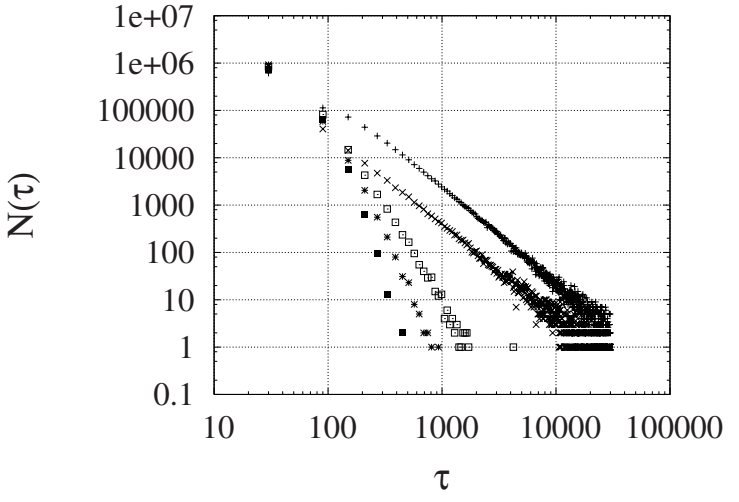


Fig. 1. The histograms of the time τ when the balance is attained, for various numbers N of nodes: $N = 3, 4, 6, 5, 7$ from right to left

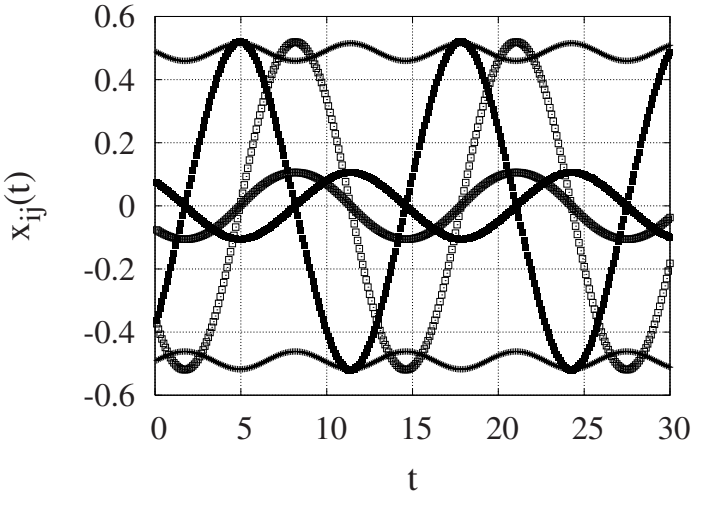


Fig. 2. A long-living behaviour of trajectories x_{ij} for $N = 3$. There, x_{23} and x_{32} preserve their opposite signs. Four other x's oscillate, two with the same phase and two with opposite phases.

$\delta = 0.5$. Both choices practically neglect $g(x)$ as long as the evolution of x remains stationary.

Numerical simulations of the time evolution governed by Eq. 1 are performed for N=3,4,5,6 and 7 nodes of a fully connected network, within a given time period $T = 3 \times 10^4$. The diagonal elements of the matrix x_{ii} are kept to be zero. Once the number of imbalanced triads falls to zero, the calculation is stopped.

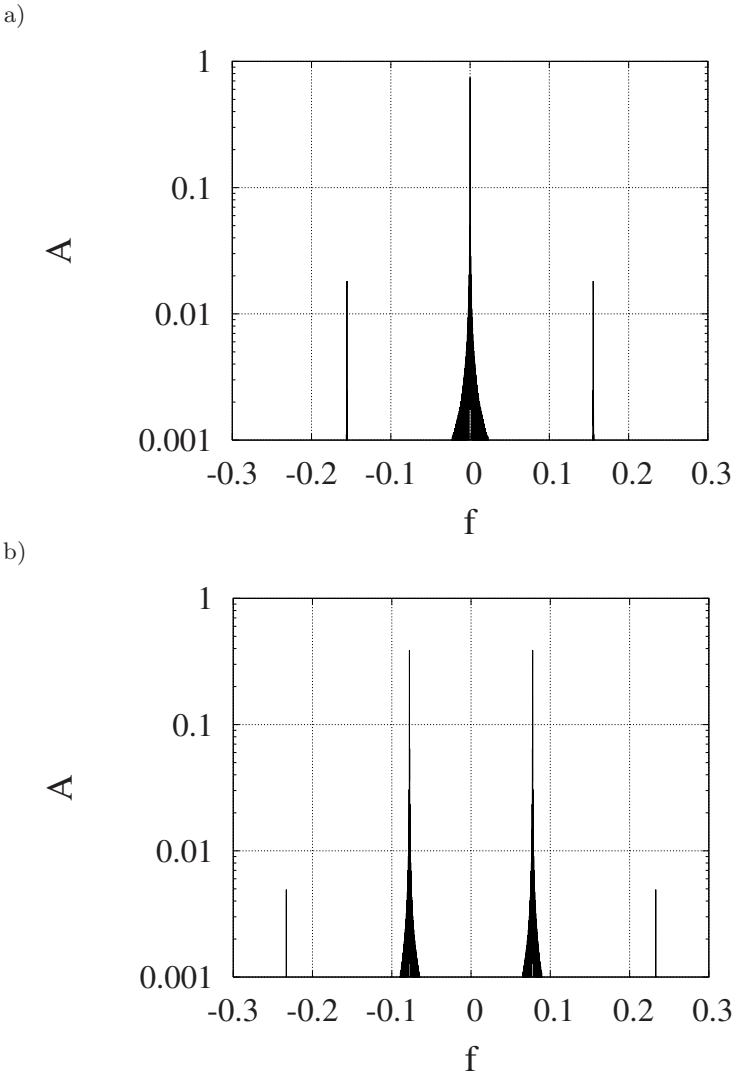


Fig. 3. Two Fourier spectra for long-living states when $N = 3$; one encountered for x_{23} and for x_{32}, and the other – for the remanining four ties

As for our experience, once the system is balanced it remains balanced. The percentage of cases, when the balance is not attained within time T, is given in Table 1 for various N. The total number if simulated trajectories is 10^6 for all values of N.

In Fig. 1 we show the probability distribution of the time τ when the balance is attained. The obtained curves indicate that the distribution functions decrease as power functions $\rho(\tau) \propto \tau^{-\alpha}$, with the exponent α dependent on the number N of nodes. The values of the exponents are given in Table 1. The percentage

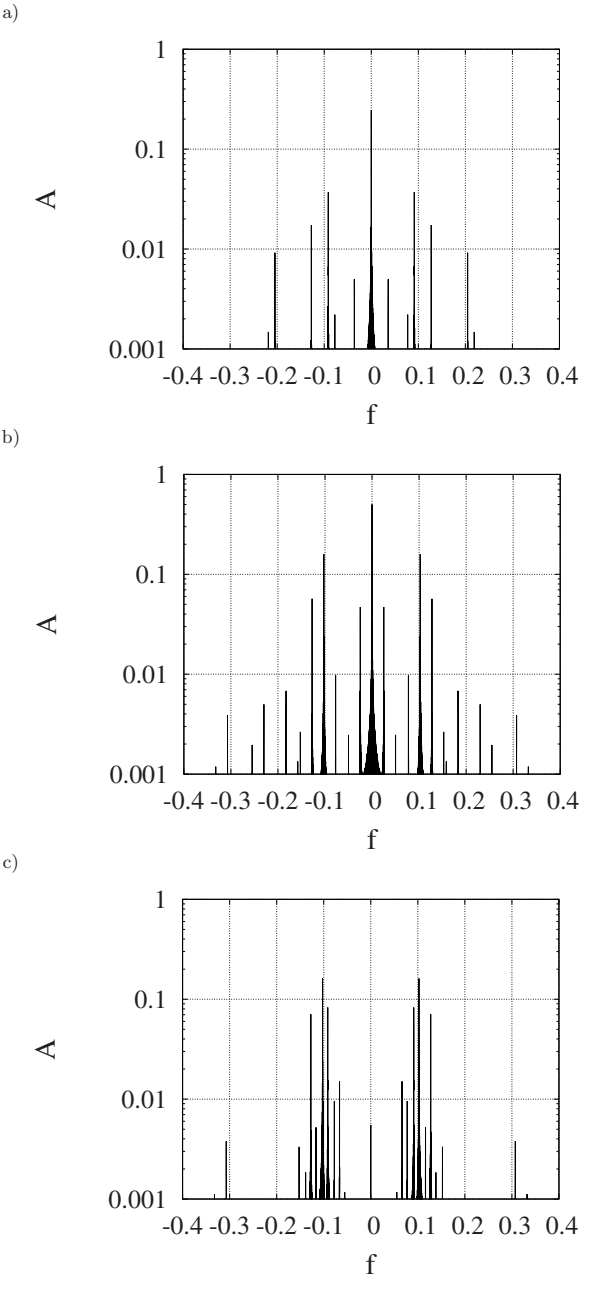


Fig. 4. Three Fourier spectra for long-living states when $N = 4$; one encountered for x_{12}, x_{21}, x_{34} and x_{43}, one for x_{13}, x_{31}, x_{24} and x_{42}, and the other – for the remanining four ties

of the cases when $\tau < T$ oscillates with N but decreases for $N = 7$; we deduce that the power distribution appears only for small N.

Searching for typical trajectories with large τ, we noticed that indeed some characteristical patterns appear which seem stationary or close to stationary. As the number of curves $x_{ij}(t)$ to be investigated depends on N as $N(N-1)$, it is easier to investigate the case when $N = 3$. Often, large τ is produced where the curves are similar to those in Fig. 2. There is also a regularity in the Fourier transforms of the selected curves. As a rule, the spectra can be divided in two groups, and those within a group are practically the same. Two spectra belong to one group, and four spectra – to the other. These two belong to the ties which join the same nodes, e.g. x_{12} and x_{21}. The spectra are shown in Fig. 3. Similar situation is found for $N = 4$. In this case, there are three different patterns of the Fourier spectra, as shown in Fig. 4. As a rule, ties with similar spectra join different nodes, e.g. x_{12} and x_{34} belong to the same pattern.

In the simplest case of $N = 3$, these numerical results suggest an approximate analytical solution of Eq. 1. Let us suppose that $x_{23} = -x_{32} = \omega$, $x_{ij} = a_{ij}\epsilon\cos(\omega t)$ for the tie $i, j = 1, 2$ and $x_{ij} = a_{ij}\epsilon\sin(\omega t)$ for the tie $i, j = 1, 3$, where ϵ is a small parameter. Substituting this to Eq. 1, we get that $a_{12} = a_{13}$ and $a_{21} = -a_{31}$. The corrections to x_{23} and x_{32} are of the order of ϵ^2, and to the other $x's$ – of ϵ^3. We can deduce that either $sign(a_{12}) = sign(a_{21})$ and then $sign(a_{13}) = -sign(a_{31})$, or $sign(a_{12}) = -sign(a_{21})$ and then $sign(a_{13}) = sign(a_{31})$. The numerical results confirm these sign rules. Also, the Fourier spectra show that the frequency of the corrections to x_{23} and to x_{32} are characterized by the frequency two times larger than the basic frequency characterizing other $x's$. This rule comes directly from the above parametrization, as $\dot{x}_{23} \propto \sin(2\omega t)$.

Some insight is possible also for $N = 4$, where the pattern of the time evolution is the same for x_{ij} and x_{ji}. In this case, all six ties are divided to three pairs, e.g. $(1, 2) + (3, 4)$, $(1, 3) + (2, 4)$, $(1, 4) + (2, 3)$. Suppose we denote these sets as a, b and c. Then the time evolution of ties a are governed by the product of elements of b and elements of c. Writing it as $\dot{a} = bc$, we have also $\dot{b} = ca$ and $\dot{c} = ab$. This regularity is observed in the simulations, and Fig. 4 is typical in this sense.

3 Discussion

As follows from Table 1, the exponent α is not universal, as it depends on N. However, as can be deduced from Fig. 1, there is no characteristic scale of time at least for small number of nodes N. Translating this result to the social psychology, the time τ of removal the cognitive dissonance can be, at least in principle, arbitrarily long. We note however that our tool – the numerical simulation – does not allow us to state that in some cases, the balance is not attained. On the other hand, the analysis of the trajectories indicates, that the large values of τ are due to one ($N = 3$) or two ($N = 4$) pairs of ties which are permanently of different sign. This rule could be written symbolically as an old exercise for beginneers in Latin: *Agricola amat puellam. Puella non amat agricolam* [14]. The lack of reciprocity seems

to create the lack of balance; other ties oscillate around zero and therefore they are not able to change the situation. As a consequence, some stable or metastable patterns appear. Their persistence depends on N, and it seems to be relatively weak for $N = 5$. We deduce it from the fact that in this case the observed times τ are relatively short. This can be due to topological properties of the fully connected graph of $N = 5$ nodes. The result suggests, that the symmetries like those discussed above for $N = 3, 4$ cannot be preserved for $N = 5$.

Trying to draw some conclusion for the social psychology, where the original concept of the Heider balance has been formulated, we can refer to some attempts to interpret within the Heider model examples drawn from history or literature. In Refs. [3,15], a fictitious situation is analysed with four persons: Hero, Blackheart, Buddy and Goodman. In the final balanced state Blackheart proved to the remaining three that they should act together. The asymmetry which maintained the evolution was that Buddy liked Blackheart. Non-reciprocated love as a motif is able to keep the action unsolved and reader's attention vivid for long time [16,17]. To kill the murderer of his father, Hamlet had to destroy Ophelia's love, as he was not able to implicate her in the conspiracy [18]. More generally, the unbalanced state is of interest as opposite to an open conflict. It is known that to activate enmity, one has to kill a commonly accepted person, as Mohandas Gandhi, Martin Luther King or Yitzhak Rabin. This method makes the situation clear for warriors. The case even more fraught with consequences – international relationships in Europe from 1872 to 1907– was mentioned in the context of the Heider balance in Ref. [19]. In fact, Otto von Bismarck maintained equilibrium by a set of bilateral relations, binding hostile states [20]. Last but not least, the message *Love Thy Enemies* [21] can be interpreted as a desperate attempt to prevent the hate from spreading. These examples suggest that in conflict preventing, some asymmetric ties can be essential.

References

1. F. Heider, *Social perception and phenomenal causality*, Psychol. Rev. **51** (1944) 358-74.
2. F. Heider, *The Psychology of Interpersonal Relations*, John Wiley and Sons, New York 1958.
3. F. Harary, R. Z. Norman and D. Cartwright, *Structural Models: An Introduction to the Theory of Directed Graphs*, John Wiley and Sons, New York 1965.
4. N. P. Hummon and P. Doreian, *Some dynamics of social balance processes: bringing Heider back into balance theory*, Social Networks **25** (2003) 17-49.
5. Z. Wang and W. Thorngate, *Sentiment and social mitosis: implications of Heider's balance theory*, Journal of Artificial Societies and Social Simulation vol. 6, no. 3 (2003) (http://jass.soc.surrey.ac.uk/6/3/2.html)
6. M. E. J. Newman and M. Girvan, *Finding and evaluating community structure in networks*, Phys. Rev. E **69** (2004) 026113.
7. T. Antal, P. L. Krapivsky and S. Redner, *Dynamics of social balance of networks*, Phys. Rev. E **72** (2005) 036121.
8. K. Kułakowski, P. Gawroński and P. Gronek, *The Heider balance - a continuous approach*, Int. J. Mod. Phys. C **16** (2005) 707.

9. P. Gawroński, P. Gronek and K. Kułakowski, *The Heider balance and social distance*, Acta Phys. Pol. B **36** (2005) 2549-58.
10. P. Gawroński and K. Kułakowski, *Heider balance in human networks*, AIP Conf. Proc. **779** (2005) 93-5.
11. L. C. Freeman, *Finding Social Groups: A Meta-Analysis of the Southern Women Data*, in R. Breiger, K. Carley and P. Pattison (eds.): *Dynamic Social Network Modeling and Analysis*, The National Academies Press, Washington 2003.
12. W. W. Zachary, *An information flow model for conflict and fission in small groups*, J. Anthropological Research **33** (1977) 452-73.
13. P. Doreian, R. Kapuscinski, D. Krackhardt and J. Szczypula, *A brief history of balance through time*, J. Math. Sociology **21** (1996) 113-131.
14. The farmer likes the girl. The girl does not like the farmer.
15. P. Doreian and A. Mrvar, *A partitioning approach to structural balance*, Social Networks **18** (1996) 149-168.
16. M. de Cervantes Saavedra, *Don Quixote*, London 1885 (http://www.donquixote.com/english.html).
17. E. Rostand, *Cyrano de Bergerac*, http://www.gutenberg.net/etext/1254.
18. W. Shakespeare, *Hamlet, Prince of Denmark*, Oxford UP, London 1914.
19. T. Antal, P. L. Krapivsky and S. Redner, *Social balance of networks: the dynamics of friendship and enmity*, presented at Dynamics on Complex Networks and Applications, Dresden, Germany, February 2006 "Physica D 224 (2006), 130-136".
20. N. Davies, *Europe. A History*, Oxford UP, New York 1996.
21. *The Bible*, Matt. 5:44.

A Hidden Pattern Discovery and Meta-synthesis of Preference Adjustment in Group Decision-Making*

Huizhang Shen[1], Jidi Zhao[1], and Huanchen Wang[2]

[1] Department of Information Systems, Shanghai Jiao Tong University, 535 Fahuazhen Rd., Shanghai, China 200052
{Hzshen,Judyzhao33}@sjtu.edu.cn
[2] Institute of System Engineering, Shanghai Jiao Tong University, 535 Fahuazhen Rd., Shanghai, China 200052

Abstract. Two aspects of group decision-making (GDM) have received much attention since its appearance, one is the organizing process of GDM which emphasizes in particular on behavioral science and qualitative analysis, and the other is the weight allocation and meta-synthesis of group preference focusing on the quantificational computation. Despite the abundant researches, the existing solutions do not take into account the dynamic change of group members' decision-making preference in the GDM process. This paper considers the preference change in GDM as a dynamical procedure, researches into the important hidden pattern in GDM process, and puts forward a GDM meta-synthesis method based on Markov Chain to mine the hidden information with both qualitative analysis and quantificational computations.

Keywords: group decision-making, hidden pattern, meta-synthesis.

1 Introduction

Based on the construction of group preference, group decision-making is a procedure [1] of synthesizing the preferences of each decision-maker in the group and sorting the schemes or choosing the best scheme in the scheme set. One of the key points in group decision-making is the decision-makers' preference relations. As Arrow pointed, the preference relation formed by group decision should satisfy five rational terms [1] which are the preference axiom, impossible axiom, completeness, Pareto optimization, and non-autarchy. Notable scholars such as Arrow [1], Dyer [2], Keeney [3] and French [4] established the theoretical foundation about group decision preference relation analysis in the group decision-making research. From their research, we know that the group preference is a function of individual preferences in the issue of the group decision-making. Preference is a term quoted from economics. In the group decision-making problems, it is used to represent the decision-makers' partiality on value [2]. There has been much research on the meta-synthetic problem. The metasynthetic algorithms accepted widely include Weighted Average Method, Bordly Multiplication [5], Bayesian Integration Method [3], Entropy Method [4],

* This research is supported by NSFC(0 0).

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 51–58, 2007.
© Springer-Verlag Berlin Heidelberg 2007

Fuzzy Cluster Analysis [2], and so on. The dynamics of the group aggregation of individual decisions has also been a subject of central importance in decision theory [6]. In order to narrow the gap between the collective opinion and each decision maker's opinion, Fan et al [8] proposed a linear goal programming model to solve group decision-making (GDM) problems where the preference information on alternatives provided by decision makers is represented in two different formats, i.e. multiplicative preference relations and fuzzy preference relations. Hochbaum and Levin [6] present a model to quantify the degree of preference which allows for flexibility in decision protocols and can take into consideration imprecise beliefs. Based on fuzzy linguistic preference relations, Herrera-Viedma et al [8] present a model of consensus support system to assist the experts in the consensus reaching process of group decision-making problems.

Despite the abundant research on group decision-making problems, existing solutions do not consider the preference adjustment as a continuous procedure and ignore the effect of the dynamic change procedure of group members' decision-making preference on group decision-making process. The procedure of forming individual preference is a decision-maker's meta-synthetic thinking procedure of perceiving all the information relating to expectation, information, sensibility, creativity, and so on, thus it is a extremely complex procedure [5]. Generally speaking, the decision-making group's preference on the scheme set will change as the decision-makers adjust their preferences under the same decision-making rules. We call it the preference convergence procedure which can be described as follows. Suppose that every member of the decision-making group has the will of coming to a unified decision. Each decision-maker in the group G presents her own preference based on the utility maximization rules. The r^{th} decision-maker DM_r will adjust and readjust her reference and feed it back to the group of decision-makers after he cognizes the preferences of the other $(l-1)$ decision-makers. This procedure of cognizing and feedback goes with each member in the group repeatedly and continuously. Such a procedure usually repeats finite times, for example, there are usually five rounds in Delphi. In the group decision-making procedure, communication among decision-makers is encouraged to increase their comprehension on the decision-making problem, on the available information and even on themselves. The decision makers are also encouraged to continuously adjust their preferences until they come to consensus on the group decision-making.

Thus in this paper, we focus on the research of how to achieve consensus quickly and reliably in group decision-making process. We consider the preference change in GDM as a dynamical procedure, research into the important hidden pattern in GDM process, study the meta-synthetic procedure of group preference based on the decision-makers' preference utility value, and put forward a meta-synthesis approach based on Markov chain for group decision-making problems.

2 Group Decision-Making and A Hidden Pattern of Group Preference Adjustment

In traditional meta-synthetic approaches on group decision-making, once the weights for every decision-maker are fixed on, the group preference value on each scheme is

determined by the sum of each decision-maker's weight multiplied by her *current preference value* of the scheme. Then all schemes are sorted in the order of their preference values. For example, if the order is $x^2 Rx^1 Rx^3 Rx^4$, the scheme x^2 will be the group's preferred scheme. Nevertheless, here a decision-maker's *current preference value* vector in the scheme set is a transient decision instead of a stationary one. That is, the decision-maker may have changed her preference value from a value into the current one in this round, and she may continue to change her preference value in the rounds hereafter. Therefore, in traditional meta-synthetic approaches, ignoring the dynamic procedure of decision-maker's preference adjustment and decision-maker's hereafter *current preference values* may lose important information in group decision-making. As the group decision-making procedure continues, the *current preference value*s are supposed to change along the time. Thus we consider the dynamic procedure of decision-maker's preference adjustment and its effect on group decision-making in this paper.

2.1 Discrete Time Markov Chains

A sequence of random variables $\{ E_n \}$ is called a Markov chain if it has the Markov property:

$$\begin{aligned} &T\{E_{n+1} = j \mid E_n = i, E_{n-1} = i_{n-1}, ..., E_0 = i_0\} = T\{E_{n+1} = j \mid E_n = i\} \\ &T_{ij} = T\{E_{n+1} = j \mid E_n = i\} \end{aligned} \tag{1}$$

Here, E_i is an event and T_{ij} is the probability to transit from state i to state j of the event. The property is called Memoryless. In other words, "Future" is independent of "Past" given "Present". Here the transition probabilities T_{ij} satisfy $T_{ij} \geq 0, \quad \sum_{j=0}^{\infty} T_{ij} = 1.$

The Chapman-Kolmogorov equation for a discrete-time Markov chain is as follows, if the distribution at "time" t_n is $\pi^{(n)}$, then the distribution at "time" t_{n+1} is given by

$$\pi^{(n+1)} = \pi^{(n)} T \tag{2}$$

2.2 Hidden Pattern of Group Preference Change Based on Markov Chain and the Construct of Markov State Transition Matrix for GDM

After the t rounds adjustment, the preference utility values in all the rounds for decision-maker DM_r are

$$\pi_r = \begin{Bmatrix} \pi_r^1(x^1) & \pi_r^1(x^2) & \cdots & \pi_r^1(x^s) \\ \pi_r^2(x^1) & \pi_r^2(x^2) & \cdots & \pi_r^2(x^s) \\ \cdots & \cdots & \cdots & \cdots \\ \pi_r^t(x^1) & \pi_r^t(x^2) & \cdots & \pi_r^t(x^s) \end{Bmatrix} \tag{3}$$

In this matrix, each row stands for the preference utility value vector in each round. Comparing the k^{th} row with the $(k+1)^{th}$ row ($\{k = 1, 2, \cdots, t\}$), if there exists $\pi_r^{k+1}(x^i) \downarrow \Leftrightarrow \pi_r^{k+1}(x^j) \uparrow$, we set the state variable $E_{ij} = E_{ij} + 1$, which shows that the decision-maker has ever changed her preference from the scheme x^i to x^j. For each decision-maker, there are at most $t-1$ times of adjustment. Packing all the adjustment for the group together, we have

$$T_r = \begin{bmatrix} 1 - \sum_{j \neq 1} \frac{E_{1j}}{E_r} & \frac{E_{12}}{E_r} & \cdots\cdots & \frac{E_{1s}}{E_r} \\ \frac{E_{21}}{E_r} & 1 - \sum_{j \neq 2} \frac{E_{2j}}{E_r} & \cdots\cdots & \frac{E_{2s}}{E_r} \\ \cdots \quad \cdots \quad \cdots & \cdots \quad \cdots & \cdots & \cdots \\ \frac{E_{s1}}{E_r} & \frac{E_{s2}}{E_r} & \cdots\cdots & 1 - \sum_{j \neq s} \frac{E_{sj}}{E_r} \end{bmatrix} \tag{4}$$

where T_r is the preference state transition matrix for decision-maker DM_r, E_{ij} denotes the preference transition times from x^i to x^j and $E_r = t - 1$ is the sample space for the state transition times.

For example, the preference utility value matrix for decision-maker DM_r is

$$\Lambda_r = \begin{bmatrix} 0.1 & 0.3 & 0.2 & 0.4 \\ 0.2 & 0.2 & 0.3 & 0.3 \\ 0.2 & 0.3 & 0.2 & 0.3 \\ 0.2 & 0.4 & 0.2 & 0.2 \\ 0.3 & 0.3 & 0.2 & 0.2 \end{bmatrix}$$

The first row of the matrix is the initial value and the sample space is $t - 1 = 5 - 1 = 4$.

Comparing the second row with the first row, we have $x^2 \to x^1, x^4 \to x^3$.

Comparing the third row with the second, we have $x^3 \to x^2$.

Comparing the fourth row with the third, we have $x^4 \to x^2$.

And, comparing the fifth row with the fourth, we have $x^2 \to x^1$.

According to equation (4), we have the preference state transition matrix T_r for decision-maker DM_r is

$$T_r = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0.5 & 0.5 & 0 & 0 \\ 0 & 0.25 & 0.75 & 0 \\ 0 & 0.25 & 0.25 & 0.5 \end{bmatrix}$$

In this matrix, $\frac{E_{11}}{E_r} = 1$ shows that the decision-maker never changes her preference on x^1.

Define the overall state transition matrix of the decision-making group in the t rounds adjustment procedure as

$$T = \frac{1}{l}\sum_{r=1}^{l} T_r \tag{5}$$

2.3 Markov Property of the Group Preference Adjustment Procedure

Because each decision-maker in the group independently puts forward her preference judgment matrix, the preference state E_n^r of decision-maker DM_r is independent of other decision-makers and the future preference state E_{n+1}^r is independent of other states except the current state E_n^r, thus the group decision-making procedure satisfies equation (1).

Obviously, the transition probabilities T_{ij} constructed from equation (4) satisfy $T_{ij} \geq 0, \quad \sum_{j=0}^{\infty} T_{ij} = 1$.

Equation (5) shows that the overall state transition probabilities matrix is the mean value of the matrices of transition probabilities of each decision-maker, thus the group property is implied in the individual properties.

Therefore, we can use the Chapman-Kolmogorov equations, equation (2), to get $\pi^{(n+1)}$ at "time" t_{n+1} from $\pi^{(n)}$ at "time" t_n.

3 Implementation of Markov-Based Meta-synthesis Approach for Group Decision-Making

A weight allocation method in group decision-making, based merely on the analysis of the objective individual preferences in the scheme set without man-made subjective

factors and taking into account both the individuals' carefulness and extremeness in the group, is presented in [9]. Shen et al [10] presented a clustering algorithm of group decision-making preference. The weight allocation method and the clustering algorithm will be introduced into the group decision-making meta-synthesis approach, which is based on Markov chain and works as follows.

(1) Publish the group decision-making problem and its background to each group member, including the event, the available data and information, the constraints, the schemes, the decision-making rules, the user handbook for GDSS, etc.

(2) Each decision-maker gives her preference judgments between each two schemes in the scheme set. All the decision-makers can publish their opinions, evidences and explanations on the message board to support their points of view.

(3) Based on the preference judgment matrix, the GDSS automatically generates the preference utility values for a decision-maker DM_r in the t^{th} round, $\left\{\pi_r^t(x^1), \pi_r^t(x^2), \cdots, \pi_r^t(x^s)\right\}$. As stated above, the individual preference adjustment in group decision-making is a continuous procedure in which the decision-makers adjust their preference in each round respectively based on the communications among the group members. The continuous adjustments make the group decisions come to converge gradually. The preference utility values vector $\left\{\pi_r^t(x^1), \pi_r^t(x^2), \cdots, \pi_r^t(x^s)\right\}$ for each decision-maker is preserved for constructing the Markov state transition matrix in step (8).

(4) Packing all the preference values got in step (3), the system has the preference utility values matrix for the t^{th} round, automatically computes the preference distance d_{ij} between decision-maker DM_i and DM_j on the scheme set, constructs the preference difference matrix D, clusters the preference difference matrix D using the algorithm given in [10] and displays the clustering results on the message board.

(5) The decision-makers in the group adjust their preferences using the clustering information given by the system. Go back to step (2) and begin another round of preference judgment.

(6) Repeat through step (2) to step (5) for $t = 7 \pm 2$ times. The choice of $t = 7 \pm 2$ is based on two foundations, one is that conventional Delphi method usually repeats more than four times, the other is psychological research shows that 7 ± 2 is an experiential value for human being's thought span. The specific value for t is given by the organizer according to the time limitation and the rigor of the problem before the beginning of group decision-making.

(7) Based merely on the decision-makers' preference judgment information, the system calculates the weight assigned to each decision-maker and yields the weight vector $W = (w_1, w_2, \cdots, w_s)$ for the group using the weight allocation method given in [9].

(8) Construct the Markov state transition matrix T using equation (4) and (5) with the saved preference values $\left\{\pi_r^t(x^1), \pi_r^t(x^2), \cdots, \pi_r^t(x^s)\right\}$.

(9) Multiply the weight vector $W = (w_1, w_2, \cdots, w_s)$ by the preference matrix Λ in the last round, and then by the Markov state transition matrix T we have

$$[w_1, w_2, \cdots, w_s]\begin{bmatrix} \pi_1^t(x^1) & \pi_1^t(x^2) & \cdots & \pi_1^t(x^s) \\ \pi_2^t(x^1) & \pi_2^t(x^2) & \cdots & \pi_2^t(x^s) \\ \cdots & \cdots & \cdots & \cdots \\ \pi_l^t(x^1) & \pi_l^t(x^2) & \cdots & \pi_l^t(x^s) \end{bmatrix}\begin{bmatrix} T_{11} & T_{12} & \cdots & T_{1s} \\ T_{21} & T_{22} & \cdots & T_{2s} \\ \cdots & \cdots & & \cdots \\ T_{s1} & T_{s2} & \cdots & T_{ss} \end{bmatrix} = [x^1 \quad x^2 \quad \cdots \quad x^s] \tag{6}$$

where $[x^1 \quad x^2 \quad \cdots \quad x^s]$ is the preference utility values vector on the scheme set X and $\max\{x^i\}$ $(i = 1, 2, \cdots, s)$ is the final decision made by the group.

Here we must clarify some points on this group decision-making support approach.

(1) Above procedure goes on entirely and merely based on the interaction between decision-makers and the system without the organization's interruption as in traditional procedure, therefore the decision-making procedure is sped up and this approach can thus meet the requirement of making decisions in limited time.

(2) Each decision maker can defense her opinion, present her explanation, put forward her questions and browse other's opinions anonymously or with her true-name.

(3) The clustering result of the preference values in each round is published and used as a reference for preference adjustment in next round.

(4) Every decision-maker is encouraged to adjust her preference in each round. If every decision-maker insists her initial decision and does not adjust her preference in the group procedure, the group can never achieve consensus and the group decision-making makes no sense.

(5) When a decision-maker adjusts her preference, she can make tiny adjustment between two schemes without changing their orders. Such a tiny adjustment represents her accept or reject on the schemes when she cannot have the cake and eat it too, and will be reflected in the state transition matrix.

4 Conclusion

This paper researches on the procedure of preference adjustment in group decision-making, discusses a description of Markov chain and its main properties, and puts forward the hidden pattern of group preference change based on Markov state transition matrix for group decision-making problems. It also proofs the Markov property of group decision-making procedure and presents the steps for implementing the meta-synthesis approach for group decision-making. Further study will firstly focus on the development of such a meta-synthesis GDM approach and the empirical investigation of its effects.

References

1. Arrow K J. Social choice and individual values[M]. New York: JohnWiley&Sons, 1963
2. Dyer et al. Group Preference Aggregation Rules Based on Strength of Preference, Management Science, 1979, 25 (9): 22~34
3. Keeney et al. Group decision making using cardinal social welfare functions, Management Science, 1975, 22 (4): 430~437
4. French S. Decision Theory: An Introduction to the Mathematics of rationality, Ellis Harwood, Chi Chester, 1986
5. Bordly et al. On the Aggregation of Individual Probability Estimates, Management Science, 1981, 27 (8): 959~964
6. Hochbaum DS and Levin A. Methodologies and algorithms for group-rankings decision, Management Science, 2006, 52 (9): 1394-1408.
7. Fan ZP, Ma J, Jiang YP, Sun YH, and Ma L. A goal programming approach to group decision making based on multiplicative preference relations and fuzzy preference relations. European Journal of Operational Research, 2006, 174 (1): 311-321.
8. Herrera-Viedma E, Martinez L, Mata F, et al. A consensus support system model for group decision-making problems with multigranular linguistic preference relations. IEEE Transactions on Fuzzy Systems, 2005, 13 (5): 644-658.
9. Shen H.Z., Zhao J.D., Wang H.C., et al. Benchmarking service quality from the inside and outside of a service organization by group decision-making, International Conference on Services Systems and Services Management, JUN 13-15, 2005, pp. 502-507
10. Shen H.Z., Hu D.P., Wang H.C., Approaching Consensus Algorithm of Group Decision Based on Clustering Analysis□Proceedings of the fifth International Symposium on Knowledge and Systems Sciences, JAIST, Japen, November 10-12, 2004, pp. 267-273.

Discussion on the Spike Train Recognition Mechanisms in Neural Circuits

Yan Liu[1], Liujun Chen[1,⋆], Jiawei Chen[1], Fangfeng Zhang[1], and Fukang Fang[2]

[1] Department of Systems Science, School of Management, Beijing Normal University, Beijing 100875, P.R.China
chen1j@bnu.edu.cn
[2] Institute of Non-equilibrium Systems, Beijing Normal University, Beijing 100875, P.R.China
bnuliuyan@163.com, bnuchenlj@163.com

Abstract. The functions of neural system, such as learning, recognition and memory, are the emergences from the elementary dynamic mechanisms. To discuss how the dynamic mechanisms in the neurons and synapses work in the function of recognition, a dynamic neural circuit is designed. In the neural circuit, the information is expressed as the inter-spike intervals of the spike trains. The neural circuit with 5 neurons can recognize the inter-spike intervals in 5-15ms. A group of the neural circuits with 6 neurons recognize a spike train composed of three spikes. The dynamic neural mechanisms in the recognition processes are analyzed. The dynamic properties of the Hodgkin-Huxley neurons are the mechanism of the spike trains decomposition. Based on the dynamic synaptic transmission mechanisms, the synaptic delay times are diverse, which is the key mechanism in the inter-spike intervals recognition. The neural circuits in the group connect variously that every neuron can join in different circuits to recognize different inputs, which increases the information capacity of the neural circuit group.

Keywords: spike train, inter-spike intervals, response delay time, neural circuit.

1 Introduction

As a complex system, the functions of the neural system, such as learning, recognition and memory, are the emergences from the elementary dynamic mechanisms. Simulating by the complex networks is a common method to discuss the emergences from the complicated structure of the neural system. In the complex networks, the number of neurons could be as similar as in the neural system and the structure could be various[1]. We focus on the properties emerging from the complex structures, so the nodes and edges in the networks often have little dynamics. While in the real neural system, the neurons (the nodes) and the synapses (the edges) have plentiful dynamic mechanisms[2,3,4,5,6], which have

⋆ Corresponding author.

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 59–66, 2007.
© Springer-Verlag Berlin Heidelberg 2007

been proved to be the substrate for the functions in brain, such as learning and memory[7,8,9]. To analyze these neural and synaptic dynamic mechanisms in large networks may be difficult, so that some simple neural circuits are developed to discuss the dynamic mechanisms in the neural system [10,11,12,13].

In the neural system, sensory systems present environmental information to central nervous system as sequences of action potentials or spikes. So it is considered that the information is expressed as the inter-spike intervals of the spikes [11,14,15,16,17]. In this paper, to discuss how the dynamic mechanisms in the neurons and synapses work in the recognition in the neural system, a neural circuit is designed to recognize the inter-spike intervals and the spike trains. Several dynamic neural and synaptic mechanisms are analyzed in the recognition.

2 Inter-spike Intervals Recognition

In the neural system, the environmental signals are expressed and transferred among the neurons as the type of spikes. When the input to a Hodgkin-Huxley neuron is a spike train composed of two spikes, the response property of the neuron will be of four kinds as follows:

(1)Responds to both of the spikes.
(2)Only responds to the first spike, but not responds to the second one. Because the neuron is in refractory period at the time of the second spike.
(3)Not responds to the first spike, but responds to the second one. Because under its parameters, at least two spikes could make the membrane potential of the neuron integrate to the threshold.
(4)Not responds to either of the spikes.

In details, the neurons satisfy the Hodgkin-Huxley equation[5],

$$C\frac{dV}{dt} = g_{Na}m^3h(E_{Na} - V) + g_K n^4(E_K - V) + g_L(E_L - V) + I_{syn} \quad (1)$$

in which V is the membrane potential. m, n, h are the gating variables, which satisfy $\frac{dX}{dt} = -(\alpha_X + \beta_X)X + \alpha_X$ for $X = m, h, n$. $\alpha_m = -0.1(25-V)/[\exp((25-V)/10) - 1]$, $\beta_m = 4\exp(-V/18)$, $\alpha_h = 0.07\exp(-V/20)$, $\beta_h = 1/[1 + \exp((30-V)/10)]$, $\alpha_n = 0.01(10-V)/[\exp((10-V)/10) - 1]$, $\beta_n = 0.125\exp(-V/80)$. The parameters are $C = 1\mu F/cm^2$, $E_{Na} = 120mV$, $E_K = -12mV$, $E_L = 10.6mV$, $g_{Na} = 120mS/cm^2$, $g_K = 36mS/cm^2$, $g_L = 0.3mS/cm^2$. $I_{syn}$is the synaptic current,

$$I_{syn} = \overline{g}_{syn}r(t)(V - E_{syn}) \quad (2)$$

where$\overline{g}_{syn}$is the maximal synaptic conductionit also shows the synaptic strength. $r(t)$ is the amount of neurotransmitter released from the pre-synapse, which is also the input signal here. Figure 1 shows that the responses of the neurons depend on the parameter of $\overline{g}_{syn}$ and the input inter-spike interval T. In the region 1, the neurons respond to both of the spikes. In the region 2, the neurons only respond to the first spike. In the region 3, the neurons only respond to the second spike. In the region 4, the neurons do not respond to either of the spikes.

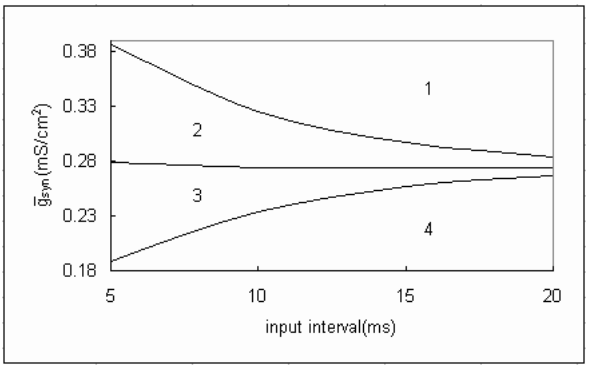


Fig. 1. The responses of the Hodgkin-Huxley neurons vary upon the $\overline{g}_{syn}$

Based on the characters above, a neural circuit is designed to recognize the inter-spike interval of the two input spikes (Fig.2). The circuit includes three layers. The first layer is the input neuron X, which fires two spikes as the input signal and the inter-spike interval is T_1. Suppose that the input spikes are at the time of t_0 and $t_1 = t_0 + T_1$. The second layer includes neurons α, β and γ. The input spikes from neuron X are transferred to the neurons α and γ through the synapses. The neuron α, with $\overline{g}^{\alpha}_{syn}$ in the region 2, only responds to the first spike (the spike at t_0) and the response delay time is τ_α. So the neuron α fires at the time of $t_\alpha = t_0 + \tau_\alpha$. The neuron γ, with $\overline{g}^{\gamma}_{syn}$ in the region 3, only responds to the second spike (the spike at t_1) and the response delay time is τ_γ. So the neuron γ fires at the time of $t_\gamma = t_1 + \tau_\gamma = t_0 + T_1 + \tau_\gamma$.The neuron β, with $\overline{g}^{\beta}_{syn}$ in the region 1, receives the spike from the neuron α and the response delay time is τ_β. So the neuron β fires at the time of $t_\beta = t_0 + \tau_\alpha + \tau_\beta$. The output layer neuron Y is a detect neuron, which receives the spikes from the neurons β and γ, and fires only when two spikes arrive within a time window of εms. Thus, when the neuron Y fires, means that $|t_\beta - t_\gamma| = |(\tau_\alpha + \tau_\beta) - (T_1 + \tau_\gamma)| < \varepsilon$.

The response delay times of the neurons α, β and γ are related with $\overline{g}^{\alpha}_{syn}$, $\overline{g}^{\beta}_{syn}$ and $\overline{g}^{\gamma}_{syn}$ that $\tau_\alpha = f_1(\overline{g}^{\alpha}_{syn})$, $\tau_\beta = f_1(\overline{g}^{\beta}_{syn})$, $\frac{df_1}{d\overline{g}_{syn}} < 0$, and $\tau_\gamma = f_2(\overline{g}^{\gamma}_{syn}, T)$, $\frac{\partial f_2}{\partial \overline{g}_{syn}} < 0$, $\frac{\partial f_2}{\partial T} > 0$. So when Y fires, it means the parameters $\overline{g}^{\alpha}_{syn}$, $\overline{g}^{\beta}_{syn}$ and

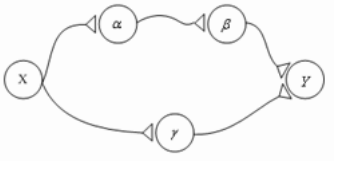


Fig. 2. The neural circuit structure for an inter-spike interval recognition

$\overline{g}^{\gamma}_{syn}$ in the neural circuit match the input interval T that $|t_\beta - t_\gamma| = |(\tau_\alpha + \tau_\beta) - (T_1 + \tau_\gamma)| < \varepsilon$, which equals to $|[f_1(\overline{g}^{\alpha}_{syn}) + f_1(\overline{g}^{\beta}_{syn})] - [T + f_2(\overline{g}^{\gamma}_{syn}, T)]| < \varepsilon$. Therefore, under different parameters of $\overline{g}^{\alpha}_{syn}$, $\overline{g}^{\beta}_{syn}$ and $\overline{g}^{\gamma}_{syn}$, the neural circuits could recognize the corresponding inter-spike intervals as $T \pm \varepsilon$.

However, a neural circuit recognizing only one inter-spike interval is not actual in the neural system. In fact, much experiments show that the neurons in brain join in different groups under different stimuli. As shown in Fig. 3, several neural circuits make up a large group. Not only the neurons $\alpha_i \rightarrow \beta_i$ and γ_i in the same circuit have an output neuron, but also the neurons $\alpha_i \rightarrow \beta_i$ and γ_j in different circuits connect to an output neuron. Such kind of structure optimizes the whole circuit group. For the delay times in different circuits are distributing in a wide range that any neurons $\alpha_i \rightarrow \beta_i$ and γ_j may combine to form a circuit, the whole circuit group can recognize the input intervals by choosing different circuit combinations. Thus, with the same number of neurons, the circuit group can process more information. For example, the group of 5 neural circuits may have 25 different combinations, which can recognize 25 different intervals.

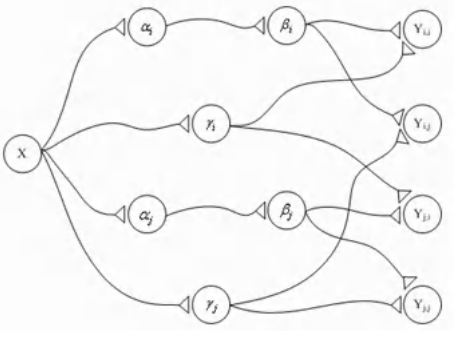


Fig. 3. The neural circuit group for inter-spike intervals recognition

Fig. 4 shows a group of 5 neural circuits recognizes the inter-spike intervals of 5ms, 10ms and 15ms. The vertical axis is the 25 output neurons. The dot, cross and circle express the corresponding neurons fire at that time, with dot for the fired neuron responding to the input interval of 5ms, and cross for 10ms, circle for 15ms respectively. As in Fig. 4, when the input inter-spike intervals are different, the circuit group would choose different circuit combinations to transfer the signals, and the output neurons would fire as different spatial patterns.

3 Spike Trains Recognition

Generally, a spike train includes several spikes and the information is expressed by a string of inter-spike intervals. A neural circuit is designed to recognize the spike train with more than one inter-spike interval (Fig. 5). Consider the simplest spike train including three spikes, with two inter-spike intervals of T_1 and T_2.

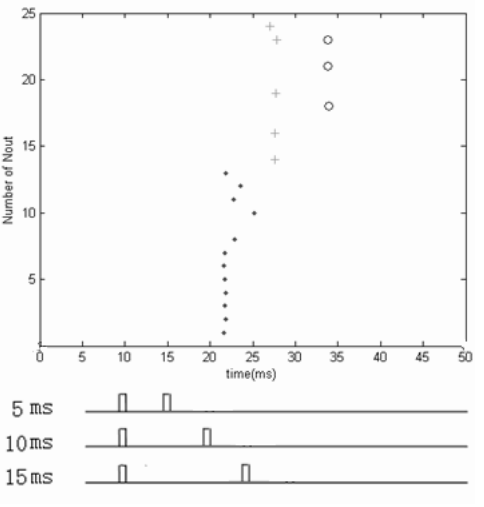


Fig. 4. Responding to the different input intervals, the output neurons fire as different spatial patterns

The circuit has three layers. The first layer is the input neuron X, which fires at t_0, $t_1 = t_0 + T_1$ and $t_2 = t_1 + T_2$. The second layer is composed of neurons α, β, γ and η. The input spikes from neuron X are transferred to the neurons α and γ through the synapses. The neuron α, with $\overline{g}^{\alpha}_{syn}$ in the region 2, only responds to the first and the third spikes (the spikes at t_0, and $t_2 = t_1 + T_2$) and the response delay time are τ^0_α and τ^2_α . So the neuron α fires at the time of $t^1_\alpha = t_0 + \tau^0_\alpha$ and $t^2_\alpha = t_2 + \tau^2_\alpha$. The neuron β, with $\overline{g}^{\beta}_{syn}$ in the region 1, receive the spike from the neuron α and the response delay time is τ_β. So the neuron β fires at the time of $t^1_\beta = t_0 + \tau^0_\alpha + \tau_\beta$ and $t^2_\beta = t_2 + \tau^2_\alpha + \tau_\beta$. The neuron γ, with $\overline{g}^{\gamma}_{syn}$ in the region 3, only responds to the second spike (the spike at t_1) and the response delay time is τ_γ. So the neuron γ fires at the time of $t_\gamma = t_1 + \tau_\gamma$. The neuron η, with $\overline{g}^{\eta}_{syn}$ in the region 1, responds to all spikes from the neuron γ and the response delay time is τ_η. So the neuron η fires at the time of $t_\eta = t_1 + \tau_\gamma + \tau_\eta$.The output layer has two neurons Y^1 and Y^2, which are the detect neurons. Neuron Y^1 receives the spikes from the neurons β and γ, and fires when two spikes arrive within a time window of εms. Neuron Y^2 receives the spikes from the neurons α and η, and fires when two spikes arrive within a time window of εms.

As shown in Fig. 5, the neural circuit has two sub-circuits to transfer the input spike train. One sub-circuit is composed of the neurons α, β, γ and Y^1,which recognize the first inter-spike interval T_1 in the spike train. The other is sub-circuit is composed of the neurons γ, η, α and Y^2, which recognize the second inter-spike interval T_2 in the spike train. The structure of the two sub-circuits is same with the circuit in Fig. 2, and they have the same inter-spike interval recognition mechanisms. So the whole neural circuit of Fig. 5 can recognize a spike train with two inter-spike intervals. In details, the output

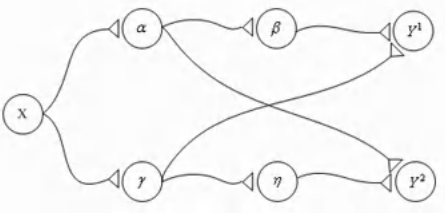


Fig. 5. The neural circuit structure for a spike train recognition

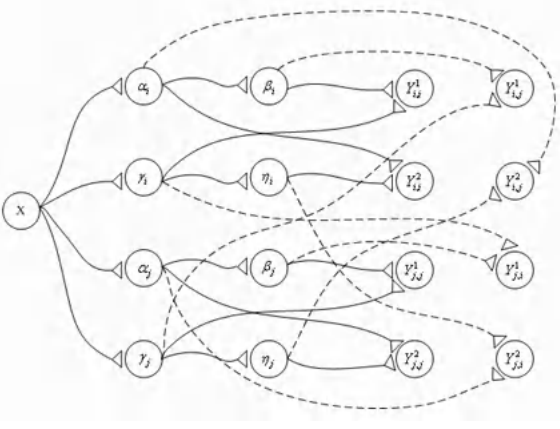


Fig. 6. The neural circuit group for spike trains recognition

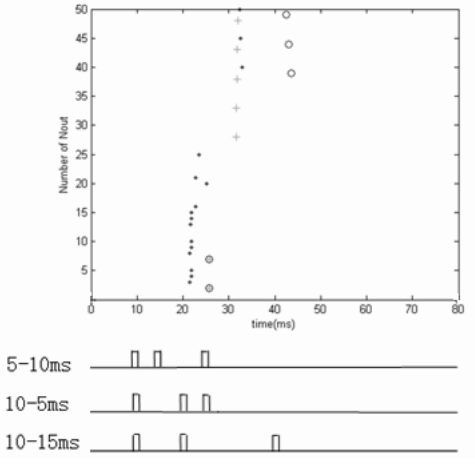


Fig. 7. Responding to the different spike trains, the output neurons fire as different spatial patterns

neuron Y^1 receives the spikes from the neuron β and the neuron γ at the times of t_β^1, t_β^2 and t_γ. Only when $|t_\beta^1 - t_\gamma| = |(\tau_\alpha^0 + \tau_\beta) - (T_1 + \tau_\gamma)| < \varepsilon$, the neuron Y^1 responds to fire. Similarly, the output neuron Y^2 receives the spikes from the neuron η and the neuron α at the times of t_η and t_α^1, t_α^2. Only when $|t_\eta - t_\alpha^2| = |(\tau_\gamma + \tau_\eta) - (T_2 + \tau_\alpha^2)| < \varepsilon$, the neuron Y^2 responds to fire. Thus, according to the output neurons Y^1 and Y^2 fire or not, the neural circuit recognizes whether the input spike train includes the inter-spike intervals of T^1 and T^2.

Several such neural circuits make up a large group. As shown in Fig. 6, not only the neurons $\alpha_i \rightarrow \beta_i$ and $\gamma_i \rightarrow \eta_i$ in the same circuit have the output neurons, but also the neurons $\alpha_i \rightarrow \beta_i$ and $\gamma_j \rightarrow \eta_j$ in different circuits connect the output neurons. Fig. 7 shows a group of 10 neural circuits with 50 output neurons recognizes three different spike trains.

The input spike trains are all composed of three spikes with inter-spike intervals of 5ms-10ms, 10ms-5ms and 10ms-15ms respectively. In Fig. 7, the vertical axis is the 50 output neurons. The dot, cross and circle express the corresponding neuron fires at that time, with dot for the neuron responding to the input spike train of 5ms-10ms, and cross for 10ms-5ms, circle for 10ms-15ms respectively. When the input spike trains are different, the circuit group chooses different circuits combinations to transfer the signals, and the corresponding output neurons fire. Therefore, the whole circuit group recognizes the input spike trains with the different output spatial patterns.

4 Conclusion

In this paper, the dynamic mechanisms in the neurons and the synapses in the learning and recognition are discussed. A neural circuit is designed to recognize the inter-spike intervals and the spike trains. There are several dynamic neural mechanisms in the recognition processes. When the input signal is a train of two spikes, the Hodgkin-Huxley neuron with dynamic synapses will response variously. With the different parameters, the response delay times are also different. Under this mechanism, the neural circuit with 5 neurons can recognize inter-spike intervals in 5-15ms. The synaptic delay time is one of the key variables in the recognition. Several neural circuits with 6 neurons make up a large group. The result shows that a group of neural circuits can recognize a spike train with three spikes. When the input signals are different, every neuron can join in different circuits and the output neurons form a spatial pattern. This structure increases the information capacity of the neural circuit, which is different with the common neural network. For the neural circuit here is still a small one, with no more than 50 neurons in each layer, the parameters and the number of neurons in the circuit are fixed. In the future work, a larger neural network will be developed, in which the parameters of neurons may be not fixed before the recognition.

Acknowledgement

This work is supported by NSFC under the grant No.60534080, No.60374010 and No.70471080.

References

1. Strogatz S. H., Exploring complex networks. Nature, 2004, 410:268-276.
2. Izhikevich E. M., Dynamical systems in neuroscience: the geometry of excitability and bursting. MIT, 2005.
3. Hebb D. O., The Organization of Behavior. Wiley: New York, 1949.
4. Bliss T. V.,Collingridge G. L., A synaptic model of memory: long-term potentiation in the hippocampus. Nature, 1993, 361: 31-39.
5. Hodgkin A. L., Huxley A. F., A quantitative description of ion currents and its applications to conduction and excitation in nerve membranes. J. Physiol, 1952, 117: 500-544.
6. Izhikevich E. M., Dynamical systems in neuroscience: the Geometry of excitability and bursting. MIT, 2007.
7. Abbott L. F., Regehr W. G., Synaptic computation. Nature, 2004, 431:796-803.
8. Mark F. B., Barry W. C., Michael A. P., Neuroscience: exploring the brain, 2nd ed. Lippincott Williamsand Wilkins, 2001.
9. Oswald A. M., Schiff M. L., Synaptic mechanisms underlying auditory processing. Curr. Opin. Neurobiol., 2006, 16: 371-376.
10. Abarbanel H. D. I., Talathi S. S., Neural circuitry for recognizing interspike interval sequences. Phys Rev Lett, 2006, 96: 148104.
11. Jin D. Z., Spiking neural network for recognizing spatiotemporal sequences of spikes. Phys Rev E, 2004, 69: 021905.
12. Mauk M. D., Buonomano D. V., The neural basis of temporal processing. Ann Rev Neurosci, 2004, 27: 307-340.
13. Large E. W., Crawford J. D., Auditory temporal computation: Interval selectivity based on post-inhibitory. J. Comput. Neurosci., 2002, 13: 125-142.
14. Buonomano D. V., Merzenich M. M., Temporal information transformed into a spatial code by a neural network with realistic properties. Science, 1995, 267: 1028-1030.
15. Lisman J. E., Bursts as a unit of neural information:making unreliable synapses reliable. Trends Neurosci., 1997, 20: 38-43.
16. Izhikevich E. M., Polychronization: Computation with Spikes. Neur Comp, 2006, 18: 245-282.
17. Buonomano D. V., Decoding temporal information: a model based on short-term synaptic plasticity. J. Neurosci, 2000, 20(3): 1129-1141.

Extended Clustering Coefficients of Small-World Networks

Wenjun Xiao[1], Yong Qin[1,2], and Behrooz Parhami[3]

[1] Dept. Computer Science, South China University of Technology, Guangzhou 510641, P.R. China
wjxiao@scut.edu.cn
[2] Center of Computer Network and Information, Maoming University, Maoming 525000, P.R. China
[3] Dept. Electrical & Computer Eng., University of California, Santa Barbara, CA 93106-9560, USA

Abstract. The clustering coefficient C of a network, which is a measure of direct connectivity between neighbors of the various nodes, ranges from 0 (for no connectivity) to 1 (for full connectivity). We define extended clustering coefficients $C(h)$ of a small-world network based on nodes that are at distance h from a source node, thus generalizing distance-1 neighborhoods employed in computing the ordinary clustering coefficient $C = C(1)$. Based on known results about the distance distribution $P_\delta(h)$ in a network, that is, the probability that a randomly chosen pair of vertices have distance h, we derive and experimentally validate the law $P_\delta(h)C(h) \leq clogN/N$, where c is a small constant that seldom exceeds 1. This result is significant because it shows that the product $P_\delta(h)C(h)$ is upper-bounded by a value that is considerably smaller than the product of maximum values for $P_\delta(h)$ and $C(h)$.

1 Introduction

Complex systems with many components and associated interactions abound in nature, prevail within society, and define many human artifacts. The interconnection or interaction structure of such systems are typically neither random (amenable to probabilistic analysis) nor regular (mathematically tractable), rendering the systematic study of their properties a rather challenging undertaking. Interactions in such systems can be modeled by networks/graphs composed of vertices/nodes and undirected or directed edges/links. A network or graph $G = (V, E)$has a set V of N vertices or nodes and a set E of M edges or links, where each edge is defined by a pair of vertices (ordered pair, for directed graphs).

Two models of actual complex networks have been studied extensively [1,2,3,4]: the small-world model and the scale-free one. Our focus in this paper is on small-world networks that feature localized clusters connected by occasional long-range links, leading to an average distance between vertices that grows logarithmically with the network size N.

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 67–73, 2007.
© Springer-Verlag Berlin Heidelberg 2007

Watts and Strogatz [2] studied mechanisms via which a regular network can be transformed into a small-world network, with little or no change in the vertex-degree distribution, and quantified the parameters that characterize the resulting structures. One feature shared by small-world networks is that their clustering coefficients are fairly high compared with random networks[2]. Clustering coefficient is defined as follows. Let a vertex v of G have $k(v)$ neighbors; that is, v has degree $k(v)$. These $k(v)$ neighbors can potentially be connected via $k(v)(k(v)-1)/2$ edges. The fraction of this maximum possible number of edges that actually exist between neighbors of v is its clustering coefficient C_v; the average of clustering coefficients over all $v \in V$ is the clustering coefficient C of the network G. A network with C close to 1 may consist of highly connected clusters or cliques, perhaps with sparser connections between the local clusters.

2 Extended Clustering Coefficients

We extend the clustering coefficient of Watts and Strogatz [2] in the following way. Define the $h-$neighbors of a vertex v as vertices of G that are at distance h (measured in number of hops) from v. Assume that v has $k_h(v)$ such $h-$neighbors, where $k_1(v)$ is the same as $k(v)$ defined earlier (see Section 1). Then there can be at most $k_h(v)(k_h(v)-1)/2$ edges connecting $h-$neighbors of vertex v. The fraction $C_v(h)$ of allowable edges that actually exist between $h-$neighbors v is the $h-$clustering coefficient of v. We assume that $C_v(h) = 1$ when $k_h(v) = 1$, which also covers the special case $h = 0$. The average of $C_v(h)$ over all $v \in G$ is the $h-$clustering coefficient $C(h)$ of G. The $1-$clustering coefficient $C(1)$ is the clustering coefficient C as defined in Section 1.

Thus, while the definition of clustering coefficient is based on the immediate neighborhood of vertices, extended clustering coefficient relates to a wider neighborhood defined by the distance parameter h. Using experimental data from a wide variety of actual complex networks, along with a deterministic model of small-world networks that we have developed, we seek to relate $C(h)$ and the distance distribution $P_\delta(h)$ of a network, defined as the probability that a randomly chosen pair of vertices are at distance h from each other.

Note that all distances referred to in this paper are shortest distances. However, in view of the results of Kim et al. [5], distances obtained from a routing algorithm with localized decisions are not fundamentally different from shortest distances in complex networks. Thus, our results are expected to remain valid when this latter definition of distance is used in lieu of shortest distance.

3 Experimental Observations

For an $N-$vertex network with M edges, we have $P_\delta(0) = 1/N$ and $P_\delta(1) = 2M/N^2 > 1/N$. Beyond $h = 1$, however, a precise expression for the value of $P_\delta(h)$ cannot be supplied, except in the case of certain regular networks. However, for many networks (small-world or otherwise), the value of $P_\delta(h)$ rises with h until it reaches a maximum value and then declines as the distance h gets

closer to the network diameter D. This is confirmed experimentally for several complex networks of practical interest in Fig1b, Fig2b.

For extended clustering coefficients, the trend begins by a decrease in clustering, from $C(0) = 1$ to $C(1) = C$, and is then followed by further reductions. This is owing to the fact that as h increases, the number q_h of nodes at distance h from a given node increases, and such nodes are apt to belong to several cliques; hence, the presence of many edges between them is quite unlikely. As h approaches D, however, a different effect may take hold. Consider, for example, one extreme case where each node in the network is at distance D from exactly one node (it has a single diametrically opposite node). This leads to $C(D) = 1$. In this same situation, $C(D-1)$ is likely to be large as well, given the common presence of multiple diametral paths to the same opposite vertex. Note that the preceding argument suggests that $C(h)$ can be large when h approaches D; it does not imply that $C(h)$ must be large in this case. Figures 1c and 2c confirm these trends.

Given the opposing trends of $P_\delta(h)$ (up, then down) and $C(h)$ (down, then possibly up), one might be led to believe that the product $P_\delta(h)C(h)$ has an upper bound. Based on the evidence presented in Fig1a and Fig2a, we conjecture that this is in fact the case. That is, for a constant c in the vicinity of and seldom exceeding 1, we have:

$$P_\delta(h)C(h) \leq clogN/N \tag{1}$$

In the special case of $h = 1$, equation (1) implies $P_\delta(1)C(1) \approx logN/N$. We have $P_\delta(1) = 2M/N^2 \approx logN/N$ for small-world networks. This is consistent with $C(1) = C$ being large for such networks.

4 Model-Based Validation

We now present additional evidence on the validity of equation (1), using a model of deterministic small-world networks that we have developed [6]. In fact, it was through this model that we first became aware of the trend represented in equation (1) and experimentally confirmed in Fig1 and Fig2. A review of our deterministic model, which is based on Cayley graphs [7], has been provided in the supporting information, where we also show that the model yields the clustering coefficient:

$$C = \frac{at(at-1)}{(at+t-l)(at+t-l-1)} \tag{2}$$

In this model, $t = log_2 N$ and $a = (2^l - 1)/t$ is a free tuning parameter that is related to the interconnection density, thereby affecting the value of C. Note that for very large networks ($N, t \to +\infty$), C tends to $a^2/(a+1)^2$ when a is a constant. By suitably choosing a, we can obtain different clustering coefficients, while maintaining a small vertex degree equal to $at + t - l = (a+1)log_2 N - 1$.

Unlike actual networks for which the computation of $C(h)$ is extremely difficult, our deterministic model is amenable to mathematical analysis that yields

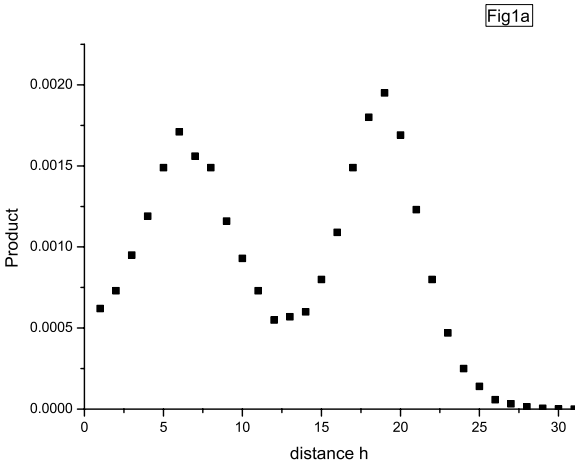


Fig. 1a. The plot of Product($P_\delta(h)C(h)$) versus h in the maximum component Δ of NCSTRL graph [9], with 6396 vertices and diameter of 31

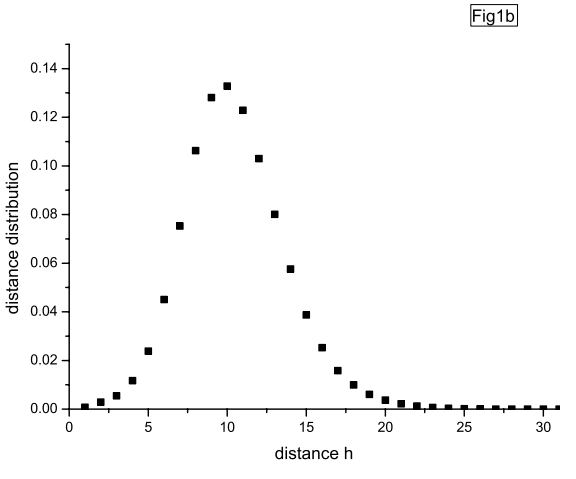


Fig. 1b. The plot of distance distribution($P_\delta(h)$) versus h in Δ

an approximate closed-form expression for the extended clustering coefficients. In our deterministic model, the number m of adjacent vertex pairs among the $h-$neighbors of any vertex is given by the expression:

$$m = (2^l - 1)(2^{l-1} - 1)\binom{t-l}{h-1} \tag{3}$$

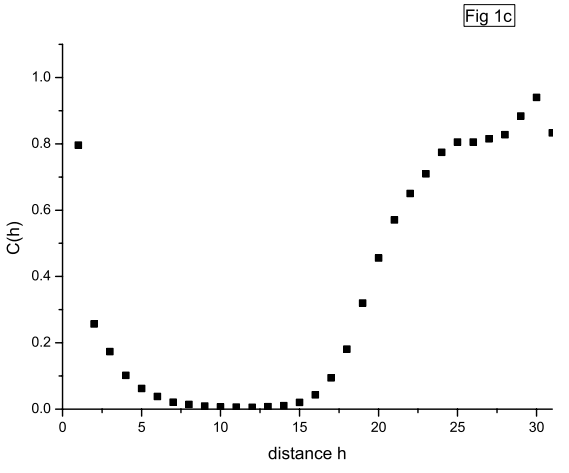


Fig. 1c. The plot of $C(h)$ versus h in Δ

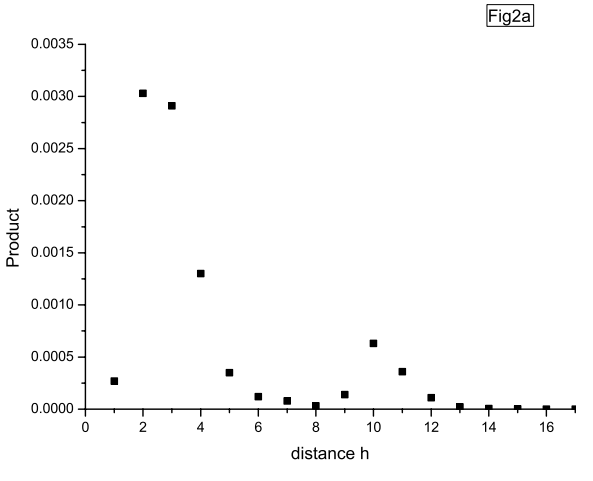


Fig. 2a. The plot of Product$(P_\delta(h)C(h))$ versus h in the maximum component $\Delta 1$ of Linux graph [10] with 5285 vertices and diameter of 17

On the other hand, the number $k_h(v)$ of $h-$neighbors of a vertex v is bounded as:

$$(2^l - 1)\binom{t-l}{h-1} \leq k_h(v) \leq (2^l - 1)\binom{t-l}{h-1} + \binom{t-l}{h} \tag{4}$$

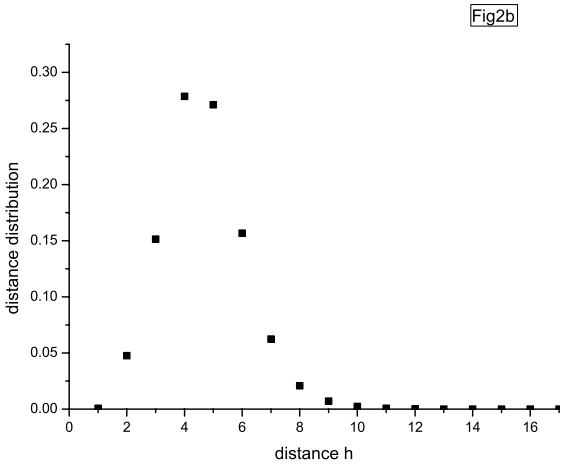


Fig. 2b. The plot of distance distribution($P_\delta(h)$) versus h in $\Delta 1$

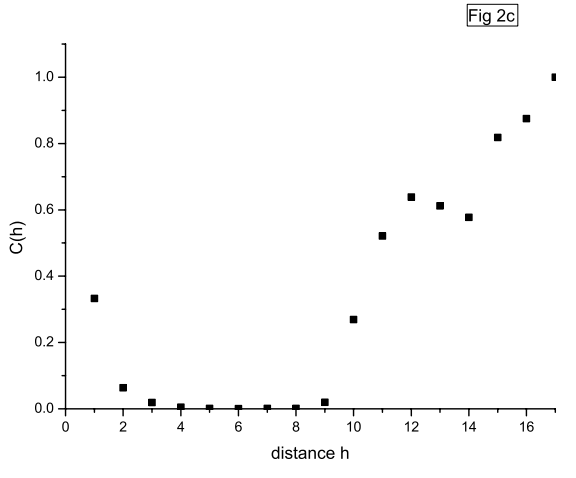


Fig. 2c. The plot of $C(h)$ versus h in $\Delta 1$

Given that the extended clustering coefficient $C(h)$ is proportional to $m/(k_h(v))^2$, we readily find:

$$C(h) \approx 1/\binom{t-l}{h-1} \tag{5}$$

In a companion paper[8], we have derived the distance distribution for small-world networks:

$$P_\delta(h) \approx \binom{logN}{h} /N \tag{6}$$

Here, we have $logN \approx D$. Because the diameter of our deterministic network model is $D = t - l + 1$, we conclude:

$$P_\delta(h) \approx \binom{t-l-1}{h} /N = (t-l+1)\binom{t-l}{h-1}/(hN) \tag{7}$$

Equations (5) and (7) lead to:

$$P_\delta(h)C(h) \leq clogN/N \tag{8}$$

Equation (8) confirms our hypothesis in equation (1), thereby supplementing the previously supplied experimental evidence of its validity.

5 Conclusion

We have shown that extended clustering coefficients are generalizations of ordinary clustering coefficient and are governed by laws that are also generalizations of those pertaining to the latter. We have presented experimental and analytical evidence that the inequality $P_\delta(h)C(h) \leq clogN/N$ holds for small-world networks. This result is significant because it shows that the product $P_\delta(h)C(h)$ is upper-bounded by a value that is considerably smaller than the product of maximum values for $P_\delta(h)$ and $C(h)$. Thus, extended clustering coefficients offer new insights into the structure of small-world networks and open up further avenues for exploration of their properties. Additionally, different shapes for the variations of $C(h)$ and $P_\delta(h)C(h)$, exemplified by Fig1 and Fig2, can be used to categorize small-world networks in order to facilitate their study.

Acknowledgments

The authors thank M. E. J. Newman for providing the NCSTRL data used in Fig. 1. Research of W. Xiao is supported by the Natural Science Foundation of China and Guangdong Province.

References

1. Barabsi, A.-L., & Albert, R. (1999) Science 286, 509-512
2. Watts, D. J., & Strogatz, S. H. (1998) Nature 393, 440-442.
3. Albert, R., & Barabsi, A.-L. (2002) Rev. Mod. Phys. 74, 47-97.
4. Newman, M. E. J. (2003) SIAM Rev. 45, 167-256.
5. Kim, B. J., Yoon, C. N., Han, S. K., & Jeong, H. (2002) Phys. Rev. E65, 027103.
6. Xiao, W. J., & Parhami, B. (2006) Info. Proc. Lett. 97, 115-117.
7. Biggs, N. (1993) Algebraic graph theory (Cambridge Univ. Press).
8. Xiao, W.J., & Parhami, B. (2005) On conditions for scale-freedom in complex networks, submitted for publication.
9. Newman, M. E. J. (2001) Phys. Rev. E 64, 016131.
10. Myers, C. R. (2003) Phys. Rev. E 68, 046116.

Detecting Invisible Relevant Persons in a Homogeneous Social Network

Yoshiharu Maeno[1], Kiichi Ito[2], and Yukio Ohsawa[3]

[1] Graduate School of Systems Management, Tsukuba University,
3-29-11 Otsuka, Bunkyo-ku, Tokyo, 112-0012 Japan
maeno.yoshiharu@nifty.com
[2] Graduate School of Media and Governance, Keio University,
5322 Endo, Fujisawa-shi, Kanagawa 252-8520 Japan
[3] School of Engineering, the University of Tokyo,
7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8563 Japan

Abstract. An algorithm to detect invisible relevant persons in a homogeneous social network is studied with computer simulation. The network is effective as a model for contemporary inter-working terrorists where large hub persons do not exist. Absense of large hub persons results in that the observed communication flow is also homogeneous. Clues regarding invisible relevant persons are hardly found in communication records. This task is, therefore, generally difficult. We demonstrate that our algorithm identifies the portion of the market baskets representing communication records, where invisible relevant persons are likely to be hidden, with good precision.

1 Problem

The activity of an organization is often under influence of invisible, but relevant persons. The influence is not directly observed. For example, a coordinator, who provides a number of activists with clever plans, communication means, attacking skill, money, etc., is hidden behind a terrorist network. The coordinator plays a role to synchronize the whole network toward a target. Understanding such invisible relevant persons from observed data provides important clues to invent hypothetical scenarios on opportunity and threat in business and social problems.

Network models such as a scale-free network (governed by a power law) [Barabási 1999] or a small-world (governed by an exponential law) [Watts 1998], have been studied as a generic abstraction of society, an economic system, and nature. It should, however, be noticed that individual systems have various concrete properties. (1) A model for contemporary inter-working terrorists, (2) a model for a self-organizing community, and (3) a model for a purposefully organized business team have quite different structures. Center-and-periphery structure, characterized by the existence of big hub persons, is a common property of the models (2) and (3). We call such a model an inhomogeneous social network. Invisible relevant persons are usually the hub persons. They can be detected by

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 74–81, 2007.
© Springer-Verlag Berlin Heidelberg 2007

investigating communication from the hub persons toward peripheral persons. On the other hand, the model (1) is different in that it does not possess representative central persons. It is compared to networks of networks. We call this model a homogeneous social network.

In this paper, we study an algorithm to detect invisible relevant persons in a homogeneous social network. The algorithm aims at identifying precisely the portion of the observed communication, which includes traces and clues to the invisible relevant persons. We generate a test data set of market basket form representing communication among visible and invisible persons from a homogeneous social network with computer simulation. We demonstrate the precision characteristics of the algorithm with the test data set.

2 Social Network Model

We present a basic structure of a homogeneous social network, compared with inhomogeneous networks. Three models mentioned in 1 are described.

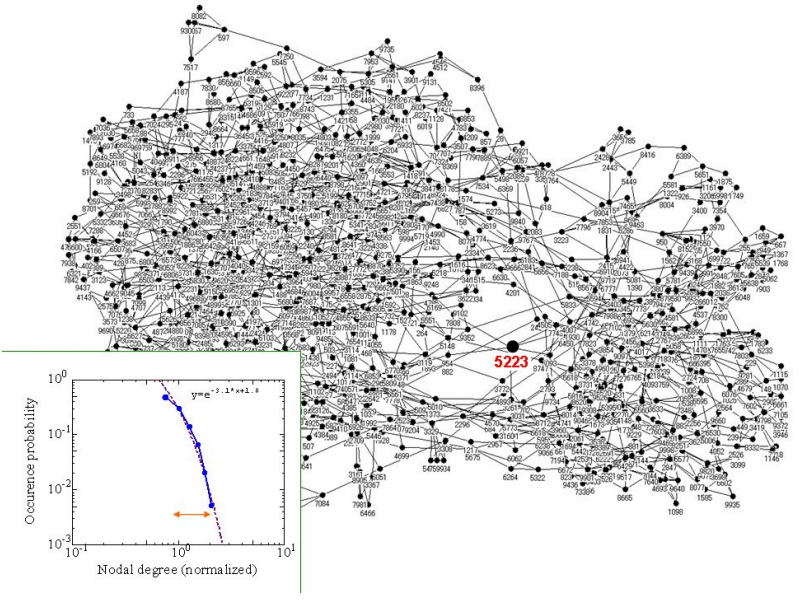


Fig. 1. Model for contemporary inter-working terrorists: a homogeneous network consisting of 995 nodes. The inset shows distribution of nodal degree. The persion ID=5223 is used in the evaluation.

A model for contemporary inter-working terrorists is illustrated in Fig.1. The network is derived from a number of empirical studies, simulation analysis, and jounalistic articles on terrorism [Popp 2006], [Singh 2004]. The network consists

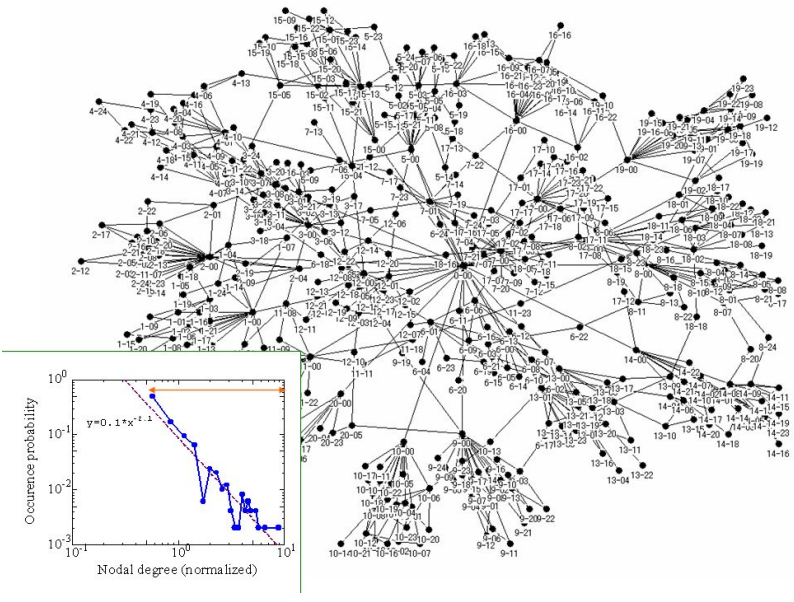


Fig. 2. Model for a self-organizing community: an inhomogeneous network consisting of 490 nodes. It is a scale-free network governed by a power low (Barabási-Albert model). The inset shows distribution of nodal degree.

of 995 nodes. As a whole, (1) the network seems to have two large groups: larger one on the left and smaller one on the right, (2) smaller groups seem to exit inside the two groups, (3) the boundary between the groups is not clear, and (4) the network does not possess big hub persons providing with a center facility among persons. The inset shows the occurrence distribution of nodal degree. The horizontal axis is normalized degree; degree devided by the average degree. It is governed by an exponential law; $y \propto e^{-3.1x}$. The degree ranges from 3 to 8. The average degree is 3.9. The deviation in the degree is small. It is the characteristics of the homogeneous network. Most persons are equivalent in that they have similar degree or centrality measure. Absense of large hub persons results in the fact that communication flow is also homogeneous. This model is used in the simulation study in 4.

A model for a self-organizing community is illustrated in Fig.2. It is a scale-free network derived from Barabási-Albert model [Barabási 1999]. The scale-free network is used to describe World Wide Web strucutre, scientist's collaboration network, etc.. The network consists of 490 nodes. The inset shows the occurrence distribution of nodal degree. It is governed by a power low; $y \propto x^{-2.1}$. The average degree is 3.6. The deviation in the degree is large. About 10 big hub persons are easily identified. The hub persons influence the way the network operates. Detecting invisible persons in such a network was studied in [Maeno 2006a].

A model for a purposefully organized business team is illustrated in Fig.3. The network is derived from empirical studies and analysis on communication via

email exchange within Enron [Keila 2006]. Enron was an energy company which ended in bankruptcy in 2001 because of the institutionalized accounting fraud. The network consists of 184 nodes. The inset shows the occurrence distribution of nodal degree. The average degree is 22.8. The deviation in the degree is large. The distribution is, however different from that for Fig.2. There are not small peripheral persons. Most persons contribute to functioning of the network. This may be a characteristic of the business team.

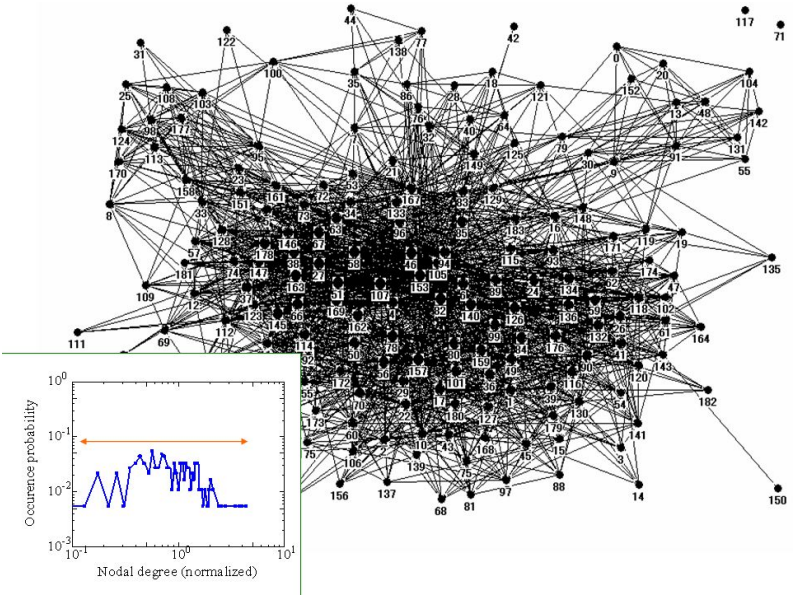


Fig. 3. Model for a purposefully organized business team: an inhomogeneous network consisting of 184 nodes. The inset shows distribution of nodal degree.

3 Algorithm

The algorithm to detect invisible persons in a homogeneous social network is presented. The algorthm is based on the crystallization algorithm [Ohsawa 2005] used in the human-interactive annealing process [Maeno 2007], [Maeno 2006b]. The input to the algorithm is observation data having market basket form; $b_i = \{e_j\}$. An individual event is denoted by e_j. A market basket is a set of events occuring simultaneously, spatially, or related strongly under a given subject. The output is ranking of b_i which indicates relative likeliness that the market basket provides clues to invisible events. The invisible event, which should have been in the market basket originally, is missing in the market baskets ranked highly.

The events are clustered into groups based on co-occurrence as a measure of similarity between events. Jaccard coefficient for the all pairs of events is

calculated from eq.(1). Jaccard coefficient is often used as a measure of co-occurrence in web mining and text analysis applications [Ohsawa 2006]. We can utilize various expertise on clustering. For example, k-medoids [Hastie 2001] or hierarchical clustering are simple, but efficient techniques. The k-medoids clustering is an EM algorithm similar to k-means algorithm for numerical data. A medoid event $e_{\mathrm{med}(j)}$ is an event locating most centrally within a cluster c_j. They are initially selected at random. Other $|E| - |C|$ events are classified into the clusters based on the Jaccard coefficint to the medoids. A new medoid is selected within the individual cluster so that the sum of Jaccard coefficints from events within the cluster to the modoid can be maximal. This is repeated until the medoid events converge. Advanced algorithms for unsupervised learning like self-oraganization mapping can also be employed.

$$J(i,j) \equiv \frac{\mathrm{Freq}(e_i \cap e_j)}{\mathrm{Freq}(e_i \cup e_j)}. \tag{1}$$

The ranking of b_i is evaluated as follows. A dummy event DE_i, representing invisible events, is inserted into b_i, that results in $b_i \rightarrow \{e_j\} \cup \mathrm{DE}_i$. The index i can be used to identify the market basket where the corresponding dummy event was inserted. The mixture of clusters resulting from the invisible events in the market basket is calculated with the Jaccard coefficient between the dummy event and the events in the clusters. It is evaluated according to eq.(2) using eq.(3). The market baskets which are ranked highly according to the largeness of eq.(2) are retrieved as candidate market baskets where invisible events should have been hidden.

$$I_{\mathrm{nu}}(i) \equiv \sum_{j=0}^{|c|-1} u(\max_{e_k \in c_j, b_i} J(\mathrm{DE}_i, e_k)). \tag{2}$$

$$u(x) = \begin{cases} 1 \text{ for } x > 0 \\ 0 \text{ otherwise} \end{cases}. \tag{3}$$

4 Evaluation

The objective of evaluation is to study how precise information regarding invisible persons the algorithm can retrieve from the test data with computer simulation. We use precision as a measure. In information retrieval, precision has been used as evaluation criteria. Precision is the fraction of relevant data among the all data returned by search.

4.1 Test Data

We describe the detail of the test data used for evaluation. It is generated from a homogeneous social network in Fig.3, as communication records among persons in the two steps below.

1. Collecting communication records into market basket: Market baskets representing communication among neighbor persons is generated. Persons within a specific distance (hop count) from a communication initiator are grouped into a market basket. This corresponds to a single conversation taking place regarding a specific subject the communication initiator concerns. An example market basket is $b_1 = \{954, 1930, 3261, 5093, 5223, 7743, 7808, ...\}$, representing communication initiated by the person ID=954.
2. Configuring invisible relevant persons: A latent structure regarding persons of interest is configured to the market basket by deleting the person from the data. Deletion made the structure invisible. As a result, the deleted persons and the links inter-connecting them become a latent structure hidden behind the market basket. The example market basket becomes $b_1 = \{954, 1930, 3261, 5093, 7743, 7808, ...\}$ if the person ID=5223 is an invisible person focused in evaluation. The resulting market baskets are like a bundle of email exchange records which lacks in oral communication from the invisible persons. The market baskets are the input to the algorithm. The algorithm attempts to identify b_1 as a candidate market basket where invisible persons should have been hidden.

For evaluation in 4.2, market baskets are configured from persons within five hops from individual persons in the first step 1. One hop is as long as one edge on the network shown in Fig.1. The number of persons within five hops is about 20% of the whole persons on the average. This is a relatively long distance communication. The latent structure of interest includes fifteen persons within two hops from the person ID=5223. These persons are remarkable in that they are equally close to every person in the network. They are not like a CEO governing a whole company in a hierarchical construct. We focus on them regarding them as a coordinating strategist group, who provides a number of activists inter-working in the terrorist network with clever plans, communication means, attacking skill, money, etc.. Persons in either the large cluster on the left or the small cluster on the right do not occupy such unbiased position. These persons are deleted from the market basket in the second step, so that they can be made invisible within the market basket data input to the algorithm.

4.2 Precision

Here, precision is evaluated by calculating the ratio of correct market baskets within the market baskets retrieved by the algorithm. The correct market baskets are those where persons had been deleted in the second step of 4.1. A single simulation condition is demonstrated in this paper. Systematic study on various conditions is planned.

Fig.4 shows the calculated precision. The horizontal axis is the number of retrieved basket data according to the ranking the algorithm outputs. The box in the figure lists 11 of highly ranked market baskets. The top 10 baskets retrieved as candidates are correct. Precision is good. This indicates the algorithm provides relevant suggestions regarding invisible persons. Clues to trace the invisible persons themselves will be found by monitoring the communication of the

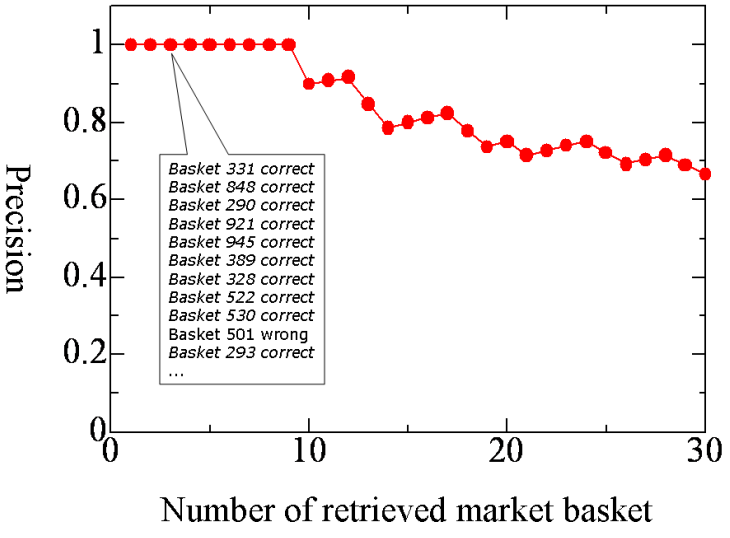


Fig. 4. Precision to identify the market baskets, where persons were made invisible, as a function of the number of retrieved market baskets

persons included in the market baskets. The algorithm is successful in a difficult task; detecting invisible persons in a homogeneous social network.

5 Discussion

An algorithm to detect invisible persons in a homogeneous social network was studied. The portion of the market baskets, where invisible persons are likely to exist, was identified with good precision. There are still remaining issues on the algorithm. We need to investigate two aspects more closely; (1) variability of social network structure and (2) communication pattern in terms of time sequence. Although we showed three models of social network, a number of alternative models exist for various domains of society, an economic system, and nature. Is a single algorithm applicable to such a wide range of models ? Communication sequence provides useful information to infer the influence flowing in a social network. Can the algorithm be modified with a time sequence analysis ? The topology of the network itself evolves as the communication pattern and consequently emerging inter-dependency between persons change. Understanding such dynamical nature is important to become aware of potential opportunity and threat arising from the latent structure hidden behind the surface behaviors.

We need to take a next step to understand the various impacts resulted from the social network, latent structures, and environment. For example, how do you avoid inconvenience caused by a specific action of a specific person in the social network influenced by an invisible relevant person's specific desicion making in a specific circumstance ? This is a hypothetical but concrete scenario invented to

obtain opportunity and to eliminate threat. For this purpose, we believe that it is essential to visualize various concepts and relationships in a map to recognize prior understanding and cognition to observation [Ohsawa 2006]. The human-interactive annealing process [Maeno 2007], [Maeno 2006b] is an effort toward such an objective.

References

[Barabási 1999] A. L. Barabási, R. Albert, and H. Jeong: Mean-field theory for scale-free random networks, Physica A **272**, 173-187 (1999).

[Hastie 2001] T. Hastie, R. Tibshirani, and J. Friedman: The elements of statistical learning: Data mining, inference, and prediction (Springer series in statistics). Springer-Verlag (2001).

[Keila 2006] P. S. Keila, and D. B. Skillicorn: Structure in the Enron email dataset, J. Computational & Mathematical Organization Theory **11**, 183-199 (2006).

[Maeno 2007] Y. Maeno, and Y. Ohsawa: Human-computer interactive annealing for discovering invisible dark events, to appear, IEEE Trans. Industrial Electronics (2007).

[Maeno 2006a] Y. Maeno, and Y. Ohsawa: Stable deterministic crystallization for discovering hidden hubs, Proc. IEEE Int'l. Conf. Systems, Man & Cybernetics, Taipei (2006).

[Maeno 2006b] Y. Maeno, K. Ito, K. Horie, and Y. Ohsawa: Human-interactive annealing for turning threat to opportunity in technology development, Proc. IEEE Int'l. Conf. Data Mining, Workshops, Hong Kong, 714-717 (2006).

[Ohsawa 2006] Y. Ohsawa eds.: Chance discovery in real world decision making. Springer-Verlag (2006).

[Ohsawa 2005] Y. Ohsawa: Data crystallization: chance discovery extended for dealing with unobservable events, New Mathematics and Natural Computation **1**, 373-392 (2005).

[Popp 2006] R. L. Popp, and J. Yen: Emergent information technologies and enabling policies for counter-terrorism, IEEE Press (2006).

[Singh 2004] S. Singh, J. Allanach, T. Haiying, K. Pattipati, and P. Willett: Stochastic modeling of a terrorist event via the ASAM system, Proc. IEEE Int'l. Conf. Systems, Man & Cybernetics, Hague, 6/5673-6/5678 (2004).

[Watts 1998] D. J. Watts, and S. H. Strogatz: Collective dynamics of small-world networks, Nature **398**, 440-442 (1998).

The Study of Size Distribution and Spatial Distribution of Urban Systems in Guangdong, China

Jianmei Yang, Dong Zhuang, and Minyi Kuang

School of Business Administration,
Institute of Emerging Industrialization Development,
South China University of Technology, P.R. China, 510640

Abstract. There are three types of urban size distribution: primate city distribution, rank-size distribution and the medium distribution between them. In order to avoid the defect of data, the paper uses both the population and built-up district area data of above county level cities in Guangdong Province in 2002 and 2004. The results prove that urban size of the urban system in the province follows rank-size distribution. In analyzing the value and the change of Zipf exponents, we point out the characteristics of the size distribution of urban system in the province. The paper also conducts spatial distribution analysis for urban system in Guangdong Province. Power law distribution is also observed, and value of the exponent belongs to the same range of value with those in the size distribution analysis. The differences of focal point between fractal dimension (the inverse of Zipf exponent) and power law exponent (prevailing in complex network analysis) are finally discussed.

Keywords: Power law distribution, urban size distribution, urban spatial distribution, complex networks.

1 Introduction

Guangdong Province is situated in the south of China P.R.C, adjacent to Hong Kong and Macao, with a total area of 1,797.57 hectares. According to the National Census report of the year 2005, there are 91.85 million permanent residents in Guangdong Province at 0:00, November 1st of the year 2005. Compared with the 86.42 million of that in the year 2000 in the fifth National Census, the population is increased by 5.43 million, amount for 6.28%, increasing 109 million per year and the annual grow rate is 1.23%. Until the end of 2004, there are totally 21 prefecture-level cities, 23 county-level cities, 41 counties and 3 autonomous counties in Guangdong Province.

Since 1995, Guangdong accelerated its steps in urbanization. At the end of 2005, the urban population of Guangdong Province reached 55.73 million, amount for 60.68% of the total population. However, there is great difference among different region in the province. The Pearl River Delta has high urbanization level with prominent urban cluster appearance, while for other regions

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 82–89, 2007.
© Springer-Verlag Berlin Heidelberg 2007

in the province, especially the northern mountain area, gaps are large compared to the former. The new development of city systems in Guangdong needs to be analyzed scientifically by using international common practice.

2 Urban Size Distribution Model and Spatial Distribution Model

Zipf distribution. We know that the larger the city, the smaller the amount, while small cities are huge in quantity. In 1949, G. K. Zipf [1] suggested the rank-size rule, which measures the city size distribution in a city system in terms of the relationship between the size and the rank of the city, and argued the power law distribution between the two variables

$$P_R = P_1/R. \tag{1}$$

In formula (1),P_R denotes the population of the city, P_1 the population of the largest city theoretically, and R the rank of city with P_R. More generally, the Zipf's law can be denoted as

$$P_i = P_1 \cdot R_i^{-q}, \tag{2}$$

where P_i denotes the population of city i, P_1 the population of the largest city, R_i the rank of city i, and q is a constant.

Pareto distribution. If we use a population yardstick r to depict the size of cities, and let $P(r)$ denote the probability of $S > r$ (i.e., a city having a population above r), $P(s)$ denote the probability of city size distribution, then we get

$$P_r \propto \int_r^{\infty} P(s)ds. \tag{3}$$

If for any $\lambda > 0$, $P(\lambda r) \propto \lambda^{-a} P(r)$, where λ denotes the proportion of scale, the scaling exponent, then we get the Pareto distribution of city size [2]

$$P_r \propto r^{-a}, \tag{4}$$

where $P(r)$ is the cumulative percentage of cities above threshold r.

If the total quantity of cities is N_0, let $P(r) = N(r)/N_0$, then (4) can be rewritten as

$$N_r \propto C \cdot r^{-D}, \tag{5}$$

where $N(r)$ is the cumulative quantity of cities, and C is a constant [2].

Compared with Hausdorff dimension formula, we know D (equals α) is the fractal dimension [3], [4]. Expression (5) is equivalent to the general form the Zipf formula

$$P(K) = P_1 K^{-q}. \tag{6}$$

In formula (6), K denotes the rank of cities, q the Zipf exponent. Theoretically, $q = 1/\alpha$ has the fractal dimensional properties, therefore it is called the Zipf dimension.

Spatial distribution model. The spatial correlation function can be expressed in the formula

$$C(r) = \frac{1}{N^2} \sum_{i}^{N} \sum_{i}^{N} H(r - d_{ij}), \text{where } i \neq j \tag{7}$$

$$\text{and } (r - d_{ij}) = \begin{cases} 1, d_{ij} \leq r \\ \\ 0, d_{ij} > r \end{cases}.$$

Here r is an yardstick, N is the total amount of cities, d_{ij} represents the Euclidian distance (or, Crow Distance) between city i and j , and H is Heaviside function[5]. To facilitate calculation, formula (7) can be modified to

$$C(r) = \sum_{i}^{N} \sum_{i}^{N} H(r - d_{ij}), \text{where } i \neq j. \tag{8}$$

If there is scale invariance in the spatial distribution of urban system, then

$$C(\lambda r) \propto \lambda^D C(r),$$

$$C(r) \propto r^D, \tag{9}$$

where D is the fractal dimension of the urban spatial distribution.

Formula (9) suggests that in an urban system, the frequency for distance between cities to be below r follows power law distribution. D represents the extent of differences in the distance between cities, or the extent of balance in the urban system spatial distribution.

Here follows the analysis of size distribution and spatial distribution of Guangdong urban system.

3 Size Distribution of Urban System in Guangdong Province and the Search for the Exponent

3.1 Calculation Method and Data Source

Calculation method. There are two ways to measure the urban size. One is to measure the size of urban population. In China, the urbanization level is generally measured by the proportion of non-agricultural population in the total population in the city, so the non-agricultural population is generally used to measure the urban size. Another method is to use the built-up district area as the measure.

The traditional rank-size distribution always uses the former method. Empirical studies show that research based on the more stable non-agricultural population data can better reflect urban size and its growth speed. However, the defects are also obvious. In the case of cities which have more floating population

than that of the permanent residents, such method can hardly reflect actual situation. For example, the floating populations in Shenzhen and Dongguan take large proportion of the total population. In statistical data from the sampled National population survey of 2005, for Shenzhen in 2004, the non-agricultural population is 1.6477 million, with registered population 1.6513 million, while the actual population (including floating population) in the city reached 8.2775 million. Therefore, merely using non-agricultural population may have high inaccuracy. The built-up district area appears stable in the development of a city, and it can better reflect the scale of the city since its statistics are less interfered by other factors. For this reason, some foreign scholars used the built-up district area in the rank-size analysis to study the development of urban cluster. When applying the rank-size rule to analyze the development of built-up district in Berlin, Schweitze etal pointed out that urban clusters are composed of many built-up districts with different area, and that the area and the rank of the city follows rank-size rule [6].

Considering the applying contexts and advantage of both methods, based on the current population situation in Guangdong Province, we use both methods to analyze the urban size distribution in the province, and the results are then compared.

Data source. The study focuses on the above county-level cities in Guangdong Province in 2002 and 2004. Data are collected from the Guangdong Statistical Yearbook. The population data concern the registered non-agriculture population. Because the built-up district areas of below prefecture-level cities are not recorded, we only use the data of above prefecture-level cities in the analysis, not including the counties and towns administrated by them.

3.2 Characteristics of the Distribution and the Exponents

According to the data collected, cities are ranked by their urban populations and built-up district areas, respectively. The results are log-log plotted in Fig. 1 and 2. From the figures, we see that all the distributions can be fitted by a straight line. Using formula

$$\ln P(K) \propto -q \ln K, \tag{10}$$

we can get the Zipf dimension q, see Tab. 1.

Table 1. Main indices of urban size distribution in Guangdong Province

Measure	Year	Quantity of cities	q	R	D
Population	2002	92	0.825	0.954	1.212
	2004	93	0.923	0.957	1.083
Built-up district area	2002	21	0.773	0.969	1.294
	2004	21	0.764	0.976	1.309

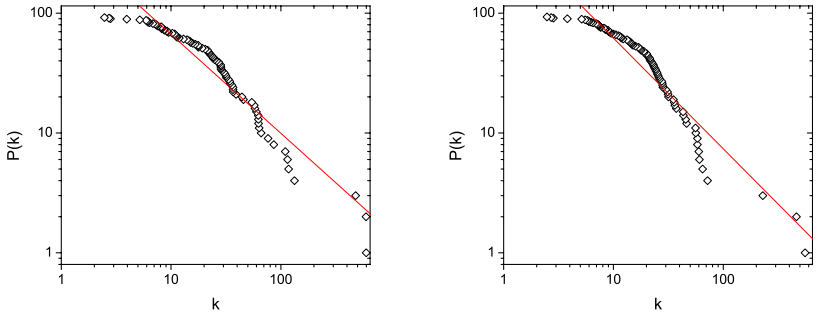


Fig. 1. Log-log rank-size distribution of urban population in Guangdong Province (k: rank of the cities; P(k): urban populations of the cities). (a) in the year of 2002; (b) in the year of 2004.

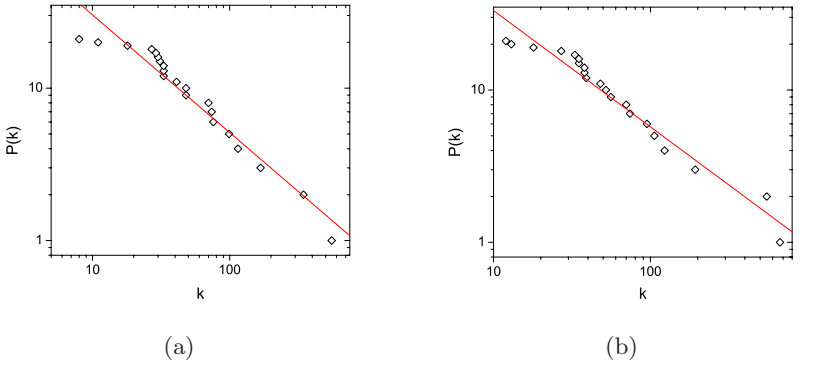


Fig. 2. Log-log rank-size distribution of built-up district areas in Guangdong Province (k: rank of the cities; P(k): urban population of the cities). (a) in the year of 2002; (b) in the year of 2004.

From the analysis above, it is proved that the urban sizes of Guangdong Province follow size-rank distribution in terms of both urban population and built-up district area, and Zipf exponents are all below 1. From 2002 to 2004, the population index increased, while the area index decreased.

4 Spatial Distribution of Urban System in Guangdong Province and the Search for the Exponent

The spatial distribution of major cities takes the above prefecture-level cities in Guangdong Province. In consideration of the particularity of the administrative

division in Foshan, two districts administrated by it (i.e., the Sanshui district and the Shunde district) are treated as cites. Then there are totally 23 cities.

In studying the spatial distances between cities, the Googly-Earth Satellite software is used to measure the accurate distances between city i and j, denoted by d_{ij}, which are filled into a 23×23 symmetric matrix. By assigning r with the multiple of 30 (i. e., $r = 30, 60, 90, \cdots, 720$ kilometers), we may obtain a series of corresponding $C(r)$ following formula 8, as listed in Tab. 2. The data are plotted and fitted on a log-log scale (see Fig. 3).

Table 2. Distance r and correlation function $C(r)$

r	30	60	90	120	150	180	210	240
$C(r)$	10	40	94	132	178	222	258	286
r	270	300	330	360	390	420	450	480
$C(r)$	310	342	362	418	442	452	464	474
r	510	540	570	600	630	660	690	720
$C(r)$	478	488	490	492	498	498	504	506

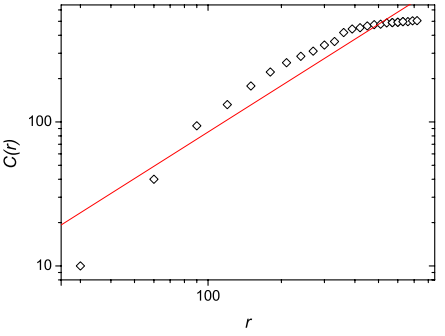


Fig. 3. Log-log plot of correlation function $C(r)$ *vs.* distance r

The regression equation is $\ln C(r) = -0.21 + 1.07 \ln r$, with correlation coefficient 0.91 and power law exponent (i.e. the fractal dimension) 1.07.

5 The Implication for the Power Law Distribution of Guangdong Urban Systems

According to existing research, all the early observed urban rank-size distributions have $q = 1$, and urban size distributions in many developed countries have dimensions approaching 1. While for developing countries, the dimensions are far from 1. That's why the science of urban planning takes 1 as the natural or

ideal Zipf exponent. Therefore, whether q equals 1 should be used to assess the rank-size distribution.

From Tab. 1, we obtained the q, R and D values for the size distribution of urban system in Guangdong Province. In terms of the science of urban planning, the distribution of population is more ideal than that of the built-up district area. The urban population distribution is getting more reasonable, while that of the built-up district area has an opposite tendency.

In terms of the complex network studies in recent years[7], cumulative power law exponent between 1 and 2 results in the coexistence of small amount of large-sized nodes and large quantities of small-sized nodes. When the value is below 1, there are quite a few large-sized nodes. In the case of study for Guangdong urban systems, the distribution follows the former. That is, a few large cities coexist with a large amount of small cities.

The fractal dimension of Guangdong urban system spatial distribution D is 1.07, even smaller than that of the population. It suggest that the variance of urban spatial distribution is greater than those of the size distributions measured by population and built-up district area.

From 2002 to 2004, the power law exponent of population distribution is increasing with that of the built-up district area distribution decreasing, suggesting the trend for the variance of population size to decrease and that of the built-up district area size to increase. The distribution exponent of population to approach 1 suggests the evolution trends for the population size distribution to getting closer to those in the developed countries and regions.

6 Conclusion and Discussion

To summarize, both the urban size and the spatial in Guangdong Province follow power law distribution, with the accumulative power law exponent between 1 and 2. It suggests the self-organized critical state [8] of urban distribution in the Guangdong urban systems, with prominent fractal characteristics, and with the coexistence of a handful of large cities with large amount of small cities.

Because floating populations are not included in the study (without access to obtain the data), there must be errors in the conclusion. Therefore, size distribution considering built-up district area should also be combined in the analysis.

We previously pointed out that the Zipf exponent q is the inverse of Pareto exponent (i.e. D), and Pareto distribution has the same implications with cumulative degree distribution in complex network. In complex network studies, networks with cumulative distribution exponents between 1 and 2 are concerned most, because their robust yet fragile feature can be significant in practice. However, in the science of urban planning, whether the Pareto exponent (i.e. D) equals 1 is concerned most(compared with urban systems having D between 1 and 2, they have more large size cities). The practical significance and theoretical context of this difference, and why urban systems with $D = 1$ are the most reasonable need more further research to be done.

References

1. Zipf, G. K.: Human Behavior and the Principle of least effort, Addison-Wesley, Reading, MA (1949)
2. Shiode, N., Batty, M.: Power Law Distributions in Real and Virtual Worlds, INET'2000 (2000)
3. Mandelbrot, B.B.: The Fractal Geometry of Nature. W. H. Freeman and Company, New York (1983)
4. Batty, M., Longley, P.A., Fotheringham, A.S.: Urban growth and form: scaling, fractal geometry, and diffusion-limited aggregation. Environment and Planning A, **21** (1989) 1447–1472
5. Liu, J.S., Chen, Y.G.: A study on Fractal Dimension of spatial Structure of Trasport Networks and the Methods of Their Determination. Acta Geographica sinica, Vol. 54, **5** (1999)
6. Schweitzer, F., Steinbrink, J.: Estimation of megacity growth: simple rules versus complex phenomena. Applied Geography, (1998) 18(1): 69–81
7. Barabási, A. L., Albert, R.: Emergence of scaling in random networks. Science Vol. 286 (1999) 509
8. Portugali, J.: Self-Organization and the City. Springer-Verlag, Berlin (2000)

Emergence of Social Rumor: Modeling, Analysis, and Simulations

ZhengYou Xia and LaiLei Huang

Department of computer science, Nanjing University of Aeronautics and Astronautics
Zhengyou_xia@yahoo.com

Abstract. We define the notion of social rumor in a standard game-theoretic framework, and assume each agent in the rumor game with individual rationality. In this framework, individual agent can interact with its neighboring agents, and word-of-mouth communication is employed during interaction. We introduce a simple and natural strategy-select rule, called behavior update rule (BUR). The BUR uses an accumulative influence force (CIF) with considering the authority influence of neighboring agents rather than simple accumulative number of information from neighboring agents. The BUR can provide rules to restrict agents' behavior to one particular strategy and lead to emergence of social rumor or anti-rumor. Most importantly, we give simple natural rules of rumor and anti-rumor information transmission, and investigate the efficiency with which social rumors and anti-rumors (agent claims that rumor information is false or doesn't exist) are achieved.

1 Introduction

Emergent behavior is a key topic in artificial life research, given that artificial life typically adopts a bottom-up approach to modeling various forms of collective behavior, belief and emergent social phenomena that are observed in the societies [1][8]. Society is regarded as emerging from interactions among individuals (agents). The action of an agent is social if it is performed towards another agent with purpose, considering the other agent also as a purposeful entity. To model social action it is necessary to go beyond the individual (single agent) level of analysis to reach a multi-agent notion of social intelligence. A more general problem is posed by the simulation studies on social dynamics, namely how given social patterns emerge from autonomous agents in a common artificial world [2] [3] [4].

Rumors are social properties and part of our everyday life. Rumors can shape the public opinion of a society by affecting and coordinating the individual beliefs of its members. Related researches about rumors include work in economics, sociology and psychology, etc. Economists [5] have looked at rumors, both from a theoretical and an empirical point of view that are focused on rumor, price and market selection. In the sociology, psychology, and policy management domain, researches about rumors are focused mainly on the effect of rumors on management, policymaking, the role of uncertainty, anxiety, and psychology of rumors [5,6]. Recent researches about rumor

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 90–97, 2007.
© Springer-Verlag Berlin Heidelberg 2007

propagation have been achieved by Zanette[7][8]. In their paper, epidemiological SIR model is employed and each element of N-element population adopts one of three possible states. To use epidemiological SIR model cannot well model some properties of social rumor. For example, how to model social rumor and anti-rumor that are coexistent in the same social network? How to model the evolution of each agent' belief about rumor (or anti-rumor) and the strategies of each agent in the rumor or anti-rumor environment? In other words, previous model [7][8] is focused on rumor propagation mechanism and cannot validly deal with the evolution about belief (or strategy) of each agent in the coexistence of rumor and anti-rumor environment.

Addressing the above problems, we present a simple model and some rules to discuss the emergence of social rumor and anti-rumor by word-of-mouth communication. We first model social rumor information by using game theory, and then discuss the emergence of social rumor and anti-rumor based on multi-agent systems. Roughly speaking, this process we aim to study is one in which individual agent interacts with one another, and information is transmitted with word-of-mouth communication. Based on its personal accumulated information from other neighboring agents, each agent updates its belief about information over time.

2 Games, Social Rumors and Agents

In this section, we lay out the framework, starting with the standard game-theoretic notion, and overlaying those with the notions of social rumors and agent.

2.1 Games

Game theory [9] is a branch of mathematical analysis developed to study decision making in conflict situations. A game involves a number of players, each of which has available to it a number of possible strategies. Depending on the strategies selected by each player, they each receive a certain payoff. As rational players, they always try to maximize their payoff. In this article, we adopt without change the notions of games, payoff matrices, and rationality as utility maximization that are described in the game theory.

Rumors are part of our everyday life, e.g. "price of butter will run up in one month". Related definitions about rumor have been described by different papers [5][6]. We adopt the standard definition and focus it on the multi-agent systems environment.

Definition 1. Rumor Information (RI): A piece of news (or information) that is passed from agent to agent through word-of-mouth communication and which may or may not be true.

In this article, we assume the initiator of rumor information will get some payoff (money or spirit satisfaction) when he generates rumors. The content of rumor information is assumed to affect the benefit of agents (money or spirit). Therefore, the initiator of rumor and individual (agent) that he hears this rumor will form a two-person general-sum game.

Definition 2. Rumor Game: The normal form of a rumor game is as follow.

$$\begin{array}{cc} & \begin{array}{cc} \multicolumn{2}{c}{Agent} \\ S1 & S2 \end{array} \\ \begin{array}{c} Initiator \\ \\ \end{array} & \begin{array}{c} rumor(true) \\ rumor(false) \end{array} \begin{pmatrix} A(true,S1), B(true,S1) & A(true,S2), B(true,S2) \\ A(false,S1), B(false,S1) & A(false,S2), B(false,S2) \end{pmatrix} \end{array}$$

Where, S1 and S2 is denoted as strategy 1 and strategy 2 of agent, respectively. A and B is denoted as real –value function of initiator and agent, respectively.

In the rumor game, initiator has two kinds of strategies, which is information that he has generated is true and false. We suppose individual (agent) has two strategies that are S1 or S2 when agent receives rumor information. We suppose that agent is rational and rumor game has no dominated strategy for agents. When an agent believes that rumor information is true, It will use one strategy to maximize itself payoff (i.e., $max\{B(true,S1), B(true,S2)\}$). When the agent doesn't believe that rumor information (i.e., agent believe that information is false), he will choose one strategy to maximize itself payoff (i.e., $max\{B(false,S1), B(false,S2)\}$). When belief of agent for rumor information is uncertainty, agent will use mixed strategies to play rumor game.

2.2 Agent and Emergence of Social Rumors

In human societies, each agent has its social authority (or position). Generally, an agent with low authority (position) may be liable to believe information said by the agent with high authority (position). In other words, when agent i transmits information to agent j, the belief of agent j about this information is affected by the social authority of agent i.

Definition 3. Social Authority of agent i can be: $A_i \in [0, \phi]$, where ϕ is a natural number.

$A_j > A_i$ means that the authority of agent j is superior to agent i. Suppose agent i transmits information to agent j, How to measure the influence force that social authority of agent i affects the belief of agent j about this information?

Definition 4. Influence Force (IF): Belief of the agent i affected by social authority of neighboring agent j can be a function: $IF_{ij} = Inf_type * e^{k(A_j - A_i)}$,

Where, A_i, A_j is social authority of agent i and agent j, respectively. k is coefficient and $k \in [0,1]$. k is denoted as the degree that agent i ratifies authority. If k=1, it means that agent i ratifies fully authority of social agent. Similarly, if k=0, it means that agent i doesn't ratify authority of social agent, in other words, the agent i thinks that its social authority is equal to agent j's. In simulation experiment of this paper, the default value of k is 0.5. Inf_type is type of information received by agent i. We assume two kinds of information types (Inf_type) that are rumor and anti-rumor

information. In this paper, we use "+1" and "-1" to denote as type of the rumor and anti-rumor information, respectively.

In the multi agent systems environment, agent i will move and interact with m different neighboring agents, which transmit information to agent i. **Cumulative Influence Force (CIF)** of agent i affected by m different neighboring agents can be computed by the following equation.

$$CIF = \sum_{j} Inf_type * e^{k(A_j - A_i)} \tag{1}$$

where , j be m different neighboring agents

In human societies, when an agent i receives information from agent j, the agent i may not believe this information. However, when agent i receives enough more same information from different neighboring agents, he will believe fully this information. Evolution of belief about information can be described by **Belief function** (Ψ) of agent i.

$$\Psi(information) = \begin{cases} CIF/\vartheta & 0 \le |CIF| < \vartheta \\ 1 & |CIF| > \vartheta \end{cases} \tag{2}$$

Where, ϑ is the threshold of agent i and each agent may have different thresholds.

In the above equation 2, If the value of **|CIF|** is equal or bigger than the threshold of agent (i.e., $\Psi(information) = 1$) and CIF is negative, agent fully believes the anti-rumor information and selects one strategy to maximize itself payoff. If the value of |CIF| is equal or bigger than threshold of agent (i.e., $\Psi(information) = 1$) and CIF is positive, agent fully believes the rumor information and selects one strategy to maximize itself payoff. If |CIF| is less than the threshold of the agent, the agent will use mixed strategies because the agent is uncertain about rumor or (and) anti-rumor information.

Definition 5. [Emergence of Social Rumor]: A belief about rumor information that restricts the agents' behavior to one particular strategy is called emergence of social rumor.

3 Behavior Update Rules

Before we discuss behavior update rule, we assume that each agent has following properties.1). Each agent is rational, which means that each agent will select one strategy to maximize itself payoff.2). Each agent can move randomly in some scope. Each agent moves and interacts with neighboring agents.3). Each agent uses word-of-mouth as communication method with neighboring agents.

We classify agents into three types, which are garrulous agent, close-lipped agent, and dumb agent. Although garrulous agent doesn't know whether information is true or not, it always likes to propagate information to their neighboring agents. The close-lipped agent is different to garrulous agent. The close-lipped agent says something that he will believe and doesn't say something that he doesn't believe. The dumb

agent doesn't say anything to others agent in any case. In this paper, we omit to consider dumb agent because the number of the dumb agent is very few in practical society

When rumor information is spread among agents, due to the lack of evidence that is typically involved in the transmission of a rumor (agents don't know whether the rumor is true or not), the probability with which an agent believes whether a rumor is true depends on the number of neighbors that communicate the same rumor information (it is not anti-rumor information). Intuitively, if you hear a story once, you may believe it or not, but if you hear it also from several persons with the same story, you become to believe that the story is true. Therefore, when every neighboring agent communicates rumor information with an agent, the probability that the agent becomes to believe the rumor increases. We use the notion of **Cumulative Influence Force (CIF)** to define behavior the update rule as following.

Definition 6. [Behavior Update Rule (BUR)]. When an agent hears information from neighboring agents, the agent **believes fully** the information and **selects one new strategy** iff absolute value of **cumulative influencing force** (**CIF**) of the agent about the information is equal or bigger than the **threshold** of the agent.

Behavior Update Rule (BUR) is a simple, natural and local update rule. In particular, it would be natural to consider update rule that uses an accumulation with considering authority influence of neighboring agents rather than simple accumulative number of information from neighboring agents.

4 Experimental Results

We describe a series of experiments in which we investigate the emergence of social rumor and anti-rumor based on multi-agent systems environment. 50 X 50 Grid and 200 agents are composed of simulation society and some agents are distributed randomly in the grid. Each agent can move random one step each time in the grid. Value of the threshold and authority of each agent in 50 X 50 grids are generated randomly.

1) Emergence of social rumor and anti-rumor

Fig1 shows comparison between the speed of emergence of social rumor and anti-rumor in the sparse environment. At the same experiment parameters, rumor and anti-rumor is emerged at about 171 and 65502 times of Iterations, respectively. Fig1 illustrates the practical social phenomenon, which is that rumor is easily emerged and anti-rumor is very difficult emerged. In the Fig1a, in the initial and middle phase, the curve is sleekly exponential increase and curve surges at final phase. The phenomenon can be explained by the following reasons. Agent will spend some time meeting its neighboring agents and it is difficult to make agents with big threshold and high authority to believe immediately rumor information. Although the above phenomenon exists, number of agents that believe rumor information mustn't decrease with increase of times of iterations in the rumor environment.

In the coexistence of rumor and anti-rumor environment, one agent may change belief from believing rumor to disbelieving rumor (believe anti-rumor) at one time when it hears anti-rumor information from neighboring agents. However, when the agent moves to another place, it may change from believing anti-rumor information to believing

rumor information at this time when neighboring agents of this agent are all to believe rumor information. Therefore, number of agents that believe anti-rumor information may be decrease with increase of times of iterations at one time. The above phenomenon also often happens in the practical society. Fig1b show the process of emergence of social anti-rumor and verifies above practical social phenomenon.

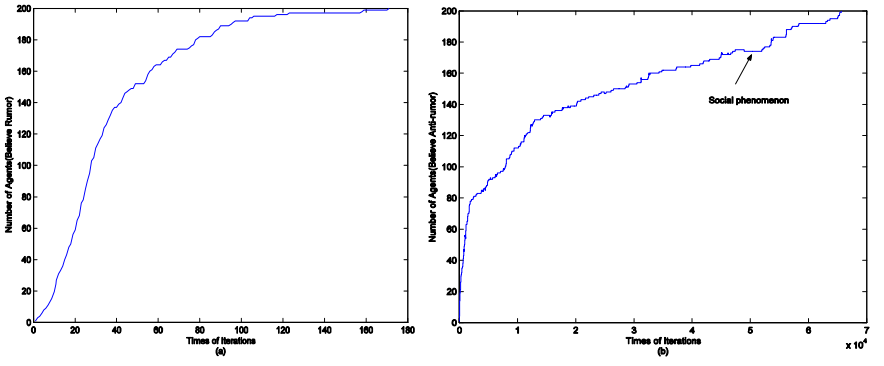


Fig. 1. Emergence of social rumor and anti-rumor :(a)Rumor. (b)Anti-Rumor

2) State of agent under coexistence of rumor and anti-rumor environment

When rumor and anti-rumor information are propagated together in practical society, there is one of three states for each person, which is to believe rumor, to disbelieve (to believe anti-rumor) and be uncertain for rumor and anti-rumor. Fig2 shows states of agents at 15000 times of iterations.

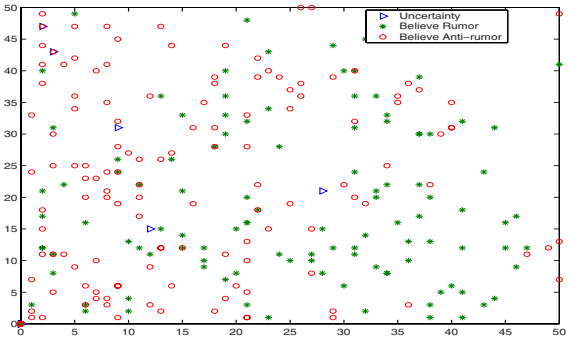


Fig. 2. State of agent

3) Belief evolution of agent

Let p be in (0,1) and n be in (-1,0). n and p are all denoted as uncertain belief of agent about information. In the rumor environment, evolution of agent's belief about rumor information is simple from 0 to 1. Evolution paths of agent' belief only include two paths, which are $0 \to 1$ and , $0 \to p \to 1$. The above phenomenon is

showed as Fig3a. However, in the coexistence of rumor and anti-rumor environment, Evolution of agent' belief about information is different to that in the rumor environment and is more complex(see Fig3b). Typical evolution paths of agent belief about information are as follows:

1. $1 \to -1$: When an agent that fully believes rumor information hears anti-rumor information from neighboring agent(s), the agent change its behavior to full believe anti-rumor information.
2. $1 \to n \to -1$: An agent full believes rumor information at one time. When the agent moves to one place and hears anti-rumor information from neighboring agent(s) at period of time, it will be in the uncertain belief for rumor and anti-rumor information. With increase of neighboring agents that believe anti-rumor information, the agent finally fully believes anti-rumor information.
3. $1 \to p \to -1$: The evolution path of agent belief is similar to the second evolution path
4. $1 \to p \to n \to -1$: The fourth evolution path is similar to the second and third evolution paths and evolves slowly than the above two paths.

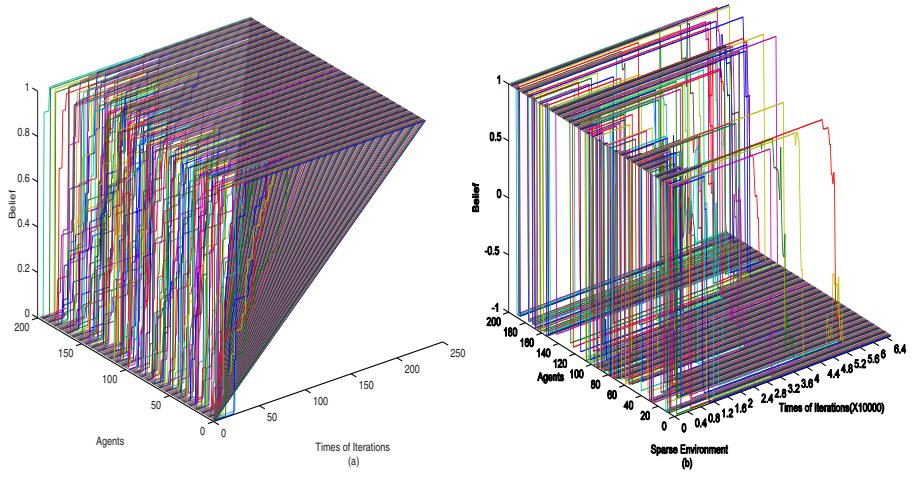


Fig. 3. Evolution of agent' belief in two kinds of environments :(a).Rumor environment. (b)Coexistence of rumor and anti-rumor environment.

There are plenty of more complex evolution paths except for aforementioned typical evolution paths in the coexistence of rumor and anti-rumor environment e.g., $1 \to -1 \to n \to p \to 1 \to p \to -1$. The above evolution path may be following situation. In the coexistence of rumor and anti-rumor environment, one agent changes from believing rumor to disbelieving rumor (believe anti-rumor) at one time when it hears anti-rumor information from neighboring agents. However, when the agent moves to another place, it may change slowly from believing anti-rumor information to believing rumor information at this time when neighboring agents of this agent are

all to believe rumor information. If the agent moves to new place and meets neighboring agents that contain anti-rumor information, the agent may become to believe anti-rumor. Fig3b illustrates different evolution paths of agent for information in dense and sparse environment, respectively.

5 Conclusion

From economist point of view, rumors were often thought to be something rather irrational based on market behavior. However, from AI and mathematics (game theory) point of view, rumors were often thought to be rational because rumors are emerged by behavior of these agents who always try to maximize themselves payoff. In this paper, we used the framework of rumor game in order to investigate the emergence of social rumors and the efficiency of that process. In the framework, each agent can move randomly and interact with neighboring agents, and information is transmitted with word-of-mouth communication. We borrow ideas of reinforcement learning in order to capture the fact that agents use local update rules (i.e., to compute cumulative influence force (CIF)) to update their strategies. We make plenty of experiment to further discuss emergence of social rumor and anti-rumor.

Acknowledgements

This paper is based upon work supported by the JiangSu Province Science Technology Foundation under Grant No. BK2006567.

References

1. Cristiano Castelfranchi, modeling social action for AI agents, artificial intelligent, 103 (1998) 157-182.
2. Rosaria Conte, Social intelligence among autonomous agents, computational mathematical organization theory 5:3 (1999) 203-228
3. J Delgado, Emergence of social conventions in complex networks, Artificial Intelligence, 141 (2002) 171--185
4. PC Buzing, AE Eiben and MC Schut, merging communication and cooperation in evolving agent societies, Journal of Artificial Societies and Social Simulation,
5. Grant Michelson, Suchitra Mouly, Rumour and gossip in organisations: a conceptual study, Management Decision, Jun 2000 Volume: 38 Issue: 5 Page: 339 – 346.
6. Michael kosfeld, rumors and markets, journal of mathematical economics, 41 (2005) 646-664
7. Zanette, Damian H., Dynamics of rumor propagation on small-world networks, Phys. Rev. E 65, 041908 (2002)
8. Zanette, Damian H., Critical behavior of propagation on small-world networks, Phys. Rev. E 64, R050901 (2001)
9. D Fudenberg, J Tirole, Game theory, MIT Press Cambridge, Mass (1991)

Emergence of Specialization from Global Optimizing Evolution in a Multi-agent System

Lei Chai, Jiawei Chen, Zhangang Han, Zengru Di, and Ying Fan*

Center for Complexity Research,
Beijing Normal University, Beijing, 100875, P.R. of China
Institute of Social Development and Public Policy,
Beijing Normal University, Beijing, 100875, P.R. of China
{lchai,jwchen,zhan,zdi,yfan}@bnu.edu.cn
http://www.springer.com/lncs

Abstract. A Markov chain model is proposed to describe the evolutionary dynamics of a multi-agent system. Many individual agents search for and exploit resources to get global optimization in an environment without complete information. With the selection acting on agent specialization at the level of system and under the condition of increasing returns, agent specialization emerges as the result of a long-term optimizing evolution.

Keywords: agent specialization, evolutionary dynamics, multi-agent system.

1 Introduction

Among novel and numerous varieties of collective activities performed by social insect societies, such as all ants and termites, and some species of bees and wasps [1], division of labor is a typical example which is classically viewed as an evolving process led by mutation and selection, and also has emergent properties of social systems. A lot of work has been done to study the formation of division of labor and the mechanism for tasks allocation [2,3,4,5]. In this paper, multi-agent system simulation has been applied and a framework of pattern formation has been borrowed to study specialization phenomena in biological, social and economic systems.

The emergence of collective behaviors in multi-agent systems has become an interesting area of complexity research [6,7,8,9,10,11]. In a distributed multi-agent system, be it natural or social, agents do not have complete information about the environment where they live, and they have to actively interact with others to reach collective or individual optimization. Even though the interactions among agents may be simple and local, they can lead to complex dynamics at the global scale. Studies of computational ecology have shown that when the agents make choices with imperfect information, the dynamics can give rise to nonlinear oscillations, clustered volatilities and chaos that drive the system

* Corresponding author.

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 98–105, 2007.
© Springer-Verlag Berlin Heidelberg 2007

far from optimality [12,13]. Moreover, when the agents in the system have a hierarchical organization that determines the task allocations, the system as a whole will be able to manage complex and various changing environments [2,14]. Holland has argued that complexity arises from the self-adaptive properties of individual agents [15] and this approach has been applied to biological ecosystems and economic systems [16,17,18].

Evolutionary processes and principles are helpful to understand the mechanisms behind the formation of specialization. As to a long-term evolution, selection acting on agent specialization must take place at the level of the colony. Some colonies survive and reproduce more than others because they have a division of labor that is better adapted to a particular environment [19]. Based on our previous multi-agent simulation model of agent division [20], this paper presents a Markov Chain model that describes a system of many individual agents that search for and exploit resources to get global optimization in an environment where they interact locally. Under the condition of increasing returns, specialization is unfailingly the result from the global optimizing evolution under certain initial conditions.

The presentation is organized into three major parts. In Section 2, the model and the results of static analysis are presented. Section 3 gives the Markov chain model describing the evolutionary dynamics of the system. The final steady state shows the specialization of the agents. The results are well consistent with that of simulation model. In Section 4, we provide a summary of our results and a brief discussion of some unresolved issues.

2 The Model and Static Analysis

Consider a system with M individual agents. Every agent is an autonomous entity which searches for and exploits resources in a given environment. There is complete information transfer among agents. Each agent knows the information about the resource distribution which is specified by every other agent's forage. For simplicity and without loss any generality, we introduce a simple, non-spatial model of evolutionary dynamics to obtain some mathematical analytical results.

2.1 Optimal Behavior for Solitary Agent

First, let's discuss the problem for a solitary agent to deal with an uncertain environment.

The live space for the agent is composed by the samples of resources valued F_0 and 0. In the beginning of every period of time, the agent can choose to search for new resources with probability q or to take the situation of last period with probability $1 - q$. When the agent determines to search new resource, it will get the resource F_0 with probability P. Then its product is $(1 - q)F_0$. The cost for the search is $(1-q)C$, where the factor $1-q$ describes the effect of learning by

doing[21]. When the agent determines to take the last situation, if the last value of the resource is F_{t-1}, the value of this period is

$$F_t = \frac{1}{a} F_{t-1}, a > 1 \tag{1}$$

where a stands for the depreciation of resources and t denotes the generation of the system's evolution. The goal of the agent is to maximize the total net returns for N periods, which is determined by each agent's parameter q for a given distribution of resources that is described by probability P.

In the beginning of searching and exploiting process noted as period 0, the agent randomly searches. For this period, the expected return is $E_0 = (1-q)(PF_0 - C)$. For period 1, the agent can choose to search for new resources with probability q. The corresponding expected benefit is E_0. Another choice is to stay in the situation of last period (with probability $1-q$). The corresponding expected return is PF_0/a. So the expected return of period 1 is

$$E_1 = (1-q)(qE_0 + (1-q)\frac{1}{a}PF_0) \tag{2}$$

Let $x = (1-q)/a$. From the similar analysis, the expected benefit of period n can be written as:

$$E_n = (1-q)(qE_0 + PF_0(x^n + q\frac{x(1-x^{n-1})}{1-x}))$$

Hence the total expected returns from period 0 to period N is

$$W = \sum_{n=0}^{N} E_n = (1+N)(1-q)qE_0 + PF_0(1-q)\sum_{n=0}^{N}(x^n + q\frac{x(1-x^{n-1})}{1-x})$$

$$= (1+N)(1-q)qE_0 + PF_0(1-q)[(1+N)\frac{qx}{1-x} + (1-\frac{q}{1-x})\frac{1+x^N}{1-x}] \tag{3}$$

From the above result, we can obtain the optimal point q_M and corresponding maximum total expected returns W for given set of parameters.

2.2 Optimal Division of Agents for a Colony

Now we turn to the colony of M agents. Each agent is characterized by a parameter q_i, which determines the searching probability of the agent. As discussed in the above solitary agent case, we assume that every searching agent finds resources F_0 with probability P. Assume there is complete information between agents. If any agent has found a resource F_0, all the others will go and exploit it. Then the product of the system is $\sum_{i=1}^{M}(1-q_i)F_0$. The searching cost for each agent is $(1-q_i)C$.

Let's compare the results on the total returns of distributed q_i with complete specialization. Hence at every time period, there will be $\sum_{i=1}^{M} q_i = m$ agents searching for new resources. The gross probability of finding at least one new

resource F_0 is the same with when there is m agents specialized in searching. We denote this gross probability as P_r. For a colony with distributed q_i, the expected product of the first period is $R_D = \sum_{i=1}^{M}(1-q_i)F_0P_r = MF_0P_r - F_0P_r\sum_{i=1}^{M} q_i = F_0P_r(M-m)$. The cost for searching is $C_D = \sum_{i=1}^{M} q_i(1-q_i)C = mC - C\sum_{i=1}^{M} q_i^2$. With $q_i \leq 1$, $\sum_{i=1}^{M} q_i^2$ is usually less than m. So we usually have $C_D > 0$. But for the colony with specialized agents, although the expected product of the first period is the same: $R_S = \sum_{i=1}^{M-m} F_0P_r = F_0P(M-m) = R_D$, hence the products of following generations are also the same, and the cost for search is 0. So the net return of a specialized system is bigger than that of the distributed one.

Assuming that there are m agents specialized in searching in the colony, all the others specialized in exploiting the resource. In each period, If any searching agent has the probability P to find the resource F_0, then the probability for m agents at least find one resource F_0 is $P_r = 1-(1-P)^m$. So the expected returns of the colony for every period are:

$$E_0 = P_r(M-m)F_0$$
$$E_1 = P_r(M-m)F_0 + P_r(1-P_r)\frac{1}{A}(M-m)F_0$$
$$E_2 = P_r(M-m)F_0 + P_r(1-P_r)\frac{1}{A}(M-m)F_0 + P_r(1-P_r)^2\frac{1}{A^2}(M-m)F_0$$
.....

Let $x = (1-P_r)/A$, where parameter $A > 1$ is also related to the diminishing of resource under agents' exploiting. The return of nth period can be written as

$$E_n = E_0(\frac{1-x^n}{1-x}) \tag{4}$$

So the total returns of N periods is

$$W(\{N_j\}) = \sum_{n=0}^{N-1} E_0\frac{1-x^n}{1-x} = \frac{E_0}{1-x}(N - x(\frac{1-x^N}{1-x}))$$
$$= \frac{P_r(M-m)F_0}{1-x}(N - x(\frac{1-x^N}{1-x})) \tag{5}$$

By equation 5, it is found that m_0 is sensitive to the probability P (as shown in Figure 1). The results are in good agreement with previous simulation results.

3 Markov Chain Model for the Evolutionary Dynamics

We assume the character space of the agents has $k+1$ states corresponding to the searching probabilities described by parameter $q_j, j = 0, 1, \ldots, k$, with $q_0 = 0$ and $q_k = 1$. The distribution of agents in every state describes the situation of the colony in macroscopic level. Let's denote this distribution as $\{N_j, j = 0, 1, ..., k\}$, and we have $\sum_{j=0}^{k} N_j = M$. For generality N_j is a positive real number instead

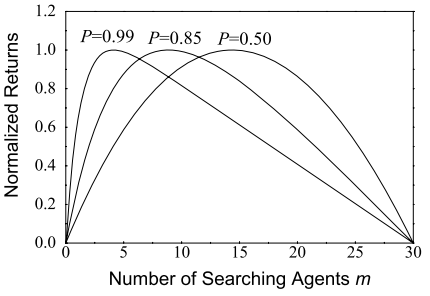


Fig. 1. Returns (normalized) as a function of number of searching agents m. Except parameter P (labeled above the corresponding curve), all the other parameters for each curve are the same.

of a positive integer. From the results in the above discussion, the total product $W\{N_j\}$ and total cost $C_D\{N_j\}$ of the colony in one generation with N periods is

$$W(\{N_j\}) = \frac{P_r(M - \sum_{j=0}^{k} N_j q_j)F_0}{1-x}(N - x(\frac{1-x^N}{1-x})) \tag{6}$$

$$C_D(\{N_j\}) = N\sum_{j=0}^{k} N_j q_j(1-q_j)C \tag{7}$$

where $x = (1-P_r)/A$, $P_r = 1-(1-P)^m$, and $m = \sum_{j=0}^{k} N_j q_j$. The total return of the system in one generation is $R(\{N_j\}) = W(\{N_j\}) - C_D(\{N_j\})$.

Within the evolution process between two generations, the state of every agent will transit among all the $k+1$ states. In the simulation model, we only make it possible for the transition between the nearest neighbors. A Markov chain process can describe the genetic variation and natural selection model. The dynamical behavior of this Markov chain is determined by following equations:

$$\begin{aligned} N_j(t+1) = N_j(t) + P(j+1 \to j)N_{j+1}(t) + P(j-1 \to j)N_{j-1}(t) \\ -P(j \to j-1)N_j(t) - P(j \to j+1)N_j(t) \end{aligned} \tag{8}$$

For $j = 0$ and $j = k$ we have

$$\begin{aligned} N_0(t+1) &= N_0(t) + P(1 \to 0)N_1(t) - P(0 \to 1)N_0(t) \\ N_k(t+1) &= N_k(t) + P(k-1 \to k)N_{k-1}(t) - P(k \to k-1)N_k(t) \end{aligned} \tag{9}$$

where $P(i \to j)$ is the probability for an agent to transit from state i to state j, and it is determined by the global optimization. Corresponding to the natural selection process described in the simulation model, the transition probability can be written as:

$$P(i \to j) = \frac{1}{2}\mu[5 + 3\ \mathrm{sgn}(R(N_i - 1, N_j + 1) - R(N_i, N_j))] \tag{10}$$

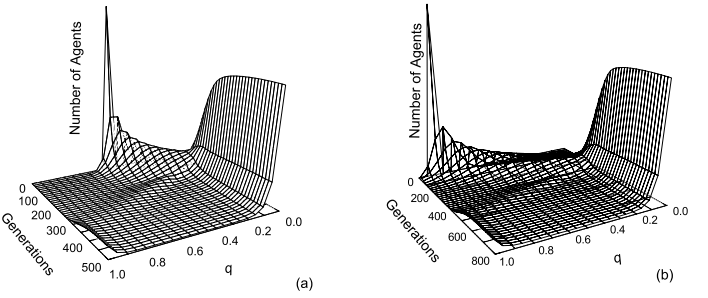


Fig. 2. Evolution of the distribution of agents. The results are qualitative similar to the computer simulations.

where μ is a parameter related to the probability of mutation. From the above equations, we can get the results of the optimal evolution of the system from any given initial conditions, which is shown in Figure 2. The Markov chain model is carried out under the following parameters. $M = 30$ agents form a colony and want to get global optimization in an uncertainty environment. The probability for every searching agent to find resource F_0 is $P = 0.7$. The value of the resource is $F_0 = 10$ and the searching cost is $C = 8$. The other parameters are $\triangle q = q_{i+1} - q_i = 0.05$, $A = 4$, $N = 54$, and $\mu = 0.02$. Given any initial condition, we get the evolution of the distribution of agents. The final optimal

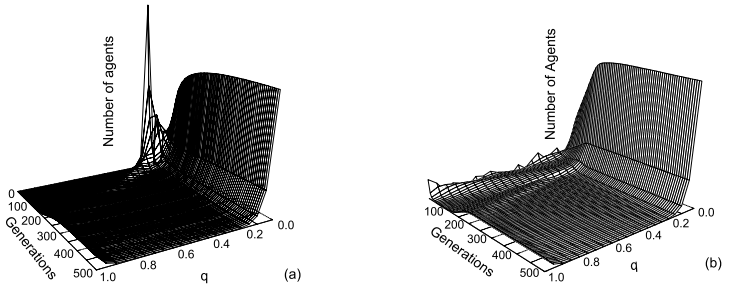


Fig. 3. Evolution from different initial conditions. (a) Homogeneous initial distribution with q_k=0.1 for every agent. (b) Random initial distribution.

distribution in the Markov chain process is stable and it could be achieved from any given initial conditions (See Figure 3). We have compared the final stable distribution of the Markov chain process with the average distribution in the generations of optimal state in our previous simulation model [20]. As shown in Figure 4, it is notable that the mathematical results are consistent well with the results of simulations.

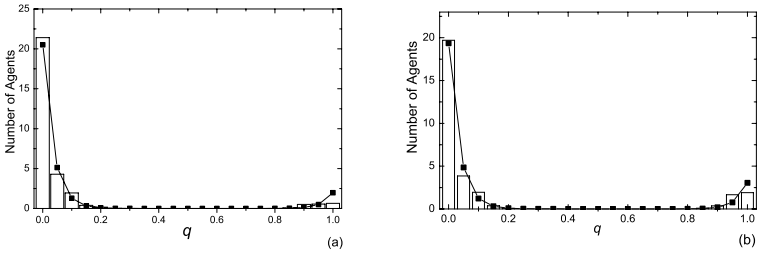


Fig. 4. Comparison of theoretical predictions (solid lines) and simulation results (column bars) of the distributions of agents' number. (a) P=0.7 (b) P=0.4

4 Conclusions

In this paper we have studied the formation of specialization of a simple multi-agent model with optimizing evolution behaviors. With a Markov chain model, the multi-agent system has provided discoveries in a concise way to perceive the evolution of labor division. First, specialization is a functional structure in macroscopic level of multi-agent systems. The model has demonstrated how a global structure emerges from the long-term optimizing evolution under the mechanism of increasing returns. Second, an evolutionary process given by transition behavior describes the mechanism of mutation and natural selection. We believe that natural selection may also serve as the basic mechanism in the labor division in social and economic systems. The last but not the least, the results reveal that the stochastic properties in the evolutionary process are necessary to generate macroscopic structure and to reach global optimization.

This work suggests a number of future directions for the study of the optimal behaviors of multi-agent systems. As mentioned in section 2, the model assumes that there is perfect information among agents and the goal of the system is global optimization. These assumptions in fact have suggested that the agents have already formed a colony. It is interesting to study the relationship between individual behavior and global optimization, to see how organization emerges from individual optimum, and to understand how a multi-agent system forms aggregate units.

Acknowledgments. This work was partially supported by the 985 project and NSFC under grant No.70471080 and No.70371072.

References

1. Theralauz G. Bonabeau E. and Deneubourg, J., The origin of nest complexity in social insects. Complexity **3** (1998) 15-25.
2. Gordon D., The Organization of work in social insect colonies. Complexity **8** (2003) 43-46.

3. Robinson G., Regulation of division of labor in insect societies. Annu. Rev. Entomol **37** (1992) 637-665.
4. Bonabeau E., Theraulaz, G. and Deneubourg J., Quantitative study of the fixed threshold model for the regulation of division of labour in insect societies. Proc. Roy. Soc. **London B 263** (1996) 1565-1569.
5. Wu J., Di Z., and Yang Z., Division of labor as the result of phase transition. Phasica A **323**, (2003) 663-676.
6. Zimmermann G., Neuneier R. and Grothmann R., Multi-Agent market modeling of foreign exchange rates. Advances in Complex Systems **4** (2001) 29-43.
7. Darbyshire, P., Effects of communication on group learning rates in a multi-agent environment. Advances in Complex Systems **6** (2003) 405-426.
8. Juanico, D., Monterola, C. and Saloma, C., Cluster formation by allelomimesis in real-world complex adaptive systems. Phys. Rev. E **71**, (2005) 041905.
9. Taylor, P. and Day, T., Behavioural evolultion: cooperate with thy neighbour? Nature **428** (2004) 643-646.
10. Nowak, M., Sasaki, A., Taylor, C. and Fudenberg, D., Emergence of cooperation and evolutionary stability in finite populations. Nature **428** (2004) 646-650.
11. Weiss G. et al. (eds.), Adaption and learning in multi-agent systems. Springer-Verlag, Berlin (1996) 20-35.
12. Kephart J., Hagg T. and Huberman B., Dynamics of computational ecosystems, Phys. Rev. A **40** (1989) 404-421.
13. Hagg T. and Huberman B., Controlling chaos in distributed systems, IEEE Transactions on Systems, Man, and Cybernetics **21** (1991) 1325-1332.
14. Bonabeau E., and G. Theraulaz, Self-organization in social insects, Trend in Ecology and Evolution **12** (1997) 188-193.
15. Holland J., Hidden Order-how adaptation builds complexity. Addison Wesley: MA (1995) 230-300.
16. Holland J. and Miller J., Artificial adaptive agents in economic theory, American Economic Review **81(2)** (1991) 365-370.
17. Andriani P., Diversity, Knowledge and complexity theory: some introductory issues. International Journal of Innovation Management **5(2)** (2001) 257-274.
18. Savage M., and Askenazi M., Arborscapes: a swarm-based multi-agent ecological disturbance model. Technical Report (STF-TR-98-06-056, Santa Fe Institute, 1998).
19. Fontana W., Buss L., The arrival of the fittest: toward a theory of biological organization, Bull. Math. Biol. **56** (1994) 1-64.
20. Di Z., Chen J., Wang Y., Han Z., Agent division as the results of global optimizing evolution, in; Shi Z., He Q. (eds.), Proceedings of International conferebce on intelligent information technology. Post & Telecom Press Beijing (2002) 40-46.
21. Arrow K.J., The economic implications of learning by doing, Rev. of Economics Stu. **29** (1962) 155-173.

A Hybrid Econometric-AI Ensemble Learning Model for Chinese Foreign Trade Prediction

Lean Yu[1,2], Shouyang Wang[1], and Kin Keung Lai[2]

[1] Institute of Systems Science, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100080, China
{yulean,sywang}@amss.ac.cn
[2] Department of Management Sciences, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong
{mskklai,msyulean}@cityu.edu.hk

Abstract. Due to the complexity of economic system, the interactive effects of economic variables or factors on Chinese foreign trade make the prediction of China's foreign trade extremely difficult. To analyze the relationship between economic variables and foreign trade, this study proposes a novel nonlinear ensemble learning methodology hybridizing nonlinear econometric model and artificial neural networks (ANN) for Chinese foreign trade prediction. In this proposed learning approach, an important econometrical model, the co-integration-based error correction vector auto-regression (EC-VAR) model is first used to capture the impacts of the economic variables on Chinese foreign trade from a multivariate analysis perspective. Then an ANN-based EC-VAR model is used to capture the nonlinear patterns hidden between foreign trade and economic factors. Subsequently, for introducing the effects of irregular events on foreign trade, the text mining and expert's judgmental adjustments are also incorporated into the nonlinear ANN-based EC-VAR model. Finally, all economic variables, the outputs of linear and nonlinear EC-VAR models and judgmental adjustment model are used as another neural network inputs for ensemble prediction purpose. For illustration, the proposed ensemble learning methodology integrating econometric techniques and artificial intelligence (AI) methods is applied to Chinese export trade prediction problem.

Keywords: Artificial neural networks, error-correction vector auto-regression, hybrid ensemble learning, foreign trade prediction.

1 Introduction

Since the reforms and opening policies was initiated in 1978, China has attracted foreign investment, brought in foreign advanced technology, increased commodity exports and accelerated economic growth. Currently, China is the world's fastest developing nation. If it grows at an average of 4% to 6% annually, China will overtake Japan as Asia's largest economy by 2040 [1]. In China's economy, the most dynamic and important part is the external sector. In 2003, the openness of economy (the ratio of foreign trade in GDP) exceeds 60 percent. About one-fourth of China's

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 106–113, 2007.
© Springer-Verlag Berlin Heidelberg 2007

industrial labors are employed by the foreign trade related industries [2]. Thus China's government pays a great attention to foreign trade.

As a transition economy, China's government has been playing a key role in its economic development. For accelerating the development of foreign trade, the government not only has issued a series of favorable policies. One important policy is "export rebate tax". Every March of the year, the Chinese highest legislation authority, the National People's Congress (NPC), will hold its annual grand session. In this session, the delegates will audit the fiscal plan of the central government. The "export rebate tax" was an important part in the fiscal plan. To make a feasible fiscal plan, it is necessary to make an accurate prediction for foreign trade. Therefore, foreign trade forecasting has an important implication for Chinese macroeconomic policymaking. For example, if export trade forecasting is accurate, the national fiscal program about "export rebate tax" is easy to be made and to be implemented. If forecasting is larger than the actual exports, some financial fund will be otiose and cannot exert its function. If forecasting is smaller than the actual exports, the financial subsidies will not be sent to export enterprises and thus affecting their enthusiasm. So, accurate foreign trade forecasting can reduce uncertainty and blindness in macroeconomic policymaking and obtain good economic benefit

However, it is difficult to predict the foreign trade accurately due to the complexity of economic system and interactive effects of the economic variables. For this purpose, this paper attempts to propose a nonlinear ensemble learning methodology hybridizing nonlinear econometric model and artificial neural networks (ANN) for Chinese foreign trade prediction. In this proposed learning approach, an important econometrical model, the co-integration-based error correction vector auto-regression (EC-VAR) model is first used to capture the impacts of the economic variables on Chinese foreign trade from a multivariate analysis perspective. Then an ANN-based EC-VAR model is used to capture the nonlinear patterns hidden between foreign trade and economic factors. Subsequently, for introducing the effects of irregular events on foreign trade, the text mining and expert's judgmental adjustments are also incorporated into the nonlinear ANN-based EC-VAR model. Finally, all economic variables, the outputs of linear and nonlinear EC-VAR models and judgmental adjustment model are used as another neural network inputs for ensemble purpose.

The major aim of this study is to present a new hybrid econometric-AI ensemble learning methodology that can significantly improve the prediction performance and thus improving their effectiveness on complex economic prediction problems. The rest of this study is organized as follows. Section 2 describes the building process of the proposed hybrid econometric-AI ensemble learning method in detail. For further illustration, the proposed ensemble learning methodology is applied to Chinese export trade prediction in Section 3. Finally, some conclusions are drawn in Section 4.

2 The Hybrid Econometric-AI Ensemble Learning Methodology

In this section, an overall formulation process of the hybrid econometric-AI ensemble learning methodology is described. First of all, the co-integration technique and error correction vector auto-regression (EC-VAR) model are briefly reviewed. Then an ANN-based EC-VAR model is presented to explore nonlinear patterns in the complex

economic phenomena. To incorporate the effects of irregular events into prediction, text mining and expert's judgmental adjustments are also considered into the ANN-based EC-VAR model. In order to integrate almost all information and implied knowledge, another neural network model is used to realize the nonlinear ensemble.

2.1 Co-integration Technique and EC-VAR Model

At one time conventional wisdom was that non-stationary variables should be differenced to make them stationary before including them in multivariate models. However, this situation was changed when Engle and Granger [3] introduced the concept of co-integration in 1987. They showed that it is quite possible for a linear combination of integrated variables to be stationary. In this case the variables are said to be co-integrated. Consider a set of variables in long-run equilibrium (static equilibrium) when $\beta_1 x_{1t} + \beta_2 x_{2t} + \ldots + \beta_n x_{nt} = 0$. The equilibrium error is then $e_t = \beta^T X_t$. If the equilibrium is meaningful it must be the case that the error is stationary. For this, co-integration technique is popular and widely used in many domains since 1987. Interested readers can refer to Engle and Granger [3] for more details.

If co-integration relationship between variables exists, an error correction vector auto-regressive (EC-VAR) model can be formulated for prediction purpose. Actually, an EC-VAR model can be seen as a co-integration-based forecasting model, which is represented by

$$\Delta y_t = \sum_{i=1}^{k_0} \alpha_i \Delta y_{t-i} + \sum_{j=1}^{m} \sum_{i=0}^{k_j} \beta_{j,i} \Delta x_{j,t-i} + \gamma \cdot EC_{t-1} + \xi_t \tag{1}$$

where y_t is dependent variable, y_{t-i} is the lag term, x_{t-i} is the lag terms of the independent variables, EC_{t-1} is the co-integration relation, or error correction term, and ξ_t is the Gaussian noise, and α, β, γ are the coefficients of different variables or lag terms, respectively. Generally, an EC-VAR model can lead to a better understanding of the nature of any nonstationarity among the different component series and can also improve longer term forecasting performance over an unconstrained model.

2.2 ANN-Based Nonlinear EC-VAR Model

In the above co-integration-based EC-VAR model, the EC is only a linear error correction term. Usually, the errors of component series contain much nonlinearity, a linear error correction is often not sufficient. For this purpose, Hendry and Ericsson [4] proposed a nonlinear EC-VAR model with the following representation:

$$\Delta y_t = \sum_{i=1}^{k_0} \alpha_i \Delta y_{t-i} + \sum_{j=1}^{m} \sum_{i=0}^{k_j} \beta_{j,i} \Delta x_{j,t-i} + \gamma \cdot f(EC_{t-1}) + \xi_t \tag{2}$$

where $f(\bullet)$ is a nonlinear function, EC_{t-1} denotes a long-term co-integration relationship, and other symbols are similar to Equation (1) . Hendry and Ericsson [4] used a thrice function of EC_{t-1} to predict the currency demand of UK and obtained good performance. In this study, we used ANN to construct a nonlinear function $f(\bullet)$ to create an ANN-based nonlinear EC-VAR model. That is, the proposed ANN-based

EC-VAR model applies the sigmoid function to determine the nonlinear function $f(\bullet)$ as the error correction term. In this paper, the sigmoid function is represented by

$$f(EC) = \frac{1}{1+e^{-(w*EC+b)}} \tag{3}$$

where w and b are unknown parameters, which is determined by ANN training. Usually, the feed-forward neural network (FNN) can realize this goal. Hornik et al. [5] and White [6] found that a three-layer FNN with an identity transfer function in the output unit and logistic functions in the middle-layer units can approximate any continuous function arbitrarily well given a sufficient amount of middle-layer units.

Generally, the ANN-based EC-VAR model is performed in the following steps:

(1) Determine the co-integration relationship between the variables, and

(2) The co-integration relationship is placed into the VAR equation with a nonlinear function form instead of linear function form or a constant. Particularly, this paper uses the sigmoid function as the nonlinear function.

2.3 Incorporating Text Mining and Judgmental Adjustments into the EC-VAR

In the complex economic system, the interactive effects are from various sources. Besides the related variables and nonlinear interactions, some irregular events, e.g., 911 terrorist attack, have an important impact on the world economy. To further improve the prediction performance, the text mining and expert's judgmental adjustment are incorporating into the forecasting model. Generally text mining refers to the process of extracting interesting and non-trivial information and knowledge from unstructured text [7]. Interested readers can refer to Sullivan [8] for more details about text mining. In this study, the main goal of the text mining is to retrieve related information from various sources and the human experts provide the judgmental adjustment information for these important information. Within such a framework, the Equation (2) can be extended as

$$\Delta y_t = \sum_{i=1}^{k_0} \alpha_i \Delta y_{t-i} + \sum_{j=1}^{m} \sum_{i=0}^{k_j} \beta_{j,i} \Delta x_{j,t-i} + \gamma \cdot f(EC_{t-1}) + JA + \xi_t \tag{4}$$

where JA represents the expert's judgmental adjustment for a specified event and other symbols are the same as the above.

2.4 Nonlinear Ensemble Forecasting Model

Considering various different economic factors and judgmental adjustments, we have so far formulated three main forecasting equations. In order to capture the effects of different variables and increase forecasting accuracy, it is important to integrate them into a single more accurate forecast. Suppose that there are k lag terms (y_{t-i}, $i = 1, 2, \ldots, k$), m related variables (x_j, $j = 1, 2, \ldots, m$), and p forecasts provided by different forecasting equations shown in the previous subsections, then nonlinear ensemble model can be represented as

$$\hat{y}_t = \varphi(y_{t-1}, \cdots, y_{t-k}; x_1, x_2, \cdots, x_m; \hat{y}_1, \hat{y}_2 \cdots, \hat{y}_p; w) + \xi_t \tag{5}$$

where w is a vector of all parameters and $\varphi(\bullet)$ is a nonlinear ensemble function.

Determining the nonlinear function $\varphi(\bullet)$ is quite challenging. In this study, a standard FNN is employed to realize nonlinear mapping [9]. Actually the paper uses the k lag terms (y_{t-i}, $i = 1, 2, \ldots, k$), m related variables (x_j, $j = 1, 2, \ldots, m$), and p forecasts as another neural network inputs to construct the nonlinear ensemble model. Fig. 1 gives an intuitive illustration for the proposed nonlinear ensemble model.

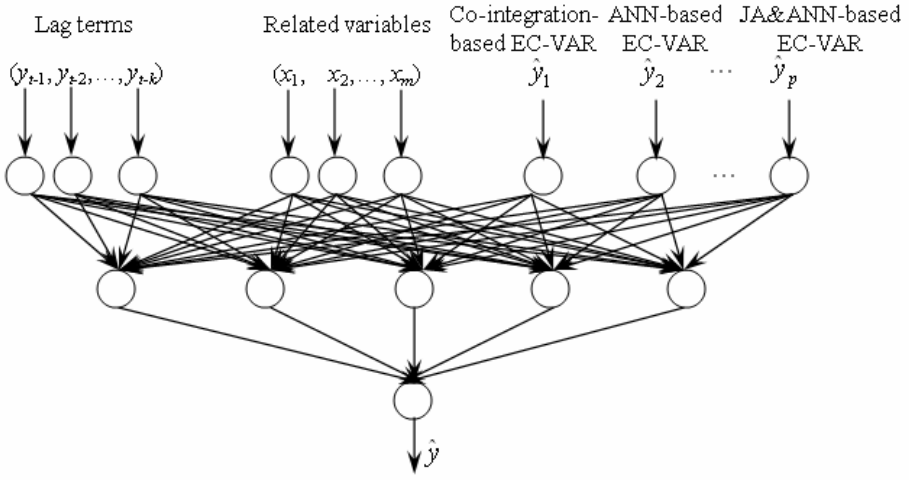


Fig. 1. An intuitive illustration for the nonlinear ensemble forecasting model

It is interesting to examine the underlying idea of the proposed hybrid econometric-AI ensemble learning methodology. For a complex and difficult forecasting problem, one single linear or nonlinear model is inadequate to model it due to interactive effect of multiple factors. Through the linear co-integration-based EC-VAR, ANN-based EC-VAR and judgmental adjustment-based nonlinear EC-VAR models, the linear patterns, nonlinear patterns, and irregular patterns are also explored. In order to formulate a comprehensive understanding, the explored patterns are aggregated into a single prediction, as indicated in Eq. (5).

From the above analysis, it is obvious that the conventional linear econometric models are insufficient for complex prediction problem because some nonlinear and irregular patterns are not captured by them. From this perspective, it is not hard to understand why conventional linear econometric models cannot always understand and predict some complex and difficult economic problem well. In this sense, using such a hybrid econometric-AI ensemble learning methodology, some complex prediction problems are easy to be implemented based on different linear and nonlinear models as well as nonlinear ensemble technique.

3 Experiment

In order to evaluate the effectiveness of the proposed hybrid econometric-AI ensemble learning methodology, a complex time series prediction task – Chinese foreign trade prediction problem – is investigated. Due to the complexity of economic system, the effects of the economic variables or factors on Chinese foreign trade are realized from many different perspectives. For this, we must select some representative variables to construct a model. The following principle is worth considering in the process of model variable selection. (a) Theoretical interpretability. This requires that the selected variable must have an explanation power for model. That is, the variable can explain why it affects the Chinese foreign trade from the theoretical viewpoint. (b) Representative. This requires us to find some representative variables to construct a Chinese foreign trade forecasting model. That is to say, the selected variables are some important variables affecting Chinese foreign trade. (c) Usability and availability. This principle requires that the related data is available for selected variables. Once choosing the trade partners and competitors, we can select some main variables affecting Chinese foreign trade, as shown in Table 1.

Table 1. Some main factors affecting Chinese foreign trade

Category	China	Trade partners	Trade competitors
Export and import indicator	The total amount of Chinese foreign trade	Import amount of main trade partners	Export amount of main trade competitors
The nominal exchange rates against the US dollar	Renminbi (RMB) exchange rate against US dollar	Hong Kong dollar, yen, Germany mark, Korean won, Singapore dollar, Taiwan new dollar.	Japanese yen, Korean won, Taiwan new dollar, and East Asian currency rates
Economic situations and investment	Inflation rates, GDP, foreign direct investment	CPI, customer confidence index, GDP and unemployment rate	CPI and GDP of these countries
Monetary policy	Currency supply, foreign exchange reserve		

In addition, the data used in this study are monthly covered from January 1985 to December 2003, which is from various sources. The data of Chinese export, foreign reserve, and foreign direct investment is collected from Statistical bulletin of the Ministry of Commerce, while the exchange rates data is from Federal Reserve Economic Data of United States (http://www.stls.frb.org/fred/). We take the data from January 1985 to December 1999 with a total of 180 observations as training set and the remainder is used as the testing set (48 observations).

To compare the forecasting performance, the normalized mean squared error (*NMSE*) is used as the evaluation criterion, which can be represented by

$$NMSE = \frac{\sum_{i=1}^{N}(x_i - \hat{x}_i)^2}{\sum_{i=1}^{N}(x_i - \bar{x}_i)^2} = \frac{1}{\sigma^2}\frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{x}_i)^2 \tag{6}$$

where σ^2 is the estimated variance of the testing data, x_i and $\hat{x}_i$ are the actual and predicted value, $\bar{x}_i$ being the mean, and N is the number of testing data. If the estimated mean of the data is used as predictor, $NMSE$=1.0 is obtained. The $NMSE$ is related to R^2 which measures the dependence between pairs of desired values and predictions by $NMSE = 1\text{-}R^2$. For comparison, the linear EC-VAR and ANN-based nonlinear EC-VAR model is used as the benchmark model. Particularly, the text mining based judgmental adjustment is also incorporated into ANN-based EC-VAR model. For example, the SARS in April 2003, an important irregular event, has a significant impact on Chinese foreign trade in May and June 2003. Such an event is hard to be included into the model, thus we go to consult the related experts to quantify the effects on Chinese foreign trade. Experts can give a judgmental adjustment value according to their experience. Table 2 reports the empirical results.

Table 2. The out-of-sample prediction results for different prediction models (2000:01-2003:12)

Prediction Model	Prediction Performance		
	NMSE	Minimum error	R^2
Linear EC-VAR	0.2923	0.0102	0.7077
ANN-based EC-VAR	0.1347	0.0156	0.8653
JA& ANN-based EC-VAR	0.0987	0.0067	0.9013
Hybrid Econometric-AI ensemble	0.0755	0.0000	0.9245

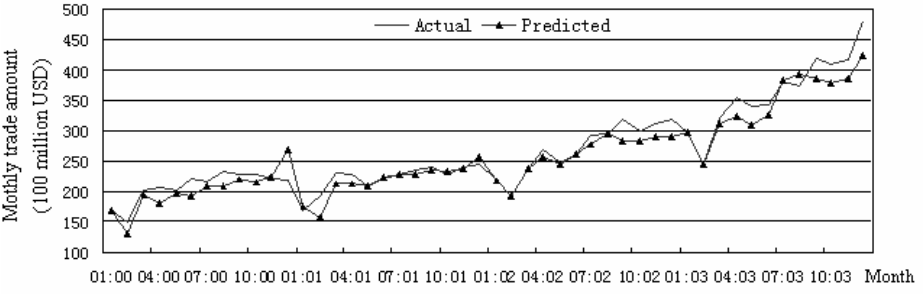


Fig. 2. Prediction results of hybrid econometric-AI ensemble learning model (2000:01-2003:12)

As can be revealed from Table 2, it is not hard to find that the proposed hybrid econometric-AI ensemble learning methodology consistently performs better than the other three prediction models. Furthermore, the minimum error of the proposed hybrid econometric-AI ensemble learning methodology is zero, implying that the proposed methodology is a feasible approach to foreign trade prediction. The main reason leading to good prediction performance may be from the following three aspects. First of all, the hybrid ensemble learning methodology integrates not only linear patterns but also nonlinear patterns as well as the effects of irregular events. Second, the hybrid ensemble learning method makes full use of both the advantage of econometric model, e.g., EC-VAR in this study and the advantages of the AI techniques such as ANN and text

mining. Third, the final ensemble model applies a nonlinear ensemble learning strategy and thus making the prediction more accurate. Concretely, the prediction performance of the hybrid econometric-AI ensemble learning model is illustrated in Fig. 2.

4 Conclusions

In this study, a new hybrid econometric-AI ensemble learning methodology is proposed for complex prediction problem. For illustration purpose, the proposed ensemble learning method integrating econometric techniques and AI methods is applied to Chinese export trade prediction problem. Experimental results reveal that the hybrid econometric-AI ensemble learning methodology can significantly improve the prediction performance over other linear and nonlinear models listed in this study, implying that the proposed hybrid econometric-AI ensemble learning methodology can be used as a feasible solution to foreign trade prediction.

Acknowledgements

This work is supported by the grants from the National Natural Science Foundation of China (NSFC No. 70221001, 70601029), the Chinese Academy of Sciences (CAS No. 3547600), the Academy of Mathematics and Systems Sciences (AMSS No. 3543500) of CAS, and the Strategic Research Grant of City University of Hong Kong (SRG No. 7001677, 7001806).

References

1. China's Future. Fortune. September 29, 1999
2. Center of China Study: Development, Cooperation, Reciprocal and Mutual Benefit: The Evaluation of International Monetary Fund on Chinese Loan (1981-2002). Research Report (2004)
3. Engle, R.F., Granger, C.W.J.: Co-integration and Error-correction: Representation, Estimation and Testing. Econometrica 55 (1987) 251-276
4. Hendry, D.F., Ericsson, N.R.: An Econometric Analysis of U.K Money Demand in Monetary Trends in the United States and the United Kingdom by Milton Friedman and Anna J. Schwartz. American Economic Review 81 (1991) 8-38
5. Hornik, K., Stinchocombe, M., White, H.: Multilayer Feedforward Networks are Universal Approximators. Neural Networks 2 (1989) 359-366
6. White, H.: Connectionist Nonparametric Regression: Multilayer Feedforward Networks can Learn Arbitrary Mappings. Neural Networks 3 (1990) 535-549
7. Yu, L., Wang, S.Y., Lai, K.K.: A rough-Set-Refined Text Mining Approach for Crude Oil Market Tendency Forecasting. International Journal of Knowledge and Systems Sciences 2(1) (2005) 33-46
8. Sullivan, D.: Document Warehousing and Text Mining: Techniques for Improving Business Operations, Marketing, and Sales. Wiley, New York (2001)
9. Yu, L., Wang, S.Y., Lai, K.K.: A Novel Nonlinear Ensemble Forecasting Model Incorporating GLAR and ANN for Foreign Exchange Rates. Computers & Operations Research 32 (2005) 2523-2541

The Origin of Volatility Cascade of the Financial Market

Chunxia Yang[1], Yingchao Zhang[1], Hongfa Wu[1], and Peiling Zhou[2]

[1] School of Information and Control Engineering,
Nanjing University of Information Science and Technology, Nanjing Jiangsu, 210044, P.R. China
[2] Department of Electronic Science and Technology,
University of Science and Technology of China, Hefei Anhui, 230026, P.R. China

Abstract. Based on the self-organized dynamical evolutionary of the investors structure, a refined dissipation market model is constructed. Unlike multifractal cascade-like ideas, this model provides a realistic (agent based) description of financial markets and reproduces the same multifractal scaling properties of price changes as the real, which indicate that the self-organized dynamical evolutionary of the investors structure may be the origin of the volatility statistical structure.

Keywords: self-organization, multifractal, cascade, financial market model, volatility.

1 Introduction

The modelling of random fluctuation of asset prices is of obvious importance in finance, which can be used in practice for volatility filtering and option pricing. The universal features of price changes have attracted widespread attention and concern to construct a useful model. The simplest feature of financial time series, uncovered by Bachelier in 1900, is the linear growth of the variance of the return fluctuations with time scale[1]. More precisely, if one denotes $V(t)$ the price of an asset at time t ,the return $r_\tau(t)$ at time t and scale τ is simply the relative variation of the price from t to $t+\tau$: $r_\tau(t) = (V(t+\tau) - V(t))/V(t) \simeq lnV(t+\tau) - lnV(t)$; if $m\tau$ is the mean return at scale τ, the following property holds to a good approximation:

$$\langle (r_\tau(t) - m\tau)^2 \rangle_e \simeq \sigma^2 \tau. \tag{1}$$

where$\langle\rangle_e$ denotes the empirical average. This behavior typically holds for τ from a few minutes to a few years, equivalent to the statement that relative price changes are uncorrelated. Linear growth of the variance of the fluctuations with time is typical of the Brownian motion which was proposed as a model of market fluctuation by Bachelier. In this model, returns are not only uncorrelated but actually independent and identical Gaussian random variables. However, due to intensive statistical studies during the last decade, this model suffers the impugnation and the challenge of actual financial data such as the real-life markets are

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 114–120, 2007.
© Springer-Verlag Berlin Heidelberg 2007

of return distributions displaying peak-center and fat-tail properties[2,3,4], one can observe volatility clustering and a non-trivial "multifractal" scaling[5,6,7], and so on. These universal features portray a world of non Gaussian random walks that Mandelbrot started exploring for us in the sixties, charting out its scrubby paths, on which increasing scientists devote themselves to look for the origin of the market fluctuation[6,8,9,10,11,12]. One of the contributions is that Mandelbrot's cascades have been used to account for scale invariance properties in statistical finance[9]. Recently, Bacry, Muzy and Delour introduced a modified model that captures the essence of cascades (BMD model) [12]. All these models explains the multi-scaling property through the notion of cascade from coarse to fine scales. But, such description using multifractal, cascades-like ideas is still mostly phenomenological. Here, without any postulate, we proposed a model with the same multifractal scaling properties as the reality which arise from a realistic (agent based) description of financial markets and help understand the underlying mechanisms.

2 Model

Cont and Bouchaud successfully applied percolation theory to modeling the financial market (CB model), which is one of the simplest models able to account for the main stylized fact of financial markets, e.g. fat tails of the histogram of log-returns[13]. Based on it, our model incorporates the following components different from the original CB model: (1) A cluster, defined as a set of interconnected investors, grows in a self-organized process. (2) The effect of "herd behavior" on the trade-volume is magnified step by step during the cluster's self-organized accumulating process. (3) Some encountering smaller clusters will form a bigger cluster through cooperating or one defeating the rivals. (4) An infinite cluster may exist without the need to tune p to p_c and its trade activity influences price fluctuation. (5) The number of investors participating in trading will vary dynamically.

2.1 Dynamic of Investor Groups

Initially, $M = 100$ investors randomly take up the sites of a $L*L$ lattice. Then for each cluster, a strategy is given: buying, selling, or sleeping, which are denoted by 1, -1 and 0 respectively. In reality, the circle of professionals and colleagues to whom a trader is typically connected evolves as a function of time: in some cases, traders are following strong herding behavior and the effective connectivity parameter p is high; in other cases, investors are more individualistic and smaller p seems more reasonable. In order to take the complex dynamics of interactions between traders into account, we assume that it undergoes the following evolution repeatedly:

(1) Growth: most of investors would like to imitate the strategies which have been adopted by many others, which induces "herd behavior" occurring. In this sense the herd behavior is amplified. Specially, the affection of the herd behavior

will be magnified gradually with the increase of the number of investors adopting this strategy, i.e., with the growth of the clusters. During cluster's growth, a number of new investors will be attracted by it and become its members. In other words, every cluster will absorb new investors with the probability:

$$P_d(\tau) = P_d(\tau - 1) + k(N_T - N(\tau - 1)). \tag{2}$$

where k is a kinetic coefficient controlling the growth speed and N_T is a threshold parameter (It has been validated that the value of the parameters k and N_T could be any value. These two parameters have no effects on the self-organization process of the clusters[14]). $N(\tau-1)$is the number of the agents along it's boundary, defined as a set made up of agents which belong to a cluster and at least border on a site which isn't part of this cluster, at the last time step $\tau - 1$. The new participating investor will take up the empty sites around the old clusters and imitate the same strategy as that of it. The probability P_d is obviously limited to the range [0, 1] so that we have to impose $P_d = 0$ and $P_d = 1$ if the recurrence relationship Equ (2) gives values for $P_d < 0$ or $P_d > 1$.

(2) New cluster's birth: some investors will randomly and independently enter the market with the probabilityP_n. These investors don't belong to an arbitrary existing cluster and will take up the empty sites.

(3) Cooperation: encountering clusters will operate cooperation and confliction between them. When their strategies are same, they are thought to cluster together to form a new group of influence. Or there would be confliction between them. The consequence of confliction is that losers would be annexed by the winner and that a new and bigger cluster whose strategy inherent the winner's will be formed. The probability of cooperation or confliction is as follow, i.e., some a cluster will cooperate with or defeat others with the probability

$$P_m(k) = \frac{|s_\tau^k|}{\sum_{j=1}^{n} |s_\tau^j|}. \tag{3}$$

where $|s_\tau^j|$ is the size of j-th cluster at time step τ.

(4) Metabolism: in reality, no matter how huge has the size of a group ever been it would collapse due to different influences such as government decision on war and peace. Some new clusters will come into the world in the wake of aging clusters' death. The probability with which an aging cluster will die is:

$$P_o = \frac{x + y}{2L}. \tag{4}$$

where x or y is the width of this cluster occurring on the lattice in the x or y direction. Equ.(4) indicates that the probability with which a cluster disbands would increase with the cluster growth. Once a spanning cluster exists, it would surely die. When a cluster disbands, all its members would leave the market and the sites where the death cluster ever occupied will be taken up by new investors

with the probability P_n. Such occupied sites form a few new clusters. Every new cluster would be given a strategy randomly.

Although each cluster could trade with others at every trading step, the evolutionary frequency of the network topology should not be so often. Thus, we assume that the network structure of the market composed by investor groups would evolve every N trading steps. With the evolutionary of this artificial stock market, the number of investors participating in trading isn't constant. The network will take on different structure; the affection of the herd behavior on the trade-volume is gradually magnified. Cooperation and confliction among clusters are always operating. Without any artificial adjustment of the connectivity probability p to p_c, spanning cluster may exist, whose activity would influence the price fluctuation.

2.2 Trading Rules

Each cluster trades with probability a (called activity); if it trades, it gives equal probability to a buying or selling with demand proportional to the cluster size. The excess demand is then the difference between the sum of all buying orders and selling orders received within one time interval. The price changes from one time step to the next by an amount proportional to the excess demand. To explain the "volatility", Stauffer introduces the feedback mechanism between the difference of the "supply and demand" and activity of the investors[15]. Whereas in our model, the difference of the "supply and demand" not only affects the activity probability but also the probability with which the active clusters choose to buy or sell. The probability a evolves following

$$a(t) = a(t-1) + lr(t-1) + \alpha. \tag{5}$$

where r is the difference between the demand and supply, l denotes the sensitivity of a to r and α measures the degree of impact of external information on the activity. $\alpha \in [-0.005, 0.005]$ is a random variable obeying Gaussian distribution of mean value 0 and standard deviation 1.Each active cluster choose to buy or sell with probabilities $2a(t)(1 - p_s(t))$ and $2a(t)p_s(t)$ respectively. For $r > 0$, $p_s(t) = 0.5 + d_1 r(t-1)$, while for $r < 0$, $p_s(t) = 0.5 + d_2 r(t-1)$. According to Kahneman and Tversky, it is asymmetry that agents make their decisions when they are in face of gain and loss [16].When $|r|$ varies within certain range, the degree of depression following some lost is twice that of happiness following the same quantity of profit. Therefore in our model we assume $d_2 = 2d_1$.The difference between the demand and supply is:

$$r(t) = \sum_{j=1}^{m} \mathtt{sign}(s_t^j)(|s_t^j|)^{\gamma}. \tag{6}$$

where m is the total number of clusters occurring on the market and $|s_t^j|$ is the size of j-th cluster at trading time step t. γ measures the degree of impact of each cluster's trade-volume on the price, $0 < \gamma < 1$ allowing for a nonlinear

dependence of the change of (the logarithm of the) price as a function of the difference between supply and demand[17]. So the evolution of the price is:

$$V(t) = V(t-1)\exp(\lambda r(t)). \tag{7}$$

where λ is a coefficient adjusting the market fluidity.

3 Simulation and Analysis

Here we set $a(0) = 0.09, r(0) = 0, V(0) = 1, P_d(0) = 0.4$, $k = 0.0001, N_T = 50$, $l = \lambda = \frac{1}{L^2}, L = 100$, $d_1 = 0.00005, \gamma = 0.8, P_n = 0.6, M = 100, N = 50$. The model can iterate for a period of any length. More simulations have been done indicating that the return distribution of the present model obeys Lévy form in the center and displays fat-tail property, in accord with the stylized facts observed in real-life financial time series. Furthermore, this model reveals the power-law relationship between the peak value of the probability distribution and the time scales in agreement with the empirical studies on the Hang Seng Index[18]. For the Lévy stable processes first suggested by Mandelbrot as an alternative, the return distribution is identical (up to a rescaling) for all τ. As mentioned in empirical studies, the distribution of real returns is not scale invariant, but rather exhibits multi-scaling, in the sense that higher moments of price changes scale anomalously with time:

$$M_q(\tau) = \langle(|r_\tau(t) - m * \tau|)^q\rangle_e \simeq A_q \tau^{\zeta_q}. \tag{8}$$

where the index ζ_q is not equal to the Brownian walk value $q/2$ [8,9,10,11,12]. References [19] illustrates the empirical multifractal analysis of Standard and Poor's Gomposite Index return. The estimated spectrum ζ_q has a concave shape that is well fitted by the parabola:

$$\zeta_q = q(1/2 + \lambda^2) - \lambda^2 q^2/2. \tag{9}$$

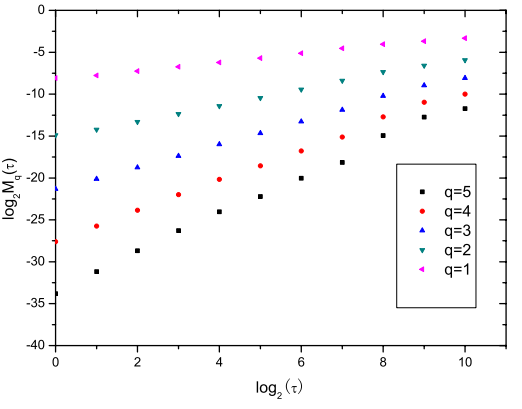


Fig. 1. First five absolute moments of the model as a function of the time period τ in double logarithmic representation

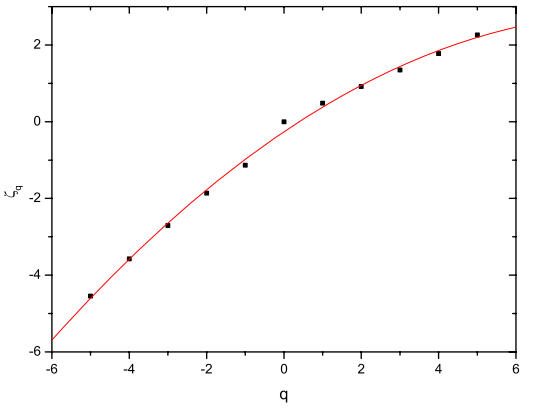


Fig. 2. ζ_q spectrum estimate versus q

where $\lambda^2 \simeq 0.03$. The coefficient λ^2 that quantifies the curvature of the ζ_q is called, in the framework of turbulence theory, the intermittency coefficient. Correspondingly in the finance parlance,λ^2 measures the intensity of volatility fluctuations.

The multifractal scaling properties of our model are numerically checked in Fig.1 and Fig.2 where one can recover the same features as that of the Standard and Poor's Gomposite Index. The fitting parabola is

$$\zeta_q = -0.265 + q(0.6 + \lambda^2) - \lambda^2 q^2/2. \tag{10}$$

where $\lambda^2 \simeq 0.074$.This indicates that our model has repuduced multifractal scaling properties of price changes.

4 Conclusion

Based on the self-organized dynamical evolutionary of the investors structure, a refined dissipation market model is constracted. In BMD model, the existence of the two independent statistical process, one for returns and another for the volatility, may not be very natural. But in this model, the two processes syncretized each other. Furthermore, unlike multifractal cascade-like ideas, this model provides a realistic (agent based) description of financial markets and repuduces multifractal scaling properties of price changes, which indicate that the self-organized dynamical evolutionary of the investors structure may be the origin of the volatility statistical structure.

Acknowledgments. This work has been supported by the National Science Foundation of China under Grant No.70471033 and 70571075, the College Science Foundation of Jiangsu Province (06KJD520122) and the Liu Da Ren Cai Gao Feng Program of Jiangsu Province (06-A-027).

References

1. Bachelier L.: Ann. Sci. Ecole Norm. Sup. 3 (1900) 21-86
2. Gopikrishnan P., Plerou V., Amaral L. A. N., et al.: Scaling of the distribution of fluctuations of financial market indices. Phys. Rev. E. 60 (1999) 5305-5316
3. Mantegna R. N., Stanley H. E.: Scaling behaviour in the dynamics of an economic index. Nature. 376 (1995) 46-49
4. Wang B. H., Hui P. M.: The distribution and scaling of fluctuations for Hang Seng index in Hong Kong stock market. Eur. Phys. J. B. 20 (2001) 573-579
5. Ghashghaie S., Breymann W., Peinke J., Talkner P., Dodge Y.: Turbulent Cascades in Foreign Exchange Markets. Nature. 381 (1996) 767-770
6. Mandelbrot B. B.: Fractals and Scaling in Finance. Springer, New York (1997)
7. Schmitt F., Schertzer D., Lovejoy S.: Multifractal analysis of Foreign exchange data. Applied Stochastic Models and Data Analysis. 15 (1999) 29
8. Mandelbrot B. B.: The variation of certain speculative prices. J. Business. 36 (1963) 394
9. Mandelbrot B. B.: Intermittent turbulence in self-similar cascades: divergence of high moments and dimension of the carrier. J. Fluid Mech. 62 (1974) 331
10. Arneodo A., Muzy J. F., Sornette D.: Direct Causal Cascade in the Stock Market. Eur. Phys. J. B. 2 (1998) 277-282
11. Muzy J. F., Delour J., Bacry E.: Modelling fluctuations of financial time series: from cascade process to stochastic volatility model. Eur. Phys. J. B. 17 (2000) 537-548
12. Bacry E., Delour J., Muzy J. F.: Multifractal random walk. Phys. Rev. E. 64 (2001) 026103
13. Cont R., Bouchaud J. P.: Herd behavior and aggregate fluctuations in financial markets. Marcroeconomic dynamics. 4 (2000) 170-196
14. Cavalcante F. S. A., Moreira A. A. et al.: Self-organized percolation growth in regular and disordered lattices. Physica A. 311 (2002) 313-319
15. Stauffer D., Jan N.: Sharp peaks in the percolation model for stock markets. Physica A. 277 (2000) 215-219
16. Kaheman D., Tversky A.: Prospect theory: an analysis of decision under risk. Econometrica. 47 (1979) 263-291
17. Farmer J. D.: Market force, ecology and evolution. Industrial and Corporate Change. 11 (2002) 895-953
18. Muzy J. F., Bacry E., Arneodo A.: The multifractalformalism revisited with wavelets. Int. J. of Bifurcat and Chaos. 4 (1994) 245
19. Borland L., Bouchaud J. P., Muzy J. F., Zumbach G.: The Dynamics of Financial Market——Mandelbrot's Multifractal Cascades, and Beyond. cond-mat/0501292 v1

Tactical Battlefield Entities Simulation Model Based on Multi-agent Interactions

Xiong Li and Sheng Dang

Department of Command and Administration, Academy of Armored Force Engineering, 100072 Beijing, China
lixiong2609@126.com, jiansijia@sohu.com

Abstract. Tactical warfare process, e.g., engagement between opposite forces, is full of unpredictability and platform-level interactions that result great difficulties in performing battlefield entities simulation. In this paper, modeling and simulation based on multi-agent interactions is applied to solve the problem. Based on the analysis on the requirement and countermeasure, the mapping from tactical warfare system's members, i.e. platform-level tactical battlefield entities, to respective intelligent agents is set up. Thus, the multi-agent platform-level tactical battlefield entities simulation system and its agent model are designed. Tactical battlefield entity agent interactions model is presented to support simulation based on multi-agent interactions by using an improved Contract Net Protocol. The established demonstration system proves the feasibility and efficiency of our model and shows its advantages in realizing platform-level military simulation.

Keywords: agent, multi-agent system, interactions, modeling and simulation.

1 Introduction

Tactical battlefield entities simulations are usually used to train soldiers to perform missions and to learn to work together in teams and across command structures, or carry out the advanced concept technology demonstrations for operational applications of equipment systems on future battlefield. How to capture tactical warfare's realism, its interactions and unpredictability, and still provide decision makers with useful insight, is an issue needing to be studied. However, conventional modeling methods can not cater for the requirement. For example, linearization, which "linearizes" problems to derive an analytical solution, comes at the price of realism since problems are not always decomposable into independent parts. The decomposition process fails to accurately capture the component interaction and these interactions dominate the real world making tactical warfare unpredictable by analytical means.

Intelligent agents and multi-agent systems that emerged as a sub-field of artificial intelligence have turned out to be useful for a wide range of application domains where difficult problems have to be dealt with. In the past few years, interest in agents has grown at an astonishing rate [1]~[8]. Multi-agent-based modeling and simulation for tactical warfare process, e.g. engagement, has been the research focus for military concept developers and military simulation systems designers.

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 121–128, 2007.
© Springer-Verlag Berlin Heidelberg 2007

But most current multi-agent-based modeling and simulation research fruits in military field usually concentrate on theory advancement [5]~[8], so they are far away from practical applications. Even some successful models have a shortage in platform-level modeling and simulation. For example, the Hierarchical Interactive Theater Model [7] constructed and exercised by U.S.A. Air Force Studies and Analyses Agency is effective, but it can only perform unit-level simulation. Similarly, the model in [8] is task-oriented, not based on platform-level modeling and simulation. The limitation would result in difficulty in describing subtly the real-time interactions of tactical battlefield entities. Moreover, almost all researches on agent-based simulation are not based on multi-agent interactions. Thus there are a lot of difficulties when the systems are implemented, since agents and multi-agent systems are complex and have many properties such as autonomy, reactivity, sociality, adaptability and intelligence. It is impossible to take all these factors into account.

Tactical warfare process has heterogeneous members, such as tanks, missile launch vehicles, armored reconnaissance vehicles, electronic reconnaissance platforms, and combat command platforms, which have administrative levels and a lot of interactions, such as sending or receiving combat orders. Thus we can think that tactical warfare system is in substance a distributed artificial intelligence system. Since an agent may have believes, desires, intentions, and it may adopt a role or have relationships with others, tactical warfare system can be looked upon as a collection of autonomous agents that are dependent upon each other. Therefore the method of modeling and simulation based on multi-agent interactions is applicable to our case. In this paper, we design a platform-level tactical battlefield entities simulation model based on multi-agent interactions to lay a foundation for the advanced concept technology demonstration of warfare activities on future battlefield.

2 Agents Model

An intelligent agent with human being properties such as autonomy, sociality, adaptability and intelligence can act as a human. Especially multi-agent systems consider how a group of intelligent and autonomous agents coordinate their capacities and plan in order to achieve certain (local or global) goals [1], [2]. Agents may be seen as a natural extension of the concept of software objects. Object-oriented programming added abstraction entities, i.e., objects, which have persistent local states to the structured programming paradigm. Similarly, agent-based programming adds abstraction entities, i.e., agents, which have an independent execution thread to the object-oriented paradigm. Thus, compared to an object, an agent is able to act in a goal-directed fashion (e.g., by interacting with other agents, reading sensors, or sending commands to effectors) rather than only passively react to procedure calls, as shown in Fig. 1.

Tactical warfare system is so alike a distributed multi-agent system in behaviors that we can set up a mapping from its internal members, i.e. platform-level tactical battlefield entities, to entity agents, e.g., tank $\rightarrow$ tank agent, combat command vehicle $\rightarrow$ combat command vehicle agent.

In order to develop effectively virtual battlefield simulation system which can be called a whole federation, in the course of the mapping we should sort not only the function agents (entity agents), but also the administration agents and service agents.

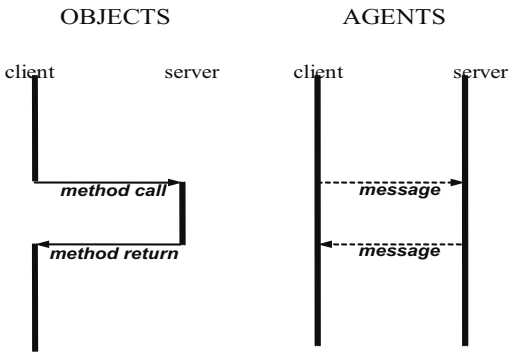


Fig. 1. Multi-threaded agents

The function agents in Red force include tank agents (TA), photo-reconnaissance vehicle agents (PRVA), radar reconnaissance vehicle agents (RRVA), armored reconnaissance vehicle agents (ARVA), cannon agents (CA), combat command vehicle agent (CCVA) and logistic support platform agents (LSPA). They are aggregated into the Red agents federation.

The function agents in Blue force are similar to those Red force agents, but some different agents, e.g., armored cavalry vehicle agents (ACVA), missile launch vehicle agents (MLVA), trench mortar agents (TMA) and information processing vehicle agents (IPVA) are designed since there are some differences in force organization. They are aggregated into the Blue agents federation.

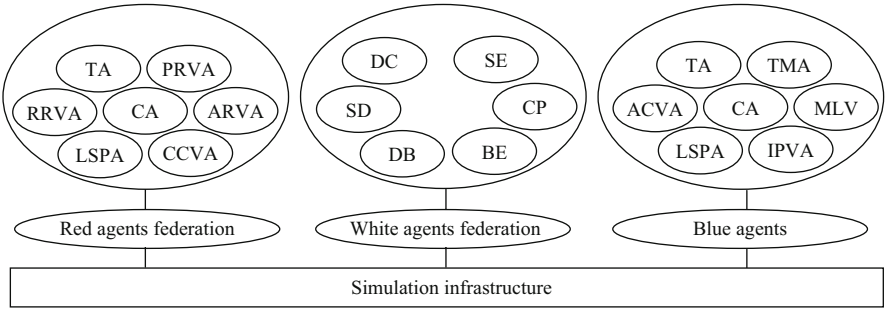


Fig. 2. Multi-agent battlefield entities simulation system architecture

Of course, we can add or cut down some function agents the Red or Blue agents federation according to the actual simulation design and development.

The administration agents and service agents include federation manager agent, declare manager agent, time manager agent, data distribution manager agent, and so on, which play the roles of demonstration control (DC), simulation evaluation (SE), data base (DB), situation displaying (SD), command practice (CP) and battlefield environment (BE). These agents can be aggregated into the "White" federation.

In this way, we can design the basic organization of distributed multi-agent platform-level tactical battlefield entities simulation system as shown in Fig. 2.

In this paper, instead of focusing our research on entity agent intelligence, we concentrated on the design of a practical framework for the development of agents capable of operating efficiently in the real simulation system. Fig. 3 shows the internal model of the entity agents in the platform-level tactical battlefield entities simulation system framework [1], [2], [4].

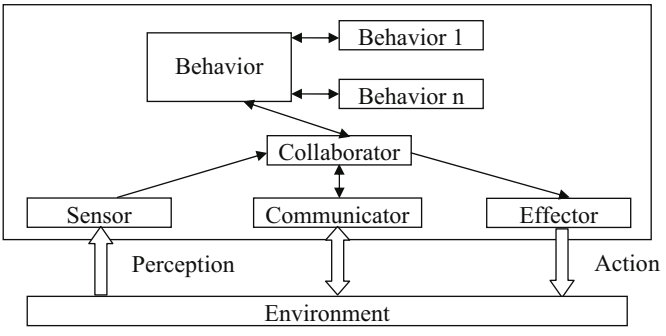


Fig. 3. Internal model of agent

In this paper, we only take one entity agent in the Red or Blue agents federation as example to illuminate the architectures of agents. In fact, the operation principium of agents in White agents federation is accordant. There are only some differences in the definitions and operation contents because of the differences in their functions.

3 Entity Agent Interactions

In this platform-level tactical battlefield entities simulation system, engagement is modeled as a distributed process among many general platforms (general entity agents) coordinated by the command platform (command entity agent). In the system, the domain data D, rules P and organizational knowledge O are based on three factors:

(1) The experience and knowledge of a general entity is based totally on its criteria (elementary belief)

(2) The general entity acquires knowledge through communication with other general entities and command entities.

(3) The general entity acquires knowledge by observing the behavior of other general entities and command entities. In practice a general entity is influenced by the above factors and the modified knowledge is incorporated in D, P and O.

Contract Net Protocol (CNP) [1], [3] proposes episodic rounds of inter-communication acts (announcements, bids, award messages) and shows its usefulness widely. The schematic representation is presented in Fig. 4. To describe unpredictability and platform-level interactions more felicitously, in this paper we use an improved CNP. In our case, tactical warfare system consists of a Red armored force unit (one combat command vehicle, nine tanks and some armored reconnaissance platforms) and a Blue army troop (one information processing vehicle, one tank, one missile launch vehicle, one trench mortar, and some

other fire platforms). The Contract Net initiator as a manager represents the combat command vehicle agent or information processing vehicle agent, and all other participants as contractors represent the other entity agents. Of course, the roles of manager and participants are changed once interaction relation changes.

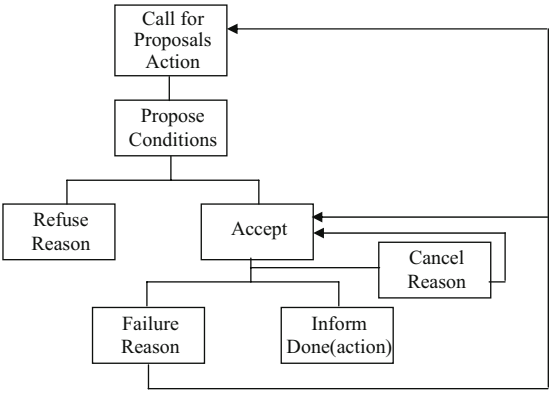


Fig. 4. Contract Net Protocol

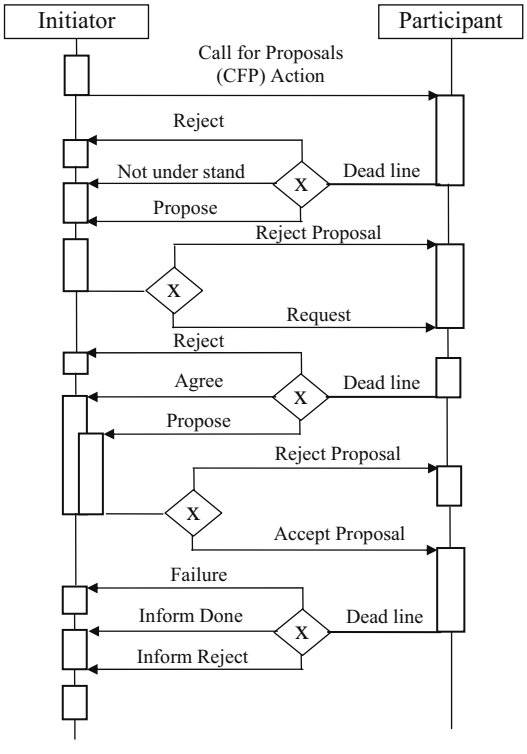


Fig. 5. Improved CNP of the Red force agents

In our model the manager wishes a task to be performed by one or a group of entity agents according to some arbitrary function which characterizes the task. The manager issues the call for proposals, and other interested agents or agents having obligation can send proposals. In contrast to the original CNP, there is no need to do anything if an agent playing a role of a participant or potential contractor is not interested or has no obligation in submitting proposals. That means that our Contract Net model from the very beginning relies on the notion of timeout, i.e. some actions need to be performed in the event of a lack of enough proposals or even in the case of a complete lack of proposals.

The proposals are collected by the manager, and then they are refused or accepted. The accepted proposals can be cancelled, either, by the manager via a cancel action, or by the contractor via a failure action. In case of cancellation other submitted proposals can be reconsidered, or a completely new call for proposals can be issued.

Fig. 5 presents the improved CNP of the Red armored force unit. The interaction is started by the combat command vehicle agent who acts as a manager issuing a call for proposals, e.g. destroying the No. 1 target in 1283 highland. These tank agents who act as participants or potential contractors respond with proposals, which the combat command vehicle agent either rejected or accepted. Accepted proposals can be either cancelled by the combat command vehicle agent, or executed by a certain tank agent, who later informs the combat command vehicle agent of success or failure of the execution.

4 Demonstration System

The demonstration system that we set up can be illustrated by Fig. 6. Fig. 6 presents the dynamic and real-time situation information during platform-level tactical battlefield entities simulation where the deployment of Red force tanks is approximately transverse. By this system, one can find out easily a certain agent's real-time state information, as shown in Fig. 7.



Fig. 6. Partial two-dimension battlefield situation



Fig. 7. A Red force agent's real-time state information

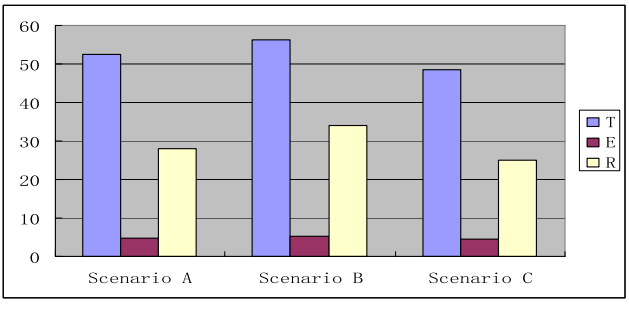


Fig. 8. Contrastive results of three scenarios

According to the military experiences on tactical warfare process on distributed battlefield, we can set appropriate data to the parameters for our system. When we run the simulation system, we can obtain some results, which are shown in Fig. 8 in which T represents total time for fulfilling the attack battle task (minute), E represents attack efficiency (min/ target) and R represents rate of destroyed force (%). In Scenario A, the Red armored force unit takes a transverse deployment. Column and triangular deployment are taken respectively in Scenario B and Scenario C. Thus by these simulation results one can find that Scenario C is the most effective attack battle plan for the Red armored force unit while Scenario C is the worst one.

We carry through Verification, Validation, and Accreditation (VV&A) for our platform-level tactical battlefield entities simulation model to analyze these results. As far as the concept model, we check whether attributes description, engagement and interactions, e.g., the entities and their tasks are consistent with real force situation. As far as the program model, emphases are put in data to verify their correctness, dependability and performance.

By the evaluation, these results that we obtained from battlefield entities simulation are accordant to real tactical warfare situation. The fact proves that our model is feasible and effectual.

5 Conclusion

Multi-agent-based modeling and simulation approaches to military simulation field gained increasing attention in recent years. However most existent models and systems can not provide enough detail to examine important dynamics in tactical warfare process, e.g., unpredictability of tactical warfare system operations and entity interactions. In this paper, a multi-agent platform-level tactical battlefield entities simulation model based on multi-agent interactions is studied. The multi-agent organization of platform-level simulation system and the architecture of entity agents are put forward, and the entity agent interactions model in this system is furthermore proposed by using an improved Contract Net Protocol. Although the established distributed simulation system model needs more research to be more practical, the demonstration system shows that our model can be used to understand the external, complicated and intelligent tactical warfare resources application and can realize the dynamic platform-level battlefield activities simulation.

References

1. Zhongzhi Shi: Intelligent Agents and Their Applications. Beijing: Science Press (2001)
2. V. Lesser: Autonomous Agents and Multi-Agent Systems. Kluwer (1998)
3. Haque, N. R. Jennings, L. Moreau: Resource Allocation in Communication Networks Using Market-Based Agents. In: Proc. 24th Int. Conf. on Innovative Techniques and Applications of AI (2004) 187–200
4. Xusheng Yang, Wanxing Sheng and Sunan Wang: Agent-Based Distribution Control and Automation System. In: Proc. Int. Conf. on Communications and Information Technologies, IEEE Press (2005) 222–225.
5. Xiong Li, Xiaobin Liu, Na Hao: Multi-agent-oriented Modeling for Intelligence Reconnaissance System. In: Proc. Int. Conf. on Parallel and Distributed Computing. IEEE Press (2005) 563–566
6. Xiong Li, Degang Liu, Hua Cong: Multi-Agent-Based Space Information Interaction Simulation Model. In: Proc. Int. Conf. on Space Information Technology, SPIE Press (2005) 598509-1–598509-5
7. Richard K. B., Gregory A. M., and Raymond R. H.: Using Agent-Based Modeling to Capture Airpower Strategic Effects. In: Proc. Int. Conf. on 2000 Winter Simulation Conference (2000) 1739–1746
8. Hou Feng, Chen Honghui and Luo Xueshan: Multi-agent based Modeling and Simulation of C^4ISR system. Electro-optic Technology Application. 3(2004) 25–30

Extensive Epidemic Spreading Model Based on Multi-agent System Framework

Chunhua Tian, Wei Ding, Rongzeng Cao, and Shun Jiang

IBM China Research Laboratory, Beijing 100094, China
{chtian,dingw,caorongz,jshun}@cn.ibm.com

Abstract. In current epidemic spreading study, continuous system dynamics model and discrete agent model are often regarded as two types of incompatible models. However these two types of models may coexist in large-scale epidemic simulation. A multi-agent based framework for such coexisting is proposed in this paper consisting of location agents with hierarchical structure, participant agents to denote host, pathogen, or medium, and event agent to model the epidemic outbreak, control policy, etc. Transformation algorithm is provided for the communication between locations with different types of epidemic models. Such a framework is illustrated by SARS (Severe Acute Respiratory Syndrome) outbreak case.

Keywords: Epidemic Spreading Model, Epidemic Simulation, Multi-Agent System.

1 Introduction

Epidemics spreading simulation is an important approach for epidemics control policy analysis. Currently there are two types of models [1], continuous system dynamics model (i.e., continuous differential equation) and purely discrete agent model with great difference in state variables. States variables in system dynamics model are the size of the groups, while that in multi-agent model are health state of agents.

From SIR (Susceptible, Infective, Recovered) model, many extensional system dynamics model are developed, such as SIS, SEIR ('E' denotes "Exposed"), MSEIR ('M' denotes "maternal antibody protection"), etc. All these models assume that the population being affected by a disease is "well mixed" and either not distributed or geographically distributed in a uniform manner.

With the application of computer simulation, more complicated and practicable models are proposed considering spatial dimension, traffic network, diversity of population and many other important factors in epidemic spreading. EpiSims [2] is an agent-based modeling platform, which relies on TRANSIMS (TRanportation Analysis SIMulation System) to emulate social behavior and traffic flow. However this approach often requires large amounts of computing resources even for small city scenario. To avoid such disadvantage, in STEM (Spatial and Temporal Epidemiological Modeler) by Ford et al. uses continuous modeling for large areas while agent based model for specific areas [3].

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 129–133, 2007.
© Springer-Verlag Berlin Heidelberg 2007

In current application, either continuous or discrete model is adopted in simulation. However, there are many scenarios requiring the coexisting of both types of models. For example, data availability (city vs. countryside), monitoring focus, multiple epidemics outbreak at the same time. The purpose of this paper is to propose a hybrid model for the common simulation platform.

2 Epidemic Spreading Model Framework

2.1 Epidemics Outbreak Scenario

The system architecture for epidemics simulation is shown in Fig. 1. At the core, there are "participant" model to emulate the behavior of host, pathogen, and medium. "Location" model to describe the geographic feature. Another key model is "social contact network" model to describe the interaction between participants and location-based activities. For modern society, transportation network has been a key channel for local epidemics to become pandemics. So the traffic model is often a key data source to construct social contact network. Demography data is used to construct participant distribution in geography, age, sex and other factors which may be needed in participant modeling. Outbreak scenario includes the flowing 3 types of factors. 1) Epidemics outbreak event to describe when certain or several types of epidemics outbreak. 2) Pathogen evolution event. 3) Environment change event such as weather, nature disaster, etc, which may impact on host/vector behaviors or epidemic spreading. Control policy event to describe the detailed policy and when it is put into action. The first three factors are often called as outbreak scenario.

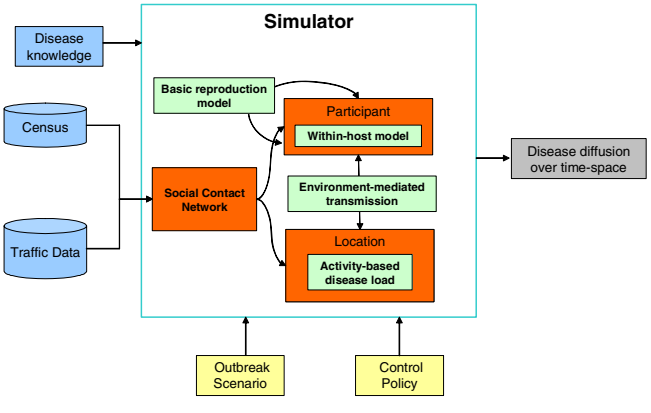


Fig. 1. Epidemic Simulation Architecture

2.2 Multi-agent Framework

Participant agents can be divided into two categories, individual or aggregate. Continuous epidemics model is often applied to aggregate participant agents, where each

agent denotes a group of host, pathogen, mediums in a location. Discrete epidemic model is correlated with individual participant agent.

However, the computation of these two types of models differs a lot. In multi-agent, there is an identifier with each agent. While system dynamics model only cares about the size of each group. So when these two types of models coexist, the transformation method is needed.

1) An agent moves from a location with discrete model to a location with continuous model

The group corresponding to the agent's state size increases 1. As long as the agent stays at the location, during each period update, the state of the agent is determined according to uniform probability. The travel behavior of the individual agent is not depended on the aggregate agents.

2) A unit flow from a location with system dynamics model to a location with multi-agent system

Generate a virtual agent with state as the flow state. Its behavior is configured according to the aggregate agent. When virtual agent leaves for another location with multi-agent model, it remains the same. When virtual agent leaves for another location with system dynamics model, it is annihilated in corresponding state group.

2.3 Metamodel

To support the co-existing and extensibility of continuous and discrete model, interface breakdown hierarchy shown in Fig. 2 is adopted. At the top of the hierarchy is the interface IDiseaseModel. It represents the abstract idea of a model of a disease process. In particular, it contains the state transition diagram although no detailed computational process that generates the transitions from one state of a particular disease model to another. IDiseaseModel is extended by the interface ISDDiseaseModel for continuous model and IMASDiseaseModel for discrete model, which expands abstraction by introducing the state transition computation process.

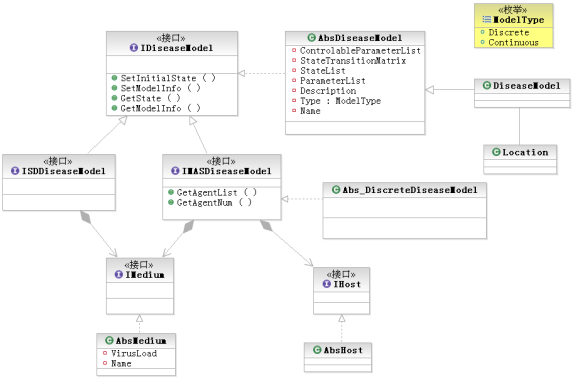
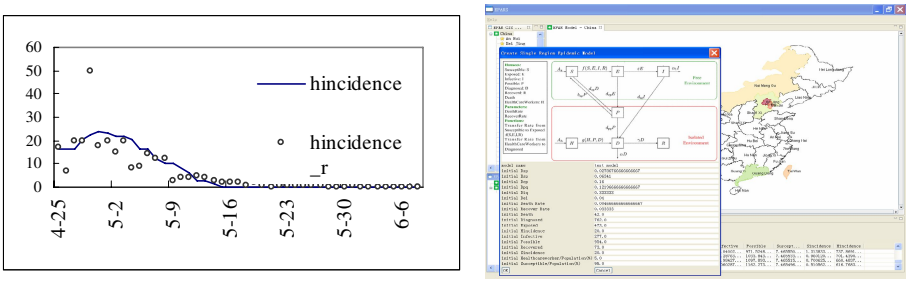


Fig. 2. Metamodel of Epidemic Spreading Model

3 Illustration

SARS outbreak case in 2003 is adopted to illustrate different modeling technologies. At first, the continuous model, BloComp(2,7) model proposed by Zhang et al. [4] is adopted to study SARS spreading dynamics in Beijing. Based on our adaptive parameter estimation algorithm, the simulation results (solid curve) have high approximation to the reported data (dotted curve) from Chinese Minister of Health [5], just as Fig. 3(a) on the incidence of healthcare workers. The difference of most days is kept within 15% except a big error in April 30.



(a) Simulation result of incidence number of healthcare workers

(b) Simulation tool snapshot

Fig. 3. SARS Spreading Simulation Tool & Result

Since most public SARS data are reported at the administration level of provinces. In the study the SARS spreading process across the China show in Fig. 3(b), Beijing is model as an agent model, while other provinces are modeled by continuous model. In the agent model, social network is generated according to the demography data such as age, job, and activity models. The people contact is stochastically generated according to the social network and agent activity model. Unfortunately traffic data across provinces is not directly available to us. For illustration purpose, national statistics on annual people transportation volume is adopted. Average daily traffic volume between two provinces is generated according to their population. The distribution in each state (that is, susceptible, exposed and infectious in free environment) is generated according to their proportion in the total population.

4 Conclusion

Due to the heterogeneity of data availability, the co-existing of continuous differential model and agent model is inevitable in epidemic simulation. Based on multi-agent framework, the architecture, method and model synthesizes the types of models are proposed in this paper. SARS outbreak case is adopted to verify the feasibility of such an approach.

References

1. Hethcote, H. W.: The mathematics of infectious diseases. SIAM Review 42 (2000) 599-653.
2. Eubank, S., Guclu, H., et al.: Modelling disease outbreaks in realistic urban social networks. Letters to Nature 429 (2004) 180-184.
3. Ford, D. A., Kaufman, J. H., Eiron, I.: An extensible spatial and temporal epidemiological modeling system. International Journal of Health Geographics 5 (2006) 4.
4. Zhang, J., Lou, J., Ma, Z., Wu, J. H.: A compartmental model for the analysis of SARS transmission patterns and outbreak control measures in China. Applied Mathematics and Computation 162 (2005) 909–924.
5. Chinese Minister of Health: SARS Report, http://168.160.224.167.

Simulation of Employee Behavior Based on Cellular Automata Model*

Yue Jiao[1], Shaorong Sun [1], and Xiaodong Sun[2]

[1] College of Management, University of Shanghai for Science and Technology, Shanghai 200093, P.R. China
[2] Antai College of Economics and Management, Shanghai Jiao Tong University, Shanghai 200052, P.R. China
jiaomooner@hotmail.com, sunshaorong@usst.edu.cn, mlshmily@sjtu.edu.cn

Abstract. The aim of this current paper is to research the interactive influence of employee behavior in a given organization. First, we define three kinds of employee behavior called Positive Behavior, Zero Behavior and Negative Behavior. Then, we give a new cellular description of behavior states and define the evolution rules for this cellular automata (CA) model. In order to find what may influence the employee behavior and how, we consider two cellular attributes: behavior's influence force, recoded as Influence and behavior's insistence force, recorded as Insistence. Finally, we use this improved CA model to simulate how employee behavior evolves, and how encouragement rules and punishment rules influence employee behavior.

1 Introduction

Employee behavior is the action of attitudes, working style, planning directly or indirectly evoked in the work. Active and energetic behavior is good for the organization to obtain their goals, and vice versa. Analyzing the employee behavior is useful for managers to lead negative behavior to positive side [1]. It is difficult to describe and look into the interactive influence among employees by applying general mathematical models. However, the self-reproducing and neighborhood rule of CA are very suitable for simulating employees and their behaviors in an organization, for they affect their neighbors' behavior and are affected by their neighbors and this process is a complex self-reproducing.

Cellular automata are simple models of computation which exhibit fascinatingly complex behavior. They have captured the attention of several generations of researchers, leading to an extensive body of work [2]. To some extent, CA can be used to reflect the behavior of human. So we apply CA to simulate employee behavior in a given organization in order to analysis how employee behavior evolves, and how encouragement rules and punishment rules influence employee behavior.

* Supported by the Program of National Natural Science Foundation of China, No. 70271005, No. 70471066; Supported by the Program of Shanghai Important Basic Research Program, No. 03JC14054; Supported by Shanghai Leading Academic Discipline Project, No.T0502.

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 134–137, 2007.
© Springer-Verlag Berlin Heidelberg 2007

2 Cellular Automata Model for Employee Behavior

Employee behavior which is encouraged and propitious to the management, production, creation and cooperation of the organization, could be called Positive Behavior(PB), such as invention of technology, retrenchment of resource; employee behavior which is the behavior not encouraged, forbidden by the rule or the culture of the organization, could be called Negative Behavior(NB), such as privilege abuse and theft[3,6]; and Zero Behavior(ZB), between Positive and Negative, is neither encouraged nor forbidden. ZB may be the leak of the rule, or is not heavy to such an extent that the object can be punished, such as absenteeism, substance abuse [4]; it may be inaction, such as do-nothing behavior.

Employees in the organization are looked as the CA space, and each employee is a cell. Distance between cells is not the distance in physics, but in psychology and behavior. Each cell is influenced by its neighbors, and at the same time influences others, which cause the evolution and the update of the employee behavior. The closer the distance is, the more influence, and vice versa.

In CA space every cell has three behavior states $S_{i,j}^t = \{1,0,-1\}$, where 1 is the PB, 0 is the ZB and -1 is the NB. For the employee in the organization, everyone has his difference. We think about two characteristics related to employee behavior: Influence and Insistence. Influence is the extent that some employee affects his neighbors. Insistence is the extent of the employee's holding his own behavior. High-Insistence employee is difficult affected by his neighbors [2]. So, each cell has two characteristics: Influence $INF_{i,j} = \{1,2,3\}$, Insistence $INS_{i,j} = \{1,2,3\}$, where each characteristic has three degrees as we hypothesis.

Employee behavior is affected by his neighbors. Different neighbor behavior makes different influence to the cell. The cumulate influences of PB, NB, and ZB neighbors on one given cell are separately called Positive, Negative, and Zero Environmental Disturbances Degree, formulated by:

$$ped_{i,j}^t = \sum_{i=i-2}^{i+2} \sum_{j=j-2}^{j+2} \frac{INF_{i',j'}}{\sqrt{(i'-i)^2 + (j'-j)^2}}, S_{i',j'}^t = 1 \tag{1}$$

$$ned_{i,j}^t = \sum_{i=i-2}^{i+2} \sum_{j=j-2}^{j+2} \frac{INF_{i',j'}}{\sqrt{(i'-i)^2 + (j'-j)^2}}, \ S_{i',j'}^t = -1 \tag{2}$$

$$zed_{i,j}^t = \sum_{i=i-2}^{i+2} \sum_{j=j-2}^{j+2} \frac{INF_{i',j'}}{\sqrt{(i'-i)^2 + (j'-j)^2}}, \ S_{i',j'}^t = 0 \tag{3}$$

We define local rules:

(1) When $S_{i,j}^t = 1$, if $ped + INS_{i,j} = \max\{ped + INS_{i,j}, ned, zed\}$, then $S_{i,j}^{t+1} = 1$; else, if $ned > zed$, then $S_{i,j}^{t+1} = -1$; if $zed > ned$, then $S_{i,j}^{t+1} = 0$; if $zed = ned$, then $P\{S_{i,j}^{t+1} = 0\} = 0.5$, $P\{S_{i,j}^{t+1} = -1\} = 0.5$.

(2) when $S_{i,j}^{t}=-1$, if $ned+INS_{i,j}=\max\{ned+INS_{i,j}, ped, zed\}$, then $S_{i,j}^{t+1}=-1$ else, if $ped>zed$, then $S_{i,j}^{t+1}=1$; if $zed>ped$, then $S_{i,j}^{t+1}=0$; if $ped=zed$, then $P\{S_{i,j}^{t+1}=1\}=0.5$, $P\{S_{i,j}^{t+1}=0\}=0.5$.

(3) When $S_{i,j}^{t}=0$, if $zed+INS_{i,j}=\max\{zed+INS_{i,j}, ped, ned\}$, then $S_{i,j}^{t+1}=0$; else, if $ped>ned$, then $S_{i,j}^{t+1}=1$;if $ned>ped$, then $S_{i,j}^{t+1}=-1$; if $ped=ned$, then $P\{S_{i,j}^{t+1}=1\}=0.5$, $P\{S_{i,j}^{t+1}=-1\}=0.5$.

In the evolution process of employee behavior, the policy of the organization plays an important role [5]. It is in that employee will strengthen his behavior intensity when policy encourages the relative behavior, and reduces when forbids. In order to find how the policy affects the employee behavior, we propose the rule:

When the organization encourages the PB of employee,

$$ped_{i,j}^{t}=\sum_{i=i-2}^{i+2}\sum_{j=j-2}^{j+2}\frac{\alpha INF_{i',j'}}{\sqrt{(i'-i)^2+(j'-j)^2}},\ S_{i',j'}^{t}=1 \tag{4}$$

When the organization punishes the NB of employee,

$$ned_{i,j}^{t}=\sum_{i=i-2}^{i+2}\sum_{j=j-2}^{j+2}\frac{\beta INF_{i',j'}}{\sqrt{(i'-i)^2+(j'-j)^2}},\ S_{i',j'}^{t}=-1 \tag{5}$$

where $\alpha\in R, \alpha>1$, and $\beta\in R, 0<\beta<1$.

3 CA Simulation

The evolution of the employee behavior in an enterprise of $100\times 100 = 10{,}000$ is simulated. The proportion of PB, ZB and NB employee and the encouragement and punishment policy is analyzed, and also their effect to the evolution of the employee behavior. The Influence and Insistence is the integer distributing uniformly in $[1,3]$. No encouragement and punishment policy. We simulate: (a) the proportion of PB, ZB and NB employee is $PS:ZS:NS=1:1:1$, distributing uniformly. The stable state of this situation is shown in Fig.1. The color in figure—black, gray and white is respectively the employee of PB, ZB and NB; (b) encouragement policy is put in force to encourage the PB, with $\alpha=1.1$; (c) punishment policy is put in force to punish the

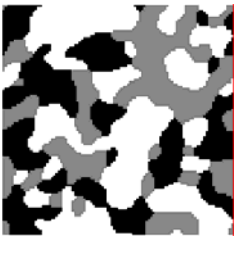


Fig. 1. Stable State

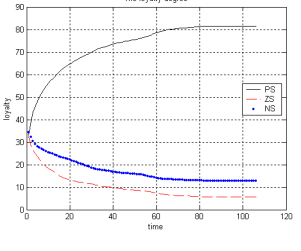


Fig. 2. Employee Proportion

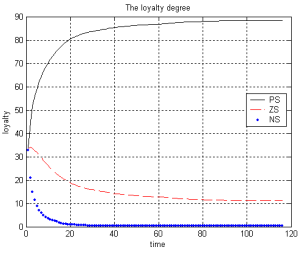


Fig. 3. Employee Proportion

NB, with $\beta = 0.9$; (d) both are put in force at the same time, with $\alpha = 1.1$, $\beta = 0.9$. We just give the proportion pictures (shown in Fig.2-3) of situation (b) and (d) because of the length limit of our paper.

Comparing the employee proportion graph of (b), (c), (d) and (a), we find the proportion of PB employee is higher than that in (a) because of the policy. The proportion of PB employee in (b) is 81%, which is much higher than that in (c), 47%; the proportion of ZB employee in (d) is 6%, which is much lower than that in (c), 46%; the proportion of NB employee in (b) is 13%, which is higher than that in (c), 7.5%. The reason is that the encouragement policy builds a hortative environment to reform behavior from Zero and Negative to Positive, while the punishment policy restricts the NB. The transfer from ZB to NB in punishment policy is merely restricted, and to positive one is not encouraged, so the proportion of ZB employee in (c) is even higher than that in (a). In (d), the proportion of PB employee is the highest, and the proportion of NB employee is the lowest. And the graph (d) changed the most quickly, in that the both policies strengthen the choice of the employee.

Our results reveal that both policies increase the proportion of PB employee, so it is necessary to make relative policies. From the simulation we find that each policy has different effect on different behaviors. In order to reduce the extra cost of the ZB, encouragement policy is better than punishment policy. But to reduce NB, the latter one is more efficient. To increase PB and reduce NB, the policies may be used together, but there may be no exact effect on controlling ZB.

4 Conclusion

In this paper we propose the conception of PB, ZB and NB, and research the interactive influence of employees' behavior in a given organization. In order to find what may influence the employee behavior and how, we consider two cellular attributes: behavior's influence force, recoded as Influence and behavior's insistence force, recorded as Insistence. Finally, we use the improved cellular automata model to simulate how employee behavior evolves, and how different encouragement rules and punishment rules influence employees' behavior.

References

1. Mark John Somers: Ethical Codes of Conduct and Organizational Context: A Study of the Relationship between Codes of Conduct, Employee Behavior and Organizational Values. Journal of Business Ethics 30(2001) 185–195, 2001
2. PALASH SARKAR: A Brief History of Cellular Automata. ACM Computing Surveys 1(2000) 80-107
3. L. A. Burke, L. A. Witt: Personality and High-maintenance Employee Behavior. Journal of Business and Psychology 3(2004) 349-363
4. Dane K. Peterson: The Relationship between Unethical Behavior and the Dimensions of the Ethical Climate Questionnaire. Journal of Business Ethics 41(2002) 313–326
5. HU Bin, ZHANG De-bin: Distance Based Cellular Automata Simulation for Employee Behaviors.Systems Engineering-Theory & Practice 2(2006) 83-96
6. Aaron Bolin, Linette Heatherly: Predictors of Employee Deviance: The Relationship between Bad Attitudes and Bad Behavior. Journal of Business and Psychology 3(2001) 405-418

Modeling, Learning and Simulating Biological Cells with Entity Grammar

Yun Wang[1,*] Rao Zheng[2], and Yan-Jiang Qiao[1]

[1] Beijing University of Chinese Medicine, Beijing, 100102, China
[2] Beijing University of Chemical Technology, Beijing, 100029, China
yonwangpku@yahoo.com.cn

Abstract. In recent years, whole cell modeling approaches, which combine existing knowledge and machine learning results, have received considerable attention. These approaches are potentially very efficient for simulating and analyzing physiological function of cells. In this work, entity grammar system is proposed as formalism for knowledge representation and multistratgy learning techniques in systems biology. Modeling biological cells with entity grammar starts from the simple grammatical models. Integrating the simple models into a cooperating entity grammar of cell facilitates real-time model learning and updating. This makes a difference with many other formalisms to build preset models. The scheme of such a platform is described in the paper and the possible applications are discussed. The proposed formalism method is open to all reasoning paradigm and can be used for studying biological complex systems.

Keywords: entity grammar system, systems biology, complex system.

1 Introduction

Computational models in biology connect molecular mechanisms to the physiological properties of cell, which is important to integrate the knowledge of cells. There have been many previous efforts to design simulators for general purpose such as GEPASI[1], E-CELL[2], Virtual Cell[3], BioDrive[4], Cellerator[5] etc. The involved formalisms include differential equations, Boolean networks, logical/graph-based (LG), Bayesian Networks, Petri-nets, π-calculus, etc. Most of the formalisms are based on the settled knowledge about cells, which deal with the problems of modeling from knowledge. To solve the problems of knowledge discovery from data, machine learning techniques are required. However, in most conditions, modeling from knowledge and knowledge discovery from data are concurrent. Combination of modeling and simulation procedures with machine learning techniques will promote and simplify both the process of model validation and data integration. Although much endeavor has been done in this direction, modeling, learning and simulating complex systems require more flexible formalism, which is also one of the main tasks in systems biology.

* Corresponding author.

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 138–141, 2007.
© Springer-Verlag Berlin Heidelberg 2007

To meet the need for new modeling formalism, learning strategies, simulation platforms and their integration, we present the formalism of Entity Grammar System (EGS) as a potential choice for these tasks.

2 Entity Grammatical Model of Biological Cell

In this section, we will recall the basic definitions of EGS. For more details, please refer to reference [6].

An entity grammar G is a quintuple, $G=(V_N,V_T,F,P,S)$, where, V_N is finite set of non-terminal symbols, V_T is finite set of terminal symbols, and $V_N \cap V_T = \Phi$, F is finite set of operations, $F=\left\{f_i \mid f_i:(E(V,F))^n \to E(V,F),1\le i\le m,\ m,n\in N\right\}$, where $V=V_N\cup V_T$, P is a finite set of productions $\alpha\to\beta$ with $\alpha\in E^+(V,F)$ and $\beta\in E(V,F)$, S is the start entities.

A cooperating entity grammar of n EGSs is denoted by

$$G=\left(g_{V_N}\left(V_{N,1},\cdots,V_{N,n}\right),g_{V_T}\left(V_{T,1},\cdots,V_{T,n}\right),g_F\left(F_1,\cdots,F_n\right),g_P\left(P_1,\cdots,P_n\right),g_S\left(S_1,\cdots,S_n\right)\right) \tag{1}$$

Where g is the cooperating functions of alphabets, sets of operations, sets of production rules and sets of start entities.

To model biological system, the cells can be represented by several cooperating systems, including the DNAs, RNAs, proteins, membrane and small molecules. The subsystems cooperate through their interactions in different levels.

Suppose we have the following entity grammatical models:

$$G_{\text{DNA}}=\left(V_{\text{DNA}},F_{\text{DNA}},P_{\text{DNA}},S_{\text{DNA}}\right) \tag{2}$$

$$G_{\text{RNA}}=\left(V_{\text{RNA}},F_{\text{RNA}},P_{\text{RNA}},S_{\text{RNA}}\right) \tag{3}$$

$$G_{\text{protein}}=\left(V_{\text{protein}},F_{\text{protein}},P_{\text{protein}},S_{\text{protein}}\right) \tag{4}$$

According to the basic definition of cooperating entity grammar, the cooperating entity grammatical model of DNAs, RNAs and proteins can be denoted as

$$G=\begin{pmatrix} g_V\left(V_{\text{DNA}},V_{\text{RNA}},V_{\text{protein}}\right),g_F\left(F_{\text{DNA}},F_{\text{RNA}},F_{\text{protein}}\right), \\ g_P\left(P_{\text{DNA}},P_{\text{RNA}},P_{\text{protein}}\right),g_S\left(S_{\text{DNA}},S_{\text{RNA}},S_{\text{protein}}\right) \end{pmatrix} \tag{5}$$

In general, the alphabets of the three grammars are different, the function g_V could be concreted as the union operation, denoted by

$$g_V\left(V_{\text{DNA}},V_{\text{RNA}},V_{\text{protein}}\right)=V_{\text{DNA}}\cup V_{\text{RNA}}\cup V_{\text{protein}} \tag{6}$$

The function g_F is the operation on the organizers of the three grammars. Overlap among the elements of the three sets of operations requires deletion of the iterative operations and addtion of new operations describing new structures. If the set of new operations is denoted by $F_{\text{DNA-RNA-protein}}$, the function g_F can be denoted by

$$g_F\left(F_{\text{DNA}}, F_{\text{RNA}}, F_{\text{protein}}\right) = F_{\text{DNA}} \cup F_{\text{RNA}} \cup F_{\text{protein}} \cup F_{\text{DNA-RNA-protein}} \tag{7}$$

Similarly, function g_P can be denoted by

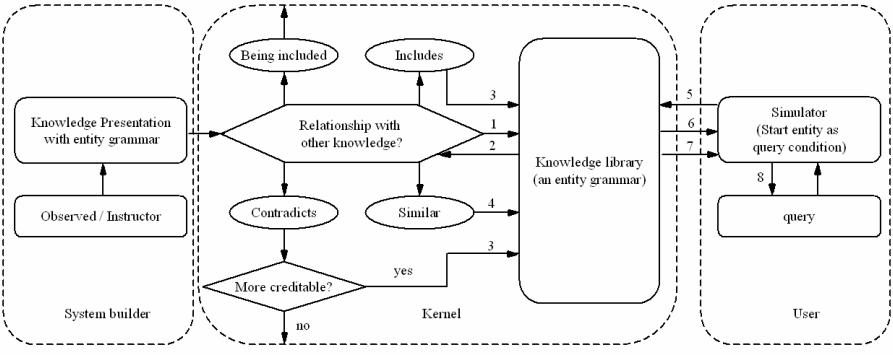
$$g_P\left(P_{\text{DNA}}, P_{\text{RNA}}, P_{\text{protein}}\right) = P_{\text{DNA}} \cup P_{\text{RNA}} \cup P_{\text{protein}} \cup P_{\text{DNA-RNA-protein}} \tag{8}$$

Where, $P_{\text{DNA-RNA-protein}}$ is the rewriting procedures involving at least two kinds of molecules. The choice of the start entities of the cooperating entity grammar depends on the purpose of research. The elements of the start entities could be DNAs, RNAs, proteins or their complexes, only if they can be described by the operations in g_F. Such a modeling method in EGS provides the possibility for establishing the models of meta-systems.

3 Entity Grammatical Model-Based Learning of Cells

In this section, we present the basic ideas of knowledge representation and learning strategy implemention based on the formalism of EGS. The approach represents the statements of knowledge as entity grammars. The collection of knowledge is the cooperating entity grammars. Entity grammatical model-based learning includes four transforms: entity generalization (induction), entity specialization (deduction), entity similization (analogy) and entity dissimilization (analogy). The process of how knowledge transmutations modify a statement in EGS can be realized in the environment of multi-programming paradigms (*e.g. Mathematica*).

The kernel of the platform includes a knowledge library in the form of entity grammars and a process of knowledge collection to build the knowledge library(Fig. 1). It provides the mechanisms of knowledge integration and the interface for knowledge input. The simulator of the platform is both the interface for query input and the channel for communicating with the kernel.



1.Update the knowledge library with new information; 2. Search and retrieve knowledge; 3. Replace; 4. Induction; 5. Input start entities; 6. Deduction; 7. Analogy; 8. Answers

Fig. 1. The architecture of entity grammar based platform

The original knowledge library is empty. With the input of knowledge from instructor and the observed facts, the model would become more and more powerful. The relationships between input knowledge and initial knowledge mainly include five types: (1) The input being included by initial knowledge would be discarded. (2) The input including initial knowledge library would be updated. (3) The input of new information would be added to the initial knowledge library. (4) For the input similar to initial knowledge, the kernel will use the induction approaches to gain the new knowledge from the input and initial knowledge and then update the library. (5) The input contradicting with initial knowledge will be reevaluated and accepted if it is more credible than initial ones. Otherwise, the input will be discarded.

The knowledge library in this system is an entity grammar with the start entities being an empty set. The definition of the start entities is left to users according to their interested problems. The simulator will input them to the kernel for simulation.

4 Concluding Remarks

In this paper, entity grammar system is presented as formalism approach for integrating diverse data and knowledge in biological cells. The platform for modeling, learning and simulating biological cells using entity grammars is described. This formalism is open to all knowledge from different cells and most of the reasoning paradigms, including induction, deduction and analogy. The cooperating entity grammar provides the methods for real-time model learning and updating. This make a difference with many other formalisms to build preset models. The approach can be used to solve problems in biological complex systems. With the development of the related concrete techniques, the formalism approach will facilitate whole cell modeling and reasoning.

Acknowledgements. This work was financially supported by the Natural Science Foundation of China (No. 30500643).

References

1. Mendes, P.: GEPASI: A Software Package for Modeling the Dynamics, Steady States and Control of Biochemical and Other Systems. Comput. Appl. Biosci. (1993) 563-571
2. Tomita, M., Hashimoto, K., Takahashi, K., Shimizu, T. S., Matsuzaki, Y., Miyoshi, F., Saito, K., Tanida, S., Yugi, K., Venter, J.C., Clyde, A., Hutchison, C.A.: E-CELL: Software Environment for Whole-cell Simulation. Bioinformatics. (1999) 72-84
3. Schaff, J., Loew, L.M.: The Virtual Cell. Proceedings of Pacific Symposium on Biocomputing '99. (1999) 228-239.
4. Kyoda, K.M., Muraki, M., Kitano, H.: Construction of a Generalized Simulator for Multi-Cell Organisms and its Application to SMAD Signal Transduction. Pacific Symp. Biocomputing. (2000) 314-325
5. Shapiro, B.E., Levchenko, A., Mjolsness, E.: Automatic Model Generation for Signal Transduction with Applications to MAP-Kinase Pathways. in Foundations of Systems Biology, Kitano, H. (ed.), MIT Press, Cambridge, MA. (2001)
6. Wang, Y.: Entity Grammar Systems: A Grammatical Tool for Studying the Hierarchal Structures of Biological systems. Bull. Math. Biol. (2004) 447-471

Chance Discovery in Credit Risk Management: Estimation of Chain Reaction Bankruptcy Structure by Directed KeyGraph

Shinichi Goda and Yukio Ohsawa

School of Engineering, The University of Tokyo, 113-8656 Japan
sgoda@mwa.biglobe.ne.jp

Abstract. Credit risk management based on portfolio theory becomes popular in recent Japanese financial industry. But consideration and modeling of chain reaction bankruptcy effect in credit portfolio analysis leave much room for improvement. That is mainly because method for grasping relations among companies with limited data is underdeveloped. In this article, chance discovery method with directed KeyGraph is applied to estimate industrial relations that are to include companies' relations that transmit chain reaction of bankruptcy. The steps for the data analysis are introduced and result of example analysis with default data in Kyushu, Japan, 2005 is presented.

Keywords: chance discovery, credit risk, chain reaction, bankruptcy.

1 Introduction

Credit risk management based on portfolio theory becomes popular in recent Japanese financial industry promoted by introduction of BIS regulation and increasing use of model-based loan decision making. Simulation method comes to be common tool for analysis of credit portfolio and simulation models have been developed for credit risk management [1,2,3,4]. However, there still remain major areas for improvement. Analysis on chain reaction bankruptcies is one of these areas.

2 Effect of Chain Reaction Bankruptcy

Chain reaction bankruptcies are common phenomenon. Its general definition is like "bankruptcy triggered by a preceding default of a company that has trade and other relations with the bankrupting company".

Most of industry experts regard it necessary to take the effect of chain reaction bankruptcy into accounts when they analyze profile of their credit risk portfolio. By introducing chain reaction factor into analysis, we can expect to better grasp the risk profile of credit risk portfolio since amount of loss caused by a default can be larger if there are other bankruptcies triggered by the default.

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 142–149, 2007.
© Springer-Verlag Berlin Heidelberg 2007

However, majority of simulation models actually used in business do not fully take the chain reaction into account. That is mainly because it is difficult to directly grasp relations among companies since available data is limited as;

1. Background information about bankruptcies publicly available is very limited, and it is not organized in way that can be used for statistical analysis.
2. Corporate data related to relations among corporation is very limited to public. Few companies make trade, financial and technological relations public, but very few with numbers.

3 Current Methods to Grasp Relations

Adjustments have been devised in models and simulators to include relation effect, but there is much room for improvement. A major method to take the effect into account is to grasp relations among companies by measuring co-relations among movements of the security price of the companies (In this method, a company's security price is regarded as representative of default probability of the company). But this method is applicable only to security issuing large companies whose number is very small. On the other hand, most of companies in a credit risk portfolio are non-security-issuing and small-mid size.

Another way is to make industry groups represent companies that belong to them, and to estimate relations among companies by grasping relations among the industries. Co-relations are measured among indexes of securities issued by companies in the industry groups and are applied to non-security-issuing ones in the groups. But the idea to estimate relations among industries with co-relations among security prices seems to be unreasonable. The most of Japanese companies are non-security-issuing ones and there is not enough evidence of the relations among security-issuing companies being same as that of non-issuing ones.

4 Grasp Chain Reaction Bankruptcy Structure by Chance Discovery Method

We proposed a method that detects relationship among bankrupted companies, without direct information of trade relations i.e., by chance discovery method in our previous work [5]. (The basic idea of this method follows the method that used for the chance discovery of earthquake by Ohsawa, Y. [6,7]).)

In this method, we estimate trade and/or other relations among companies defaulted in a geographical area within a certain time period, by visualizing relations among industry groups that include the defaults with KeyGraph. The method is based on the assumptions as follows;

1. There should be relations that transmitted factors causing defaults among companies that defaulted in a geographical area within a certain time period.
2. The default transmitting relations among companies should be mainly based on and represented by relations among industry groups. As seen in the above

definition of chain reaction bankruptcy, "trade relation" is generally regarded as one of the most influential relations. Trade relation between a pair of industry is generally universal among companies in the paired industries.
3. Default transmitting relations among industries could be paths to transmit default of a company in an industry to other companies in other industries.

Suppose that cloth retailer A and cloth wholesaler B defaulted within a month successively in Kansai district. Suppose other sets of cloth retailer and wholesaler those located in same districts defaulted successively within a month repeatedly. We can estimate with high confidence that there were sets of trade relation between the cloth retailer and the wholesaler defaulted successively and that the sets of trade relation between the two companies caused the successive default, even if there is no public information of sets of trade relations between the two companies. We can estimate so based on expert's knowledge about cloth trading industry and on the observed default patterns analyzed with KeyGraph.

5 Methodology

5.1 Original Method

First, we explain the original method proposed in our previous work [5] and introduce its basic idea. Steps are as follows (see Table 1);

Step1. Preparation of data

1. Data of defaults: each default event has attributes of default date, geographical area in which the defaulted company is located and an industry to which it belongs.
2. Sorting: group the defaults by area and sort the events in each area by default dates.
3. Select companies that seemed to have triggered chain reaction.

Step2. Transformation of company data to industry data

Transform company data prepared in Step1. to industry data by replacing a company's name to a code of an industry to which the company belongs.

Step3. Transformation of data to sentence form

1. Make the default events grouped by area in sentence form by order of their default dates. Each event is denoted by industry name and spaced.
2. Form one sentence starting from a triggering company and ending at a company whose default date is after the default date of the starting company.

Step4. Discovery of relations among industries by KeyGraph

1. Extract co-occurrence of default events by using KeyGraph with sentences formed default data.

2. Interpret a result of KeyGraph and estimate relations among industries. It is important to read relations expressed in a KeyGraph with experts' knowledge.

Examples of experts' knowledge about factors that are supposed to work behind relations extracted by KeyGraph are as listed below.

a. Technological and business relations among industries
   (example) An automobile is made of steel, glass, tires and electronic parts.
b. Commonality of customers among industries
   (example) Consumers living in Kyushu region shop at cloth/food retailers and eat restaurant both located in Kyushu.
c. Ownership relation:

Table 1. Example of data and sentence

Step1. Data (1) sorted by date

No.	date of default	comp.	ind. code	area
1	2005/1/4	A	101	north
2	2005/1/15	B	210	south
3	2005/1/30	C	303	east
	⋮			
N	2005/12/28	ZZZ	330	north

Step2. Data (2) grouped by area, and sorted by date

No.	date of default	comp.	ind. code	area
3	2005/1/30	C	303	east
8	2005/3/20	H	440	east
	⋮			
1	2005/4/1	A	101	north
	⋮			
N	2005/12/28	ZZZ	330	north
	⋮			
9	2005/3/29	I	211	west

Step3. Transform to sentence form
Series starting from D1 = 101 110 418 ...
Series starting from D2 = 210 105 330 ...
Series starting from D3 = 303 240 440 ...
⋮

5.2 Time Order Method (with Directed KeyGraph)

The original method had two points to be improved;

1. Time order among defaults is not captured. That makes the estimation of causal relation among defaults difficult.
2. The criteria for selecting trigger defaults were not clear enough.
   In our previous work [5], we made a sentence with default events in a month that starts from a hypothetical trigger default that were selected from a list of defaults that was made by Japanese SME Agency, based on the size or impact of defaults.

We newly introduce time order method to deal with above points. The basic idea of the new method is to try to better distinguish the causal relations among defaults with the time order among them.

Table 2. Method for making time ordered pair of defalts

Default date		1/1	1/10	1/20	2/1	2/5	2/25	2/28		12/1	12/20	12/30
Name of company		A	B	C	D	E	F	G	···	X	Y	Z
Industry code		10	20	30	40	10	30	20		40	10	50
Pair	Start from 10(A)		S10_20	S10_30								
	Start from 20(B)			S20_30	S20_40	S20_10						
	Start from 30(C)				S30_40	S30_10						
	Start from 40(D)					S40_10	S40_30	S40_20				
	Start from 10(E)						S10_30	S10_20	···			
	···									···		
	Start from 20(W)									S20_40	S20_10	
	Start from 40(X)										S40_10	S40_50

With time order method, by making a sentence include only a pair of defaults with distinction of earlier event of the two, the time order among defaults can be expressed by KeyGraph with direction – we name it as "directed KeyGraph". Detailed steps are as follows (see Table 2);

1. Make a sentence of a pair of defaults.
2. Put "S" to industry code of the earlier default in a pair to distinguish it as a starting default.
3. Make series of pairs with a selected starting default indicated by "S", and ending defaults each of which follows the starting default within a period.
4. Select another default that occurred next to the first default as the second starting event and take step 3.
5. Repeat step 3. and 4. until all the defaults in the analyzed period are selected as starting events.
6. Make and include linking pairs to link the nodes those are captured as ending events but at the same time are starting ones. For example, in Table 2, default E, that is coded as 10, is included in pair "S20_10", "S30_10", "S40_10" as an ending default and also included in pair "S10_30", "S10_20" as a starting one.

When the paired data is analyzed with directed KeyGraph, starting defaults are indicated by "S" put on the industry code. When a node, "S20" for example, is linked to another node, "10", it means that defaults in the industry 20 occurred before defaults in 10, indicating that defaults in 20 trigger the defaults in 10. In case two nodes of same industry are linked, like "10 – S10", it means ending defaults in "10" are at the same time starting defaults.

6 Case Study – Analysis of Chain Reaction Structure of Bankruptcies in Kyushu, 2005

As a case study, we applied time order method described above to data of bankruptcies in Kyushu district, Japan, 2005. The reason for limiting area for analysis only to Kyushu district is simply to make the size of data controllable.

A. Contents of data

Samples are 343 defaults in Kyushu, a local part of the Japanese samples described below. About 10,400 pairs are made from the above selected 343 default samples.

1. Japanese samples, about 3,400 defaults, are randomly selected from the all defaulted companies that defaulted based on bankruptcy laws and published in the official gazette in Japan, 2005. The area consists of 9 districts including Kyushu. Samples are categorized in about 200 mid-level industry groups by author, based on the industry categories defined by Teikoku Databank, Ltd.
2. Period for pairing one starting default with ending ones is set to one month.
3. About 6,200 linking pairs are added.

B. Analysis

First, the data was analyzed by original method and then by time order method.

(1) Original method (see Fig. 1 for result)

With this KeyGraph by original method, we can understand that defaults in linked industries occurred in a close time period. But it is difficult to estimates whether the co-occurrence is based on chain reaction. That is partly because this graph lacks of information about causal relation among defaults.

(2) Time order method with directed KeyGraph (see Fig. 2 for result)

In Fig. 2, time orders among defaults in the industries are shown by arrows indicating causal relations among them. Nodes consist of starting defaults are indicated by "S_" put on industry codes. Time order is expressed by arrows from a node with "S_" to another node without "S_". Circled pair of two nodes, one with "S_", are in same industry. For example, an arrow from S_C61 goes to C63, showing defaults in C61 occurred before defaults in C63 in major case, not vice versa. C63 is linked to S_C63 and two nodes are circled. An arrow then goes from S_C63 to C62. It shows that defaults in C63 those occurred after defaults in C61 are, at the same time, defaults those occurred before those in C62. It indicates that defaults in C63 caused by those in C61 then triggered defaults in C62.

When we see Fig. 2 this way, we can better estimate causal relations among defaults. The number of arrows go out from C61 (= S_C61) and C62 (= S_C62) are greater than those go into C61 and C62, that indicates defaults in civil engineering/construction and civil engineering industry caused defaults in other industries in major case, not vice versa. The defaults in C61, C62 might then trigger defaults in variety of related industries like C54 (brick layer work), C77 (pipe work), T31 (transportation_trade), i.e. Arrows from S_W03 (textile/cloth/others wholesale) go to C62 and to C63, indicating defaults in W03, caused by depressed consumer spending for cloth, were source of defaults in C62 and in C63 other than decreased public construction work. Many arrows go out from R39, indicating defaults of super markets, caused by depressed consumer spending, triggered defaults of groceries, toy/accessory shops in the super markets and of electric machinery wholesalers who trade with the market.

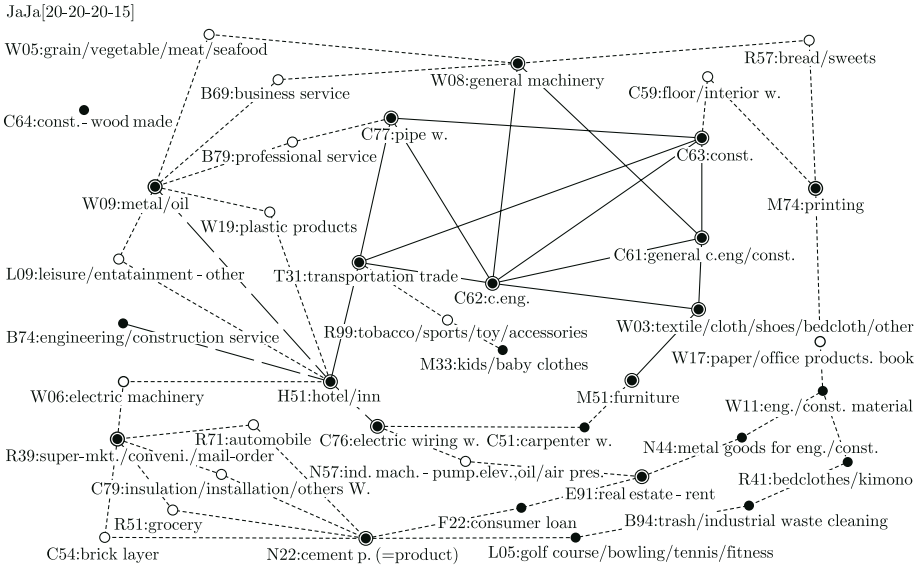


Fig. 1. KeyGraph by original method

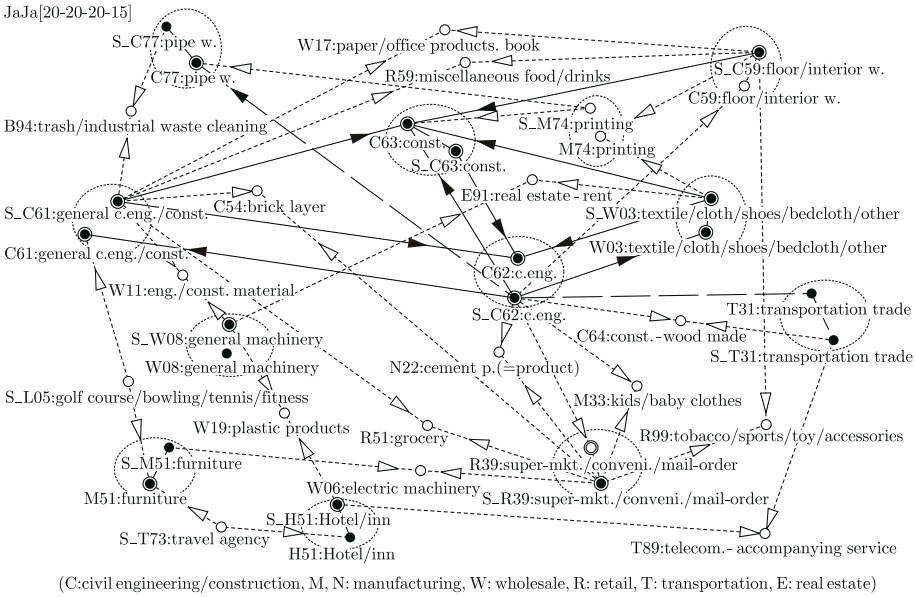


Fig. 2. KeyGraph by time order method ("S_" indicates starting default)

7 Conclusion

In this article, we applied chance discovery method to estimate structure of industrial relations that are to transmit chain reaction of bankruptcy. We introduced time order method with directed KeyGraph to capture/express time order of defaults and to better estimate causal relations among defaults. The result of the analysis of default data in Kyushu 2005 was promising. With further accumulation of analyses and improvement of method, a structure estimated by chance discovery method will sufficiently be a base for risk analysis and risk management of a credit portfolio.

The areas for further improvements are;

1. techniques for extracting appropriate time range between start/end of defaults
2. measurement of influence over a default probability of an industry/company of a default event to be transmitted through estimated industrial relations
3. modeling of estimated industrial relations in network structure for the use for risk analysis and risk management of a credit portfolio

References

1. Saunders, A.: Credit Risk Measurement (Japanese), Kinyu Zaisei Jijo Press, 2001.
2. FISC: "Working Report on Risk Management Model", 1999.
3. Nakabayashi, A. and Sasaki, M.: "Models for credit risk measurement and its application to Japanese bank", FRI Review, **2** No. 2, 1998.
4. Torii, H.: "Portfolio based credit risk management models – its effectiveness and limit", Weekly Kinyu Zaisei Jijo Magazine, June, 1998.
5. Goda, S. and Ohsawa, Y.: "Chance Discovery in Credit Risk Management – Estimation of chain reaction bankruptcy structure by chance discovery method", Proceedings of the 2006 IEEE International Conference on Systems, Man, and Cybernetics (6 Volumes), October 8–11, 2006.
6. Ohsawa, Y.: "Discovering risky subduction zones of earthquake", Information Technology for Chance Discovery, 314–325, Tokyo-Denki University Press, 2005.
7. Ohsawa, Y.: "Visualizing relations between chances and surrounding events", Information Technology for Chance Discovery, 121–153, Tokyo-Denki University Press, 2005.

Text Classification with Support Vector Machine and Back Propagation Neural Network

Wen Zhang[1], Xijin Tang[2], and Taketoshi Yoshida[1]

[1] School of Knowledge Science, Japan Advanced Institute of Science and Technology, 1-1 Ashahidai, Tatsunokuchi, Ishikawa 923-1292, Japan
{zhangwen,yoshida}@jaist.ac.jp
[2] Institute of Systems Science, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100080, P.R. China
xjtang@amss.ac.cn

Abstract. We compared a support vector machine (SVM) with a back propagation neural network (BPNN) for the task of text classification of XiangShan science conference (XSSC) web documents. We made a comparison on the performances of the multi-class classification of these two learning methods. The result of an experiment demonstrated that SVM substantially outperformed the one by BPNN in prediction accuracy and recall. Furthermore, the result of classification was improved with the combined method which was devised in this paper.

Keywords: text classification, SVM, BPNN, Xiangshan Science Conference.

1 Introduction

Automated text classification utilizes a supervised learning method to assign predefined category labels to new documents based on the likelihood suggested by a trained set of labels and documents. Many studies have been taken to statistical learning methods to compare their effectiveness in solving real-world problems which are often high-dimensional and have a skewed category distribution over labeled documents.

XiangShan Science Conference (XSSC) is well known in China as a scientific forum for the invited influential scientists and active young researchers about frontiers of scientific researches. The major discussions are summarized and posted at XSSC Web-site (http://www.xssc.ac.cn) and then aggregated into a large and valuable repository with all kinds of information which are related to scientific research and development in China. Some studies about XSSC have been undertaken from the perspective of enabling knowledge creation and supporting it by information tools [1-2]. Augmented information support (AIS) is one of those tools, for which Web text mining technologies are applied and basic tasks such as Web crawler, feature selection and indexing, and text clustering [3].

Performance examinations of statistical learning methods are usually carried out on the standard data set. Generally speaking, there is no superior algorithm in the statistical learning area for text classification problems. Even with the same classifier, different

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 150–157, 2007.
© Springer-Verlag Berlin Heidelberg 2007

performances may be revealed with different types of data sets because no statistical analysis was conducted to verify the impact of difference in the data on the performance variation of these classifiers [4]. For this reason the practical XSSC datasets are applied to examine the performance of SVM and BPNN.

The rest of this paper is organized as follows. Section 2 describes XSSC Web document representation and clustering of the datasets for XSSC data preprocessing. Section 3 describes the experiment design. The experimental results are presented in Section 4 and a comparison between SVM and BPNN is also conducted on the multi-class texts classification in the same section. The combined method is devised to integrate the results of SVM and BPNN and its performance is also demonstrated. Finally, concluding remarks and further research are given in Section 5.

2 XSSC Data Preprocessing

This section describes the preprocessing for the performance evaluation of both SVM and BPNN.

2.1 XSSC Web Documents Representation

Based on our prior work, 192 Web documents were collected from XSSC Website and a set of keywords for the collection of the whole documents was created to represent the Web documents. That is,

$$\text{Doc(i)} = (k_{i,1},...,k_{i,j},...,k_{i,m}),\ \text{let } k_{ij} = \begin{cases} 1, \text{ if keyword j exists in the } i\text{th document} \\ 0, \text{ if keyword j does not exist in the } i\text{th document} \end{cases} \quad (1)$$

m is the total size of the combined keywords collection.

Thus, 192 Boolean vectors were obtained to represent the 192 Web documents mentioned above initially. Secondly, a cosine transformation was conducted with these Boolean vectors to represent the Web documents more accurately and objectively. That is, let $\bar{k}_{i,j} = \dfrac{Doc(i) \bullet Doc(j)}{|Doc(i)||Doc(j)|}$ ($\bullet$ describes an inner product of vectors) and the representation vectors for the 192 Web documents were replaced with the newly generated immediate vectors using cosine transformation $\overline{Doc(i)} = (\bar{k}_{i,1}, \bar{k}_{i,2},..., \bar{k}_{i,192})$. Here, $\overline{Doc(i)}$ is the final and newly adopted representation vector for ith document. All following data preprocessing and the latter performance examination are all carried out on these transformed representation vectors.

2.2 Data Clustering

Clustering techniques are applied when there is no class to be predicted but the instances are required to be divided into natural groups. Two methods hierarchical clustering method and heuristic tuning method are applied to cluster the 192 documents into the predefined classes for a testing data set for the multi-class classification performance examination of SVM and BPNN in the next section. The testing data set is constructed by the following two steps:

Step 1: The similarity vectors representing the Web documents are processed by hierarchical clustering analysis in SPSS and a dendrogram is generated to describe the overall distribution of the documents on the given categories.

Step 2: Heuristic method is employed by manually adjustment on the references of documents clusters obtained in Step 1 to make them appropriately categorized i.e. to provide a standard classification for these documents.

Table 1 is the standard documents clustering generated by the above mentioned processing method. A skewed category distribution can be seen and a general trend of research focus currently existing among all the categories in XSSC can be drawn out. Here, 5 outliers were detected during clustering excluded from the following processing.

Table 1. Standard documents clustering on XSSC data set

Category ID	Subject of discipline	Total	Percentage
1	Life Science	60	31.25
2	Resource and Environment Science	31	16.15
3	Basic Science	21	10.94
4	Scientific Policy	16	8.33
5	Material Science	15	7.81
6	Transportation and Energy Science	11	5.48
7	Information Science	8	4.17
8	Space Science	6	3.13
9	Complexity Science	6	3.12
10	Outliers	5	2.60
11	Aeronautics & Astronautics	4	2.08
12	Micro-electronic Science	3	1.56
13	Safety Science	3	1.56
14	Other	3	1.56
Total	------	192	100.00

3 Experiment Design

The purpose of experiment design is to devise a controlled study on SVM and BPNN multi-class text classification performance. Regarding the skewed data distribution as shown in Table 1, the problem of the unbalanced data is dealt by assigning the number of training data and the number of test data in each class with same proportion.

3.1 Multi-class Text Classification Experiment Design

For multi-class text classification, four types of experiments are designed to test the performance of SVM and BPNN. Here three-class examination is conducted because the classification of more than three classes is very similar with three classes; on the other hand, the number of sample data is not enough to carry out classification more than three classes as well. The test strategy here is that the number of training sample is fixed as twice as the number of the testing sample whereas the categories from where the data set were selected are varied from each other. Table 2 shows the experiment design for the multi-class examination by using SVM and BPNN where

"30/20/20" means that 30 data samples were selected out randomly from "Life Science" as training data, 20 from category No.2 and 20 from category No.3 as training data. It can be seen that the numbers of training samples and test samples follow a decreasing trend because we also want to study on the performance of SVM and BPNN when the numbers of training and testing set are varying.

Table 2. Experiment design for multi-classification

Test No.	Test 1	Test 2	Test 3	Test 4
Selected Categories	No.1/ No.2/ Other classes	No.2/ No.3/ No.4	No.3/ No.4/ No.5	No.4 / No.5/ No.6
Numbers of Training data	30/20/20	20/14/11	14/11/10	11/10/8
Numbers of Testing data	15/10/10	10/7/5	7/5/5	5/5/3

3.2 SVM and BPNN Design

SVM is a classifier derived from a statistical learning theory by Vapnik and Chervonenkis (called VC theory hereafter) and it was firstly introduced in 1995 [5]. Based on VC theory and a kernel theory, SVM was proposed that was equivalent to solve a linearly constrained quadratic programming problem.

For multi-classification the one-against-all (OAA) method is adopted because of its same computation complexity with the one-against-one (OAO) in a SVM classifier and usually performs well [6]. Some deficiencies and improvement of this k-class (k>2) classification method were discussed in Ref. [7, 8] such as majority vote, SVM decision tree and so on. Here polynomial kernel $K(s,t) = ((s \bullet t)+c)^d$ (c=1,d=2) is used as the kernel function of SVM classifier because of its better learning ability compared with other kernels in the validation examination of our training data which is used to select the network of BPNN.

BPNN is a method known as back propagation for updating the weights of a multi-layered network undergoing supervised training [9, 10]. The back propagation algorithm defines two sweeps of the network: first a forward sweep from the input layer to the output layer and then a backward sweep from the output layer to the input layer. A three-layer fully connected feed-forward network which consists of an input layer, ahidden layer and an output layer is adopted here. With the previous basic training mentioned above, the "tansigmod" function is used in the hidden layer with 5 nodes and "purelinear" function for the output layer with 3 nodes [11]. The network of BPNN is designed as shown in Figure 1.

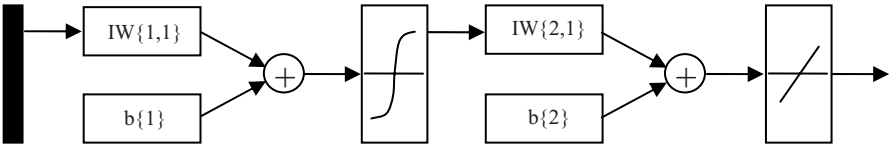


Fig. 1. BPNN with 5 nodes in hidden layer and 1 node in output layer

3.3 Combined Method

A combined method for the prediction of the unlabeled samples is designed to investigate if there are any improvements in prediction accuracy with combined results of SVM and BPNN. If the unlabeled sample is predicted with the same class by both SVM and BPNN, it would be then labeled by this predicted class. Otherwise, it will be given no label and cannot be assigned to any class. The accuracy of a combined method is calculated out by the following formula:

$$Accuracy(Combined\ Mehtod) = \frac{\left|S_{L(\mathrm{SVM})=L(\mathrm{BPNN})=L(\mathrm{Standard})}\right|}{\left|S_{L(\mathrm{SVM})=L(\mathrm{BPNN})}\right|} \tag{2}$$

Here, $S_{L(\mathrm{SVM})=L(\mathrm{BPNN})}$ is defined as the set of those samples that SVM and BPNN give the same predicted class label. By analogy, $S_{L(\mathrm{SVM,\,i})=L(\mathrm{BPNN,\,i})=L(\mathrm{Standard,\,i})}$ is defined as the set of those samples that SVM, BPNN and standard data set give the same class label. Then the *accuracy* means that to how much extent the prediction which the SVM and BPNN give the same label is reliable, i.e. the right answer for the unlabeled class.

4 Results of Experiments

Experiments were conducted according to the design described in section 4. We constructed the SVM and BPNN classifiers with the help of mySVM [11] and MatLab Neural ToolBox [12]. All the tests were conducted iteratively for 10 times and the average value of indicators was calculated to observe the performances of SVM and BPNN.

4.1 The Results of SVM and BPNN on Multi-class Text Classification

The results of SVM and BPNN on multi-class text classification are shown in Table 3. The indicators as accuracy and recall are employed to measure the classification of SVM and BPNN. Take the Test 1 for example, we obtained the accuracy as 0.7143 which comes from that 10 test samples from category No.1, 9 test samples from category No.2 and 6 test samples from category No.3 of all the 35 test samples designed in Test 1 mentioned in Section 3 were given the right labels. And recall as "10/16/19" means that 10 test samples were given label No.1, 16 as No.2 and 9 as No.3 in the multi-class classification in Test 1.

Table 3. Accuracies and recalls of SVM and BPNN on multi-class text classification

Classifier \ Test No.		Test. 1	Test .2	Test. 3	Test. 4
BPNN	Accuracy	0.7143 (10/9/6)	0.5909 (8/2/3)	0.4706 (2/3/3)	0.6923 (3/3/3)
	Recall	10/16/9	10/5/7	4/8/5	5/4/4
SVM	Accuracy	0.7714 (11/8/8)	0.6364 (9/3/2)	0.4706 (5/1/2)	0.8462 (4/4/3)
	Recall	14/11/10	11/7/4	9/3/5	5/4/4

From Table 3 it can be said that SVM has outperformed BPNN on the task of XSSC Web documents multi-class classification. The result from SVM classifier demonstrated convincingly better than that the one from BPNN on whatever the accuracies and recalls are.

4.2 The Result of the Combined Method

The combined method introduced in Section 4 was conducted with SVM and BPNN on multi-class text classification in order to examine the performance of our combined method. Table 4 shows the combined result of SVM and BPNN. Take Test 1 for example, the accuracy is 0.9200 which comes from that 25 test samples in this test is given the same label by SVM and BPNN and of them 23 test samples has the same labels as the standard data set. It can be seen that the combined accuracies were improved significantly in comparing SVM with BPNN with the contrast with the result shoed in Table 3. Also a particular comparison in accuracy between the combined method, SVM and BPNN is plotted in Figure 2.

Table 4. Accuracies of the combined method for multi-class text classification

Test No. / Classification	Test 1	Test 2	Test 3	Test 4
Multi-class classification	0.9200(23/25)	0.6875(11/16)	0.5714(4/7)	0.8889(8/9)

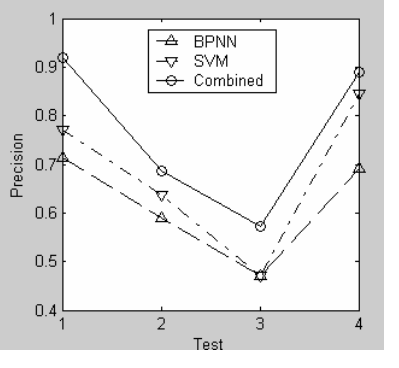


Fig. 2. Accuracies in the combined method, SVM and BPNN on multi-class text classification

5 Concluding Remarks

In this paper some experiments are carried out for multi-class text classification by using SVM and BPNN. Unlike the usual performance examinations the data sets of the experiments taken here are from a real practical application, that is, XSSC data set. A combined method of synthesizing the results from SVM and BPNN is developed to study whether there is an improvement in accuracy if the prediction result from different classifiers is combined. The experimental results demonstrated that

SVM showed a better performance than BPNN on the measure of accuracy and recall. The adaptation of the combined method achieved the improvement of accuracy for the multi-class text classification task.

Although the experimental results have provided us with some clues on text classification, a generalized conclusion is not obtained from this examination. Our work is on the initial step and more examination and investigation should be undertaken for more convincing work

One of the promising directions in text mining field is concerning the predictive pattern discovery from large amount of documents. In order to achieve this goal, we should introduce not only the required learning algorithms but also the semantics into the text mining field. More attention will be concentrated on the areas of semantic Web and ontology-based knowledge management, especially on the work that employs ontology to describe the existing concepts in a collection of texts in order to represent documents more precisely and explore the relationships of concepts from textual resources automatically.

Acknowledgments

This work is partially supported by the National Natural Science Foundation of China under Grant No.70571078 and 70221001 and by Ministry of Education, Culture, Sports, Science and Technology of Japan under the "Kanazawa Region, Ishikawa High-Tech Sensing Cluster of Knowledge-Based Cluster Creation Project".

References

1. Tang, X.J., Liu, Y.J., Zhang, W.: Computerized Support for Idea Generation during Knowledge Creating Process. In: Khosla, R. J. Howlett, and L. C. Jain (eds.): Knowledge-Based Intelligent Information & Engineering Systems (proceedings of KES'2005, Part IV), Lecture Notes on Artificial Intelligence, Vol.3684, Springer-Verlag, Berlin Heidelberg (2005) 437-443.
2. Liu, Y.J., Tang, X.J.: Developed computerized tools based on mental models for creativity support. In: Gu, J. F. *et al*.(eds.): Knowledge and Systems Sciences: toward Knowledge Synthesis and Creation. (Proceedings of KSS2006), Lecture Notes on Decision Support, Vol. 8, Global-Link, Beijing (2006) 63-70.
3. Zhang, W., Tang, X.J.: Web text mining on XSSC. In: Gu, J. F. *et al*.(eds): Knowledge and Systems Sciences: toward Knowledge Synthesis and Creation. (Proceedings of KSS2006), Lecture Notes on Decision Sciences, Vol.8, Global-Link, Beijing (2006) 167-175
4. Yang, Y.M., Lin, X.: A re-examination of text categorization methods. In: Proceedings on the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Berkeley,C A, (1999) 42-49..
5. F., Mulier.: Vapnik-Chervonenkis (VC) Learning Theory and Its Application. IEEE Trans on Neural Networks. 10(5) (1999) 5-7
6. Christophier, J., C, Burges.: A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery. Kluwer Academic Publisher, Boston (1998) 121-167

7. Rennie, J. D., Rifkin, R.: Improving Multi-class Text Classification with the Support Vector Machine. Master's thesis. MIT (2001)
8. Weston, J., Watkins, C.: Multi-class support vector machines. In Proceedings ESANN. Brussels (1999)
9. Rob, C.: Artificial Intelligence. Palgrave Macmillan. New York (2003) 312-315.
10. Rumelhart, D. E., Hinton, G. E. and Williams, R. J.: Learning internal representations by error propagation. In Parallel Distributed Processing, Exploitations in the Microstructure of Cognition, Vol. 1. Cambridge, MA: MIT Press (1986) 318-362
11. Stefan, R.: mySVM-Manual. Online: http://www-ai.cs.unidortmund.de/software/mysvm. (2000)
12. Neural Network Toolbox for MATLAB. Online: http://www.mathworks.com/products/neural-net/

Construction and Application of PSO-SVM Model for Personal Credit Scoring

Ming-hui Jiang and Xu-chuan Yuan

School of Management, Harbin Institute of Technology, 150001 Harbin, China
{jiangmh@cast.cn,yuanxuchuan}@126.com

Abstract. The parameters of support vector machine (SVM) are crucial to the model's classification performance. Aiming at the randomicity of selecting the parameters in SVM, this paper constructed a PSO-SVM model by using particle swarm optimization (PSO) to search the parameters of SVM. The model was used for personal credit scoring in commercial banks and particles' fitness function was used to control the type II error which costs huger loss to commercial banks. Compared with BP NN, the application results indicate that PSO-SVM gets higher classification accuracy with lower type II error rate and the model shows stronger robustness, which presents more applicable for commercial banks to control personal credit risks.

Keywords: support vector machine, particle swarm optimization, personal credit scoring.

1 Introduction

Support vector machine (SVM) [1] developed as a new machine learning method, compared with artificial neural networks, is more applicable for small samples classification based on statistic learning theory (SLT). SVM has widely used in text recognition, face recognition, image compression and so on for the model has no rigid requirements on variables' distribution and it performs well on generalization based on structural risk minimization (SRM). Through applying kernel functions to map the sample data from a low-dimensional feature space to a high-dimensional space, SVM constructs an optimal separating hyper-plane to classify the data. However, the parameters of SVM which are crucial to the model's classification performance are selected randomly or by cross validation which is time consuming. So searching optimal parameters is crucial to use SVM successfully.

The particle swarm optimization [2] (PSO) is a parallel evolutionary computation technique which has been widely used in function approximation, pattern recognition as well as neural network's training. In this paper, we make a combination of PSO and SVM to construct a PSO-SVM model by using PSO to search SVM's parameters, and then we use the model for personal credit scoring. In the end, we make a comparison between PSO-SVM and BP NN on classification results to examine the model's classification performance.

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 158–161, 2007.
© Springer-Verlag Berlin Heidelberg 2007

2 Construction and Application of PSO-SVM

2.1 Samples and Variables

In this paper, we use the personal credit data from one commercial bank in Shenzhen. The input variables figured with x_i and one output variable y listed in Table 1.

Table 1. Inputs and outputs

variable	index	value
x_1	Education	elementary 1; middling 2; advanced 3
x_2	Monthly Income	1: 1000~5000; 2: 5001~10000; 3: 10001~15000; 4: 15001~20000; 5: 20001~30000
x_3	Organization Character	national department 1; scientific education cultural and healthy department 2; trade and business 3; post and communication 4; financial and insurance 5; social service 6; supply of water, electricity and gas 7; industry 8; real estate 9; other 10
x_4	Career	manager 1; technique 2; officer 3; jobless 4; other 5
x_5	Spouse	yes 1; no 2
x_6	Loan Amount	1:6000~10000; 2:10001~50000; 3:50001~100000; 4:100001~300000; 5:30001~500000; 6: 500000~600000
x_7	Time limit	1: 24~36; 2: 37~49; 3: 50~60
x_8	Return mode	actual value 1; corpus 2
x_9	Surety	pledge 1; impawn 2; other 3
x_{10}	Age	1: 20~30; 2: 31~40; 3: 41~50; 4: 51~60
y	Default or not	yes -1; no 1

As the collectivity has many samples, which are complex and different, we use equal proportion allocation in stratified random sampling here. First, we divide the collectivity with the value y into two groups and then we extract about 500 samples respectively, thus to make the proportion of y=-1 and y=1 to be nearly 1:1. Finally, we get 1057 samples for the model, with 528 training samples composed of 257 default ones and 271 non-default ones and 529 testing samples composed of 248 default ones and 281 non-default ones.

2.2 Construction and Application of PSO-SVM

In practice of credit scoring, there exist two types of errors: type I error which mistakes the good applicants as bad ones and refuses to offer the loan and type II error which supplies loan to applicants with bad credit. The loss caused by type II error is huger, so we should concern more about type II error than type I error.

According to the principle of SVM [1], we know the parameters need to be selected include the kernel function's parameter and the penalty parameter C. In order to control the type II error, we set the penalty parameters of type I and type II errors as C_1 and C_2 respectively with the relation $C_2=K*C_1$, $K>=1$. Among four commonly used kernel functions, RBF kernel function has been used widely, for it has been proved with strong non-linear mapping and there is only one parameter in the kernel.

So we choose RBF function as the kernel. So the parameters in PSO-SVM need to be optimized are σ^2, C_1 and K. We set the searching range as [1, 100].

In order to get a better performance, we use a modified PSO algorithm with a inertial weight w [3], that is

$$v_{ij}(t+1) = w \cdot v_{ij}(t) + c_1 \cdot r_1 \cdot (p_j(t) - x_{ij}(t)) + c_2 \cdot r_2 \cdot (g_j(t) - x_{ij}(t)) \quad . \tag{1}$$

$$x_{ij}(t+1) = x_{ij}(t) + v_{ij}(t+1) \quad . \tag{2}$$

The inertial weight w is linearly decreasing [4] according to Formula (3), that is

$$w_t = w_{\max} - \frac{w_{\max} - w_{\min}}{t_{\max}} * t \quad . \tag{3}$$

Here we set $w_{\max}$ as 0.9 and $w_{\min}$ as 0.4 and the maximum iteration as 100 to terminate the algorithm. We set the population size m as 20, the acceleration constants c_1 and c_2 as 2 [5], $v_{\max}$ as 0.4. In order to control the type II error efficiently, we set PSO to maximize the particle's fitness function defined as Formula (4):

$$f = M(1 - \frac{n_1}{m_1} - k * \frac{n_2}{m_2}) \quad . \tag{4}$$

where, n_1 and n_2 represent the number of type I error and type II error respectively; m_1 and m_2 represent the number of non-default samples and default samples respectively. So n_1/m_1 and n_2/m_2 represent type I and type II error rates respectively; k is a variable and here we set k>1. Through the experiments with different numbers, we choose k=2 in the final program. M is set as 100 used to make the change of the fitness observable.

Using MATLAB to run the program of PSO-SVM and after 100 iterations we get the optimal parameters as σ^2=85.351, C_1=10.685, K=4.677 and C_2=K*C_1=49.497. The classification results of PSO-SVM are shown in Table 2.

In order to make a comparison, here we construct a forward three-layer BP NN model with 10 neurons in input layer, 1 neuron in output layer and 7 neurons in hidden layer which is commonly used in personal credit scoring for the same samples. Besides, the transfer functions in hidden and output layers are *tansig* and *logsig*. We set the epoch as 1000 to terminate the training and the performance function as *MSE* (Mean Square Error). We choose an improved BP algorithm with an adaptive learning rate and a momentum to train the network. The classification results of BP NN are shown in Table 2 with the critical value 0.5.

Table 2. Classification results

Model	Training samples			Testing samples		
	Type I error	Type II error	Accuracy	Type I error	Type II error	Accuracy
PSO-SVM	24(8.86%)	2(0.78%)	95.08%	26(9.25%)	5(2.02%)	94.14%
BP NN	6(2.21%)	2(0.78%)	98.86%	27(9.61%)	12(4.84%)	92.63%

3 Result Analysis

From Table 2, we can see on training samples, the classification accuracy of PSO-SVM is lower than that of BP NN, but on testing samples, PSO-SVM gets a higher accuracy than BP NN does. From the other aspect, we can see the difference on classification accuracy of BP NN is 98.86%-92.63%=6.23%, however, PSO-SVM is only 95.08%-94.14%=0.94%. This indicates PSO-SVM gets a better performance on robustness than BP NN. In a dynamic environment of personal credit, a model with strong robustness is more significant for commercial banks to keep away from credit risks and expand the personal credit market. From this aspect, we can say PSO-SVM is more applicable than BP NN. Besides, from the results of the two types of errors, we can see PSO-SVM classifies the default samples with lower type II error rate both on training and testing samples, especially on testing samples on which the result of PSO-SVM is much lower than that of BP NN. The results indicate that the fitness function of PSO is effective to control the type II error which is more significant for commercial banks to keep away from personal credit risks. The reason is that in order to increase the fitness of the particle in PSO, SVM enhances the punishment on type II error to achieve the goal.

4 Conclusions

This paper makes a combination of PSO and SVM to construct a PSO-SVM model and used the model for personal credit scoring. We get the conclusions from the empirical results: first, using PSO's global search to optimize the parameters of SVM, we can get the optimal parameters of SVM and get over the randomicity; second, through particle's fitness function in PSO, we can control the type II error which brings huger loss, that is more significant for commercial banks to keep away from credit risks; third, compared with BP NN, the difference of accuracies on training samples and testing samples of PSO-SVM is smaller, that indicates the model gets a stronger robustness. In a word, we can use the PSO-SVM model for personal credit scoring to get a good result.

References

1. Vapnik V N.: The nature of statistical learning theory. Springer-Verlag, New York (1995)
2. Kennedy J, Eberhart R C.: Particle Swarm Optimization. Proceedings of IEEE International Conference on Neutral Networks, Perth, Australia (1995) 1942-1948
3. Shi Y H, Eberhart R C.: A Modified Particle Swarm Optimizer. IEEE International Conference on Evolutionary Computation, Anchorage, Alaska (1998) 69-73
4. Shi Y H, Eberhart R C.: Empirical study of particle swarm optimization. Proceedings of the 1999 Congress on Evolutionary Computation (1999) 1945-1950
5. Shi Y H, Eberhart R C.: Parameter Selection in Particle Swarm Optimization. Proceedings of the Seventh Annual Conf. on Evolutionary Programming (1998) 591-601

Feature Description Systems for Clusters by Using Logical Rule Generations Based on the Genetic Programming and Its Applications to Data Mining

Jianjun Lu[1,2], Yunling Liu[1], and Shozo Tokinaga[2]

[1] China Agricultural University, Beijing 100083, China
[2] Graduate School of Economics, Kyushu University, 812-8581 Japan

Abstract. This paper deals with the realization of retrieval and feature description systems for clusters by using logical rule generations based on the Genetic Programming (GP). At first, whole data is divided into several clusters and the rules are improved based the GP. The fitness of individuals is defined in proportion to the hits of corresponding logical expression to the samples in targeted cluster c, but also in inversely proportion to the hits outside the cluster c. The GP method is applied to various real world data by showing effective performance compared to conventional methods.

1 Introduction

This paper deals with the feature description systems for clusters by using rule generations based on the Genetic Programming (GP) and its applications to data mining [4][5]. In the method, we prepare various kinds of logical expression (having tree structure and are called as individuals) for the data (called as samples in the following) in the underlying cluster by using variables for categorical values, then we improve the logical expressions by using the GP. As the fitness of each individual, we use the number of hits of individual (cases where the logical expression corresponding the individual) for the samples in cluster[1]-[3].

As applications, we apply the method to the evaluation of decision making of personal loan by showing the effectiveness of the method. Moreover, the feature description system is applied to eight groups of samples which are arbitrarily collected from various data bases.

2 Feature Description System for Cluster Based on the GP

The overview of the feature description system for cluster based on the GP treated in the paper is shown in the following [4].
(1) Description of samples by categorical variables

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 162–165, 2007.
© Springer-Verlag Berlin Heidelberg 2007

Generally, there are two kind of variables (numerical variables and logical variables) to characterize the sample, but we assume that in the system we have only logical variables. The numerical variables are transformed into categorical variables by using conventional methods of discretization (details are omitted here)[6].

(2) generation of initial individual for logical expressions

It is assumed that the logical expressions represented by the categorical variables are used to describe the feature of samples in the cluster which correspond to the individuals. For the sake of tractability, it is assumed that the logical expression has a binary tree structure. At the beginning of GP procedure, we generate the initial pool of individuals (say, 1000 individuals) by using random numbers.

(3) definition of fitness of individuals

The definition of the fitness of k th individual in the GP procedure is given by the number of hits denoting the number of samples in the cluster to which the logical expression corresponding to the individual is true. Moreover, the number of hits for the samples outside the underlying cluster c in the whole dataset is also used. We at first defines following index.

$$y_k = T - h_k^2/n_k \tag{1}$$

The notations included in equation (1) are defined as follows.

n_k : number of samples in the whole dataset for which the logical expression of the individual k is true

h_k : number of samples in the underlying cluster c for which the logical expression of the individual k is true

T : total number of samples in the cluster c

The fitness of individual k (denoted as f_k) is defined by adding a certain positive number to y_k, and taking the inverse of the number as.

$$f_k = (a + y_k)^{-1} \tag{2}$$

If the number of hits for logical expression covering samples in the cluster becomes larger, then accordingly, the measure defined in equation (1) becomes to be close to zero.

3 Logical Rules for Feature Description and the GP

The prefix representation is equivalent to the tree representation of arithmetic expressions[6][7]. For example, we have the next prefix representation.

$$(6.43x_1 + x_2)(x_3 - 3.54) \longrightarrow \times + \times 6.43x_1x_2 - x_3 3.54 \tag{3}$$

We apply the GP procedure to approximate the functions f(,) using the observations y_n.

To keep the consistency of genetic operations, the so-called stack count (denoted as *StcckCount*) is useful. The *StackCount* is the number of arguments it

places on minus the number of arguments it takes off the stack. The cumulative *StackCount* never becomes positive until we reach the end at which point the overall sum still needs to be 1. The basic rule is that any two loci on the two parents genomes can serve as crossover points as long as the ongoing *StackCount* just before those points is the same. The crossover operation creates new offsprings by exchanging sub-trees between two parents.

Usually, we calculate the root mean square error ($rmse$) between $x(t)$ and $\tilde{x}(t)$, and use it as the fitness. The fitness S_i of i th individual is defined as the inversion of $rmse$.

We assume that the all of samples are characterized by the logical variables, then we use a relatively simple method to generate logical expressions. Assume that there are categorical variables $v_1, v_2, ..., v_m$, and these variables can have the values $s_1, s_2, ...$ For example, if the logical variables v_1, v_2 take the value s_3, s_5, then we have

$$v_1 = s_3, v_2 = s_5 \tag{4}$$

Then, these binary expressions are used as predicates included in the logical expressions in the GP. For example, we define new logical variables X_{kj} represented by input variable v_i such as

$$X_{kj} = \begin{cases} True, & \text{if } v_k = s_j \\ False, & \text{otherwise} \end{cases} \tag{5}$$

We also define the fitness of individuals as the accuracy of rule generated by the rule corresponds to the underlying individual. To improve the fitness of individuals, we apply the GP operations to the logical expressions.

4 Applications

4.1 Applications to German Credit Data

The experiment on real-life credit-risk evaluation is carried out using the German credit data. The German credit data is obtainable from the website. The data consist of 1000 records of personal loan, and the input variables for one record include 7 numerical data and 13 categorical data.

Even though the original purpose of the dataset is the generation of accept/deny rules for personal loans, but we use the dataset to examine the ability of the GP method of the paper. At first, we select 100 samples at random from the dataset, and classify them into three clusters by using following seven numerical variables based on the conventional software package.

Then, we assume one cluster (say cluster c) is the target cluster whose feature must be described, and other 6 samples belonging two another clusters are regarded as samples outside cluster c.

It is concluded that after about 500 generation of the GP procedures the feature extraction (description) is completed, and the logical expression finally obtained describes the true feature of cluster c.

4.2 Applications of Feature Description to Real Data

In the following, we explain the simulation studies of the feature description of the paper applied to multiple real dataset, and discuss the performance in the mean.

At first, we select about 100-300 samples from the dataset at random, and then we divide these samples into three cluster by using conventional numerical method of clustering. Then, at the next step, we apply the GP procedure of the paper for the extraction of cluster and feature description to these three clusters, independently. That means if we focus on the cluster c, the samples belonging two clusters d different from the underlying cluster c are regarded as the samples outside the cluster c. The number of GP generation N_F necessary to obtain final result of feature description is omitted. The GP procedure to extract the clusters and to give feature descriptions can work effectively by spending 500 or 600 GP generations even for real world data, despite wide variations. We must note that by using the feature description method of the paper, finally we obtain 100% correct classification of samples to the underlying clusters.

5 Conclusion

This paper treated the realization of retrieval and feature description systems for clusters by using logical rule generations based on the GP. As applications, the GP method was applied to various real world data. For future works, it is necessary to apply the method of transformation of logical expressions to natural language. Further works by the authors will be continued.

References

1. G.Piatetsky and W.J.Frawley,"Knowledge discovery in database:An overview,"in Knowledge Discovery in Database, AIII/MIT Press, 1991.
2. A.A.Freitas, Data Mining and Knowledge Discovery with Evolutionary Algorithms, Springer-Verlag, 2002.
3. S.Tokinaga,J.Lu and Y.Ikeda, "Neural network rule extraction by using the Genetic Programming and its applications to explanatory classifications," IEICE Trans.Fuadamentals, vol.E88-A,no.10,pp.2627-2635,2005.
4. M.L.Wong and K.S.Leung,Data Mining Using Grammer Based Genetic Programmign and Applications,Kluwer Academic Publisher, London,2000.
5. J.Lu, Y.Kishikawa and S.Tokinaga,"Realization of Feature Descriptive Systems for Clusters by using Rule Generations based on the Genetic Programming and its Applications(in Japanese)," IEICE Trans.Fuadamentals, vol.J89-A,no.12,pp.2627-2635,2006.
6. J.Lu, S.Tokinaga, and Y.Ikeda, gExplanatory Rule Extraction Based on the Trained Neural Network and the Genetic Programmingh, Journal of the Operations Research Society of Japan, Vol.149, No.1, pp.66-82, 2006.
7. Y.Ikeda and S.Tokinaga,"Chaoticity and fractality analysis of an artificial stock market by the multi-agent systems based on the co-evulutionary Genetic Programming", IEICE Trans.Fundamentals, vol.E87-A,no.9,pp.2387-2394, 2004.
8. J.R.Koza,Genetic Programming,MIT Press, 1992.

Artificial Immunity-Based Discovery for Popular Information in WEB Pages

Caiming Liu, Xiaojie Liu, Tao Li, Lingxi Peng, Jinquan Zeng, and Hui Zhao

School of Computer Science,
Sichuan University
610065 Chengdu, China
liucaiming@gmail.com

Abstract. An artificial immunity-based discovery method for popular information is proposed. Principles of evolution and concentration of antibodies in artificial immune system are simulated. Key words in web pages are extracted and simulated as antibody and antigen. Antibodies are evolved and excreted dynamically. Concentration of antibodies is computed to attain accurately the degree of popular measurement in quantity. The proposed method improves the intelligent degree of information discovery and provides a new way to discover WEB information.

Keywords: artificial immune, popular information, information discovery, antibody, concentration.

1 Introduction

WEB pages comprise a great deal of information. How to extract useful knowledge from WEB pages is focused by researchers. Intelligent methods are needed to compute degree of popular measurement in quantity.

Artificial immune system (AIS) which simulates the characters of learning and self-adaptation [1, 2] of biological immune system uses the way of clone selection and mutation principles to recognize foreign harmful antigens [3, 4]. Furthermore, if an antigen damages body greatly, the antibody which recognizes the antigen excretes many similar antibodies [5, 6]. Concentration of this kind of antibody will increase quickly. The antibody concentration can be used to judge the harmful measurement of relative antigen [7].

In this paper, the former principles is used for reference. Key words in the WEB pages are extracted to be simulated as antibodies which will be evolved in our model automatically. Their concentration is computed to judge the degree of popular measurement of WEB information.

The rest of the paper is organized as follows. In Section 2, the proposed principle is introduced. In Section 3, simulations and experimental results are showed. Finally, Sections 4 contains our conclusions.

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 166–169, 2007.
© Springer-Verlag Berlin Heidelberg 2007

2 Proposed Principle of Information Discovery

2.1 Information Simulation

Key words are extracted to find popular information hid in WEB pages. The key words extracted from WEB pages are simulated as antigens. Let $U = \bigcup_{i=1}^{\infty} G^i$ be the state space of antigen, where is Chinese or English word library.

Let antibody represent information of key words. Antibody is defined as $D = \langle Keyword, age, count, type \rangle$. D is a four-tube group. It is consisted of key word information (*keyword*), the live generations of antibody (*age*), the amount of antigens which antibody matches (*count*) and the classification of the information (*type*).

Antibody is consisted of mature antibody and memory antibody. Mature antibody needs to match enough number (δ) of antigens in its lifecycle λ_M to evolve into memory antibody. Or else, it dies. The type of antibody which is extracted in the first time is the key word self. Mature antibody *M* is described in equation (1).

$$M = \{x | x \in D \wedge x.age < \lambda_M \wedge x.count < \delta\} \tag{1}$$

Where, λ_M is the lifecycle of antibody, δ is threshold of evolution.

Memory antibody *R* is evolved by mature antibody. It denotes that information emerges frequently recently. Memory antibody can live in a long lifecycle λ_R even it doesn't match any antigens in long period of time.

2.2 Evolution of Antibody

Mature antibody is extracted from WEB pages newly or generated by clone selection to recognize new information. If mature antibody can not match enough antigens in its lifecycle, it tends toward death. Once its matching count arrives at the threshold δ (i. e. $count \geq \delta$), it will be activated and evolved into memory antibody. During the period λ_M, some mature antibodies match few antigens. This kind of antibody will die soon. If a key word appears in a high frequency, matching count of its relative antibody will be accumulated to threshold δ sharply. Its relative antibody will be evolved into memory quickly. This represents the key word is popular.

Memory antibody denotes a key word is popular. It has long lifecycle λ_R. However, λ_R is not infinite. The past popular information may not appear again or emerge in low frequency in future. When a memory antibody can not match any key words in the period λ_R, its affinity *count* decreases in a time interval $T_{\text{int}erval}$ (*count* is decreased by 1). When the affinity is equal to 0, memory antibody is degenerated as mature antibody. If the memory antibody matches antigen again, its affinity increase (*count* is added by 1). If it becomes popular again, its affinity ascends quickly.

2.3 Degree Computation of Popular Measurement

To discover relative information, memory antibody excretes new mature antibody rapidly by the way of copy, mutation, and etc. By computing the concentration of antibody, the popular degree of relative information can be worked out. According to the type of antibody, its concentration is obtained. The concentration represents the popular degree. Equation (2) describes antibody concentration.

$$hr(r_i) = c\left(\frac{2}{1+e^{-\left(\sum r_i.type+\sum m_i.type\right)}} - 1\right) \tag{2}$$

Where, $hr(r_i)$ denotes the popular degree of information similar to memory antibody r_i, the lager the $hr(r_i)$, the relative key word is more popular; m_i is mature antibody and in the same kind of r_i; c is a constant number, $c > 0$.

3 Experimental Results

In our experiments, two portal web sites and two BBSes are chosen to test the efficiency of proposed method. Choices of monitor spots are shown in table 1. The four columns of table 1 represent the following parameters respectively: name of web site, lifecycle λ_M (see section 2.1) of mature antibody, lifecycle of memory antibody λ_R (see section 2.1) and degeneration time interval $T_{\text{int}\,erval}$ (see section 2.2) of memory antibody.

Table 1. Monitor spots and configuration of parameters

Monitor spot	λ_M	λ_R	$T_{\text{int}\,erval}$
Web site 1	One week	One day	Half day
Web site 2	One week	One day	Half day
BBS 1	Four days	Five hours	1.5 hours
BBS 2	Four days	Two hours	1.5 hours

A great lot of key words are extracted from pages of former web sites. These key words are simulated as antibodies. We monitor the four web sites from September to November in 2006. Through evolvement of antibody, concentration of antibody is worked out. According to the concentration, we obtain the popular degree of some information by equation (2), as is shown in figure 1. The larger popular degree is, the more popular the key word is. It is National Day of Chinese in October 1. So, people pay much attention to the key word “national day” whose concentration ascends sharply.

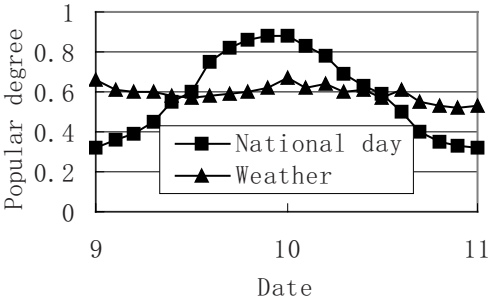


Fig. 1. The measurement of popular information

4 Conclusion

In this paper, key words in WEB pages are simulated as antibodies and antigens. Antibody evolution and clone selection principles in AIS are used to evolve antibodies dynamically. Degeneration of memory antibody is proposed to simulate popular information degradation. Concentration of antibody forms automatically. The concentration represents popular degree of information in quantity. The experimental results show that our method can get the popular degree of information efficiently.

Acknowledgments. This work was supported by the National Natural Science Foundation of China under Grant No.60373110, No.60573130 and No.60502011, the National Research Foundation for the Doctoral Program of Higher Education of China under Grant No.20030610003, the New Century Excellent Expert Program of Ministry of Education of China under Grant No. NCET-04-0870, and the Innovation Foundation of Sichuan University under Grant No.2004CF10.

References

1. Li, T.: Computer Immunology[M]. Publishing House of Electronics Industry, Beijing (2004)
2. Forrest, S., Hofmeyr, S. A., Somayaji, A.: Computer immunology[J]. Communications of the ACM, Vol. 40(10) (1997) 88-96
3. De Castro, L. N., Timmis, J. I.: Artificial Immune Systems: A Novel Computational Intelligence Approach[M]. London: Springer-Verlag, 2002
4. Hofmeyr, S. A., Forrest, S.: Architecture for an Artificial Immune System[J]. Evolutionary Computation, Vol. 8(4) (2000) 443-473
5. Hofmeyr, S. A., Forrest, S.: Immunity by design: an artificial immune system[C]. In: Genetic Evolutionary Computation Conf, San Francisco, CA, (1999) 1289-1296
6. Timmis, J., Neal, M., Hunt, J.: An artificial immune system for data analysis. Biosystems, Vol. 55(1/3) (2000) 143-150
7. Li, T.: An immunity based network security risk estimation. Science in China(Series E), Vol. 35(8) (2005) 798-816

Network Structure and Knowledge Transfer

Fangcheng Tang[1,2]

[1] School of Economics and Management, Tsinghua University
Beijing 100084, P.R. China
[2] School of Economics and Management, Xi'an Technological University
Xi'an 710032, P.R. China
fctang@163.com

Abstract. This study employs single layer perceptron model (SLPM) to explore how the topological structure of intra-organization networks affects knowledge transfer. The results demonstrate that in the process of knowledge transfer, both the disseminative capacity of knowledge senders and the absorptive capacity of knowledge receivers should be taken into consideration. While hierarchical networks can enable greater numbers of organizational units to acquire knowledge, they reduce the speed and efficiency of knowledge transfer, whereas scale-free networks can accelerate transfer of knowledge among units.

Keywords: Knowledge transfer, Hierarchical network, Scale-free network.

1 Introduction

This paper utilizes a single-layer perceptron neural network model (SLPM) [1][2] to simulate the knowledge transfer process in two network topologies of an organization. Specifically, our study focuses on the potentially different performance of knowledge transfer between two organizational network topologies: hierarchical and scale-free network which represent informal and formal structures in an organization. We argue that the nature of knowledge transfer within intra-organizational networks can be modeled with the tools of computational learning theory (CLT) as developed by computer scientists to study learning and problem-solving by machines. As the knowledge is transferred within an intra-organizational network, network nodes learn from their senders and become proficient at recognizing new and related knowledge.

2 Virtual Experiment Design

2.1 The Structure of Intra-organization Networks

An intra-organization network can be described by a graph theoretic approach in this form: G=(V,E), where Vdenotes the set of units and E is sets of the relationships in the networks.Let V=1,...,N denote a finite set of units.for $\forall\, i, j \in V$,

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 170–173, 2007.
© Springer-Verlag Berlin Heidelberg 2007

define the binary varible $\chi(i,j)$ taking value $\chi(i,j)=1$ if there is a connection between i and j, otherwise $\chi(i,j)=0$.

Thus, a symmetrical hierarchical structure for vertical organizations where the number of layers, the number of nodes in the first layer and the number of connections from each node to the next layer are set. A scale-free model which is defined by the Watts and Strogatz [3]. The network evolves into a scale-invariant state, the probability that a vertex has k edges following a power law with an exponent $\gamma_{model}=2.9\pm0.1$[4].

2.2 Node Status Variables

We assuming that individual in the organization can be characterized by three variables α,β,ρ. The first is a signal variable α, which demonstrates whether individual acquired specific knowledge, and value 1 means the individual successfully gets access to the knowledge and 0 means the opposite. The second variable β, which is independent randomized uniformly distributed on [0, 1], represents the capacity of the individuals to learn and absorb specific knowledge. Larger values mean substantial capacity, which can be determined by the genetics, background, or experience of these members. The last variable ρ shows the disseminative capacity of one member to transfer knowledge to others. Variable ρ is also random uniformly distributed on [0, 1] and members who have strong capacity to spread knowledge get high numerical values.

2.3 Interaction Mechanism and NN Model

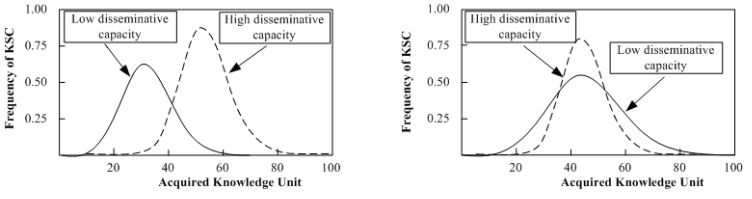
Here we use the SLPM of neural network to simulate the knowledge transfer. Every node in the network is regarded as a threshold logical cell, input $X=(X_1,X_2,...,X_N)$, each X_i equals to the output Ω_i of vertex i, shows the variable S of node i, $W=\{\omega_{1j},\omega_{2j},...,\omega_{vj}\}$ and ω_{ij} comes from the variable F_i , θ is the threshold value which is derived from the ingredients variable D of the node. This model can be interpreted as that a node which has difficulty D in comprehending the knowledge can only learn the knowledge under the condition that the sum of the F of the input node which already has the knowledge is bigger than D of this node, so output $\Omega_i=\{0,1\}$ is calculated as:

$$\Omega_j=f(\sum_{i=1}^{n}\omega_{ij}X_i-\theta_i),f(u_i)=\begin{cases}0 & \text{if } \sigma_i\leq 0,\\ 1 & \text{if } \sigma_i>0.\end{cases} \tag{1}$$

3 Results

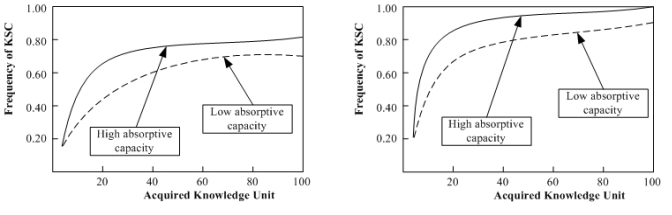
A comparison of results of the computational experiments is summarized in Fig. 1 and Fig.2. In Fig1 (a) and (b), as the disseminative capacity of the dominant individual is set at 1, the average number of knowledge-acquired individuals is increased, and the number of knowledge-acquired nodes is more with higher

absorptive capacity (dashed line) than that with lower absorptive capacity (solid line) in hierarchical and scale-free network. Regardless of high or low disseminative capacity, the frequency of KSC tends to increase at first and then decline as the number of knowledge-acquired individuals increases. But the extent of frequency of KSC increases with low absorptive capacity is less than that with high absorptive. High absorptive capacity enables the number of knowledge-acquired individuals to increase more, and it seems that the frequency of scale-free network follows the assumption of normal distribution.



(a) hierarchical network (b) scale-free network

Fig. 1. The relationship between the frequency of KSC and acquired knowledge member with different disseminative capabilities of knowledge holders



(a) Hierarchical Network (b) Scale-free Network

Fig. 2. Relationship between the frequency of KSC and acquired knowledge member with absorptive capacity

Fig. 2 shows that, regardless of the higher or lower absorptive capacity, the frequency of KSC of knowledge-acquired units in hierarchical and scale-free network increases. Simultaneously, the extent of frequency of KSC with high absorptive capacity increases is more than that with low absorptive capacity. In hierarchical networks, average knowledge state change with high absorptive capacity tends to be stable at interval $0.60 \sim 0.80$. In scale-free networks, however, the tendency to increase will reach the maximum value 1.00. In addition, the frequency of KSC with low absorptive capacity in scale-free networks tends to be more than that of hierarchical networks. Therefore, as an important channel to knowledge transfer, scale-free networks enable knowledge transfer to be more efficient than does hierarchical networks.

4 Conclusions

This study demonstrated that efficient and effective knowledge transfer must consider two important aspects: knowledge senders' disseminative capacity and knowledge receivers' absorptive capacity. Cohen and Levinthal argue that organizations have to build specific absorptive capacity to identify, assimilate, and exploit external knowledge [5]. But an organizations also need to build strong disseminative capacity to communicate their knowledge in such a way that knowledge workers in the network can easily understand and thus assimilate the required knowledge, hence to maximize the value of knowledge. A corollary of this finding is to understand the role of the dominant knowledge holder in the knowledge transfer process. Also, an interesting effect of the model is that 100 per cent knowledge transfer in an intra-organizational network is never reached. We believe that this situation is the result of interaction of nodes with different disseminative capacity and absorptive capacity. The implication for multiunit organizations is that 100 per cent knowledge sharing in work-teams or groups is rarely achieved.

But current study has never considered the different types of knowledge, e.g., tacit and explict knowledge. In addition, the strength of ties between the players is not considered. These interesting problems will be remained as future research.

Acknowledgments. I would like to thank the anonymous reviewers for their insightful comments on previous versions of this manuscript.I was also grateful to the funding of National Science Fund of China under contract No. 70602004 and Postdoctoral Foundation of China under contract No. 2005038072 for their general support.

References

1. Hertz J., Krogh, A., Palmer, R. G. Introduction to the theory of neural computation. Redwood City: Addison-Wesley (1991).
2. Lippman, R. P. An introduction to computing with neural nets. IEEE ASSP Magazine 4 (1987) 4-22.
3. Watts, D.J., Strogatz, S. H. Collective dynamics of small world networks. Nature 393 (1998) 440-442.
4. Barabśi, A-L., Albert, R., Jeong, H. Scale-free characteristics of random networks: the topology of the world-wide web. Physica A 281 (2000) 69-77.
5. Cohen, W., Levinthal, D. Absorptive capacity: A new perspective on learning and innovation. Administrative Science Quarterly 35 (1990) 128-152.

Information Relationship Identification in Team Innovation

Xinmiao Li[1], Xinhui Li[2], and Pengzhu Zhang[3]

[1] School of Information Management and Engineering,
Shanghai University of Finance and Economics,
200433, Shanghai, China
xinmiaolisunny@hotmail.com
[2] DongBeiWang, Haidian,
100094, Beijing, China
miaoxinhuili@gmail.com
[3] Antai Management School, Shanghai Jiaotong University,
200052, Shanghai, China

Abstract. In the 1990s, innovation has become the main resource of competitive advantage for firms. Computer-mediated team innovative activities provide much more information, thus resulting in information overload. In this paper, the method for identifying the affirmative and negative relationship between the solution and its comment is proposed. The identification model is set up and applied in team innovation. In the model, we improve the traditional feature extraction method from three aspects: classifying feature words into six levels according to their affirmative and negative tone, introducing the new feature of the prior comment's attitude and the new feature of the number of subsentences. The results show that the method realizes the identification of the affirmative and negative relationship effectively. It can benefit organizing the large amount of team comments effectively and increase the efficiency of information organization.

1 Introduction

From the 1990s, innovation, replacing efficiency and quality, is regarded as the main resource of competitive advantage for firms [1]. Computer-mediated team innovative activities generate much more redundant information [2][3], thus results in information overload [4][5]. In this case, it cannot satisfy the needs of team innovation by using manual method. Therefore, it is necessary and imperious to study the organization of team information effectively by use of the advanced information technology. However, the process of team innovation is unstructured and complex, which results in the complicated information relationship and difficult information organization. On the other hand, due to the limitation of Chinese text mining technology, complete automation of Chinese information organization is difficult to realize. The essence of information organization is the classification of information. This paper particularly studies the identification of the affirmative and negative relationship between the solution and its comments during team innovation.

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 174–177, 2007.
© Springer-Verlag Berlin Heidelberg 2007

2 Literature Review

Traditional team activities usually depend on the secretary to record and pack up the team information, which costs time and effort. With the development of information technology, more and more scholars pay attention to the research on effective methods and tools to support team information organization during team activities. Some scholars study information classification produced in e-meeting by mining the comparability of information topics [6][7]. Most researches about information classification usually make the concepts involved in information as classes [8]. In team innovation, aiming at a solution, some comments bear a clearly and definitely affirmative or negative relationship with the solution. If the comments are picked up and identified, the team can recognize and master the tendentiousness of the comments, which favors the solution selection and decision. Therefore, the identification of the affirmative and negative relationship between the solution and its comments deserves the research. Moreover, most researches deal with English text. Because of the particularity of Chinese, those researches cannot treat Chinese text effectively. On the basis of text mining and Neural Network (NN), this paper studies the identification of the affirmative and negative relationship between a solution and its comments.

3 Model of Team Information Relationship Identification

In the real process of team innovation, there are various relationships between a solution and its comments. As for the solution selection and decision, the affirmative or negative relationship is what the team is concerned about most. In this paper, the model of information relationship identification is proposed. The model is used to identify the affirmative or negative relationship between the solution and its comment. Firstly, team comment is picked up from the team information base as the target information for identification. Then, the model extracts the features of comments. The next step is to put those features into the NN. The relationship between the comments and solution is identified by the NN. Finally, the results identified are put back into the team information base and used to the team information organization.

4 Application of the Identification Method to Team Innovation

The method above has been applied to an innovative project. During the project, a design solution is selected randomly. 100 comments on the solution are selected randomly as training samples. Another 100 comments on the solution are selected randomly as testing samples.

4.1 Identification Model

Feature Extraction of Team Information

The characteristic vector of comment should be extracted first during the identification. The algorithm of feature extraction has a direct impact on information relationship identification. Vector Space Model (VSM) is mainly used. However, the traditional

method is not appropriate to address the problem mentioned in this paper. We improve the traditional method of feature extraction.

Firstly, words and phrases representing affirmative or negative meaning will be selected as characteristic items.

Secondly, because the comments studied by this paper are short, many dimensions of the characteristic vectors deriving from the traditional extracting method are zero. In order to reduce the dimensions of characteristic vectors greatly and avoid serious losing of features, based on the result of experiments, the feature words are sorted into six levels representing different degrees of affirmation or negation. The six levels separately express the meanings of feature words: extremely affirmation, affirmation, somehow affirmation, extremely negation, negation, and somehow negation.

Thirdly, in order to further improve the precision of the comment attitude, the following two features are particularly added.

(1) *Number of sub-sentences*: This feature describes the number of sub-sentences included by a comment. The sub-sentences are separated by punctuations. This feature can represent the distribution of feature words in a text, which can usually reflect the attitude of the current text.

(2) *Attitude of the prior text*: This feature reflects the implied syntax relationship between neighbored texts in continuous comments. It is mainly used to improve the identification effect on the condition of continuous comments.

BPNN Classifier

A three-layer Back Propagation Neural Network (BPNN) is used to identify the relationship of team information. The input layer has eight neurons. They are the weight of extremely affirmative feature words, the weight of affirmative feature words, the weight of somehow affirmative feature words, the weight of extremely negative feature words, the weight of negative feature words, the weight of somehow negative feature words, the number of sub-sentences, and the attitude of the prior text. Among which, the values of the first seven neurons are {0, 1, 2, 3, 4, 5}. The value of the last one is {-1, 0, 1}. There is one neuron in the output layer. The value of the output node is {-1, 1}. There are four neurons in the hidden layer. *tansig* is the action function of the hidden layer. *purelin* is the action function of the output layer.

4.2 Results of the Identification Model

The experimental results show that the correct ratio of the training set is 100%, and the correct ratio of the testing set is 73.33%. The applied effect can be further described by Recall and Precision. The Precision and the Recall of the affirmative relationship is separately 76.47% and 72.22%. The Precision and the Recall of the negative relationship is separately 75% and 78.95%. The average Recall is 75.74%. The average Precision is 75.59%. The results show that the model can support the identification of the affirmative or negative relationship between the solution and its comments. It increases the efficiency of information organization.

5 Conclusions

In this research, we explore the identification method of the affirmative and negative relationship between the innovation solution and its comments during team innovation. In the identification model, we improve the traditional feature extraction method from three aspects. The results show that the method realizes the identification of the affirmative and negative relationship between the solution and its comment effectively. It can increase the efficiency of information organization.

References

1. Bolwijn, P.T., Kumpe: T. Manufacturing in the 1990s-Productivity, Flexibility and Innovation. Long Range Planning (1990) 44-57
2. Michael Parent, R. Brent Gallupe. Knowledge Creation in Focus Groups: can Group Technologies Help? Information & Management (2000) 47-58
3. Schilling, M.A., Hill, C. Managing the New Product Development Process: Strategic Imperatives. Academy of Management Executive (1998) 67-81
4. Trauth E M, Jessup L M.: Understanding Computer-Mediated Discussions Positives and Interpretive Analyses of Group Support System Use. MIS Quarterly (2000) 43-79
5. Grise M, Gallupe R B.: Information Overload Addressing the Productivity Paradox in Face-to-Face Electronic Meetings. Journal of Management Information Systems (1999-2000) 157-185
6. H. Chen, K.J. Lynch,: Automatic Construction of Networks of Concepts Characterizing Document Databases. IEEE Transactions on Systems, Man, and Cybernetics (1992) 885- 902
7. H. Chen, P. Hsu, R. Orwig L. Hoopes, and J. F. Nunamaker.: Automatic Concept Classification of Text from Electronic Meetings. Communications of the ACM (1994) 56-73
8. Dmitri Roussinov, J. Leon Zhao.: Automatic Discovery of Similarity Relationships through Web Mining. Decision Support Systems (2002) 1-18

Agile Knowledge Supply Chain for Emergency Decision-Making Support

Qingquan Wang and Lili Rong

Institute of Systems Engineering, Dalian University of Technology, 116024 Dalian, China
dlwqq@hotmail.com, llrong@dlut.edu.cn

Abstract. Facing complex and changing emergencies decision makers need to obtain sufficient background knowledge to make effective decisions. We outline the characteristics of requirements for decision-making around knowledge sources, agility and the nature of knowledge products in quick responses to emergencies. We characterize the process of knowledge management in emergencies as quotation, manufacture and supply of a special product -- Emergency Knowledge Product. From the point of view of achieving agility, we draw on the operational mechanism of the Agile Supply Chain (ASC) to construct Agile Knowledge Supply Chain (AKSC), a first in this paper. According to Similarities between ASC and AKSC, here we depict definition and architecture of AKSC. AKSC can explore a new approach to knowledge-based quick responses in the emergency decision-making support.

Keywords: Emergency Decision-Making, Knowledge Product, Knowledge management Agility, Agile Knowledge Supply Chain.

1 Introduction

Nowadays, humans are threatened by various unexpected disasters which include all kinds of terror attacks, epidemics, hurricanes, tsunamis, earthquakes, air crashes, collective food poisoning and industrial accidents. With increasing technology, population and deterioration of the environment, losses of such disasters are increasing exponentially. An outbreak of these disasters causes immeasurable losses in lives and property, like the 911 terror attack, SARS, the bird flu, the Indian Ocean tsunami, the Katrina hurricane, and the Pakistan earthquake. Quick and effective decision- making is crucial in emergency responses [1], and is closely related to background knowledge of circumstances and experiences [2].

In the last decade, Emergency Decision Support Systems (EDSS) have gradually introduced and integrated the technologies of knowledge management. For examples, advanced knowledge models to support environmental emergency management in [3], aggregation of the knowledge of geographically separate experts in [4], the EDSS raised in [5], the expert system developed for knowledge acquisition of chemical incidents [6], the agent-based environmental emergency knowledge system in [7]. The knowledge-based EDSS is still in its primary stage, but the maturation of knowledge-based technologies is facilitating this enterprise. [5].

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 178–185, 2007.
© Springer-Verlag Berlin Heidelberg 2007

The speed and quality of background knowledge are key factors in quick-response emergency decision-making. Mendonca deems that speed of response and quality of knowledge is crucial factors for effective emergency response [4]. And Viviane emphasized time pressure in Collaborative Emergency Responses, described a system which aims at storing and disseminating contextual knowledge to an emergency [8]. Therefore, there are the problems of agility in emergency responses.

Agility of knowledge system is a new research subject in the knowledge management field. Some researchers have started focusing attention on the agility of knowledge system; for example, KWM, in improving the agility of knowledge management [9], agile knowledge sharing [10], agile knowledge workflow [11], agile knowledge engineering methodology, RapidOWL [12], and project SAKE launched by European Community [13]. Agility is much more prominent in the knowledge-based EDSS, especially acquisition of background knowledge, and it is an important indicator for evaluating the aided effectiveness of decision-making background knowledge in relieving time pressure.

The agile manufacturing field gives us a hint that the Agile Supply Chain (ASC) can adjust itself to fit the continuous changes in market demands. A Supply Chain is a loosely related group of companies formed to enable collaboration in achieving mutually agreed upon goals [14], and includes activities and processes to supply products or services to end-users. The concept of ASC was introduced by Stanford University. Researches of ASC mainly concentrate on virtual enterprises, information flow, Supply Chain reconstruction, etc. [15], [16]. From the point of view of the agility of providing background knowledge to the decision-maker, we outline the characteristics of requirements of decision-making and apply emergency knowledge management to the process of quotation, manufacture and supply of a special product. We draw on the operational mechanism of ASC to construct the Architecture of the Agile Knowledge Supply Chain (AKSC) for emergency decision-making support.

In this paper the requirements of emergency decision-making including knowledge resources, characteristics of knowledge requirements and product requirements, are introduced in section 2. In section 3, the nature of ASC, the definition of AKSC and their similarities are discussed. In section 4, we present AKSC architecture. Finally, a summary of this paper is given in section 5.

2 Requirements of Emergency Decision-Making

2.1 Requirements of Emergency Knowledge Sources

"Knowledge is not the same as information. Knowledge is information that has been pared, shaped, interpreted, selected, and transformed"-- E. Feigenbaum.

Processing emergency raw information into knowledge is similar to processing raw materials into products. Knowledge processing has specific sources of information. In knowledge-based EDSS there are three main sources of 'raw material' information: emergency environments, contributing emergent incidents and emergency documentation, shown in figure 1.

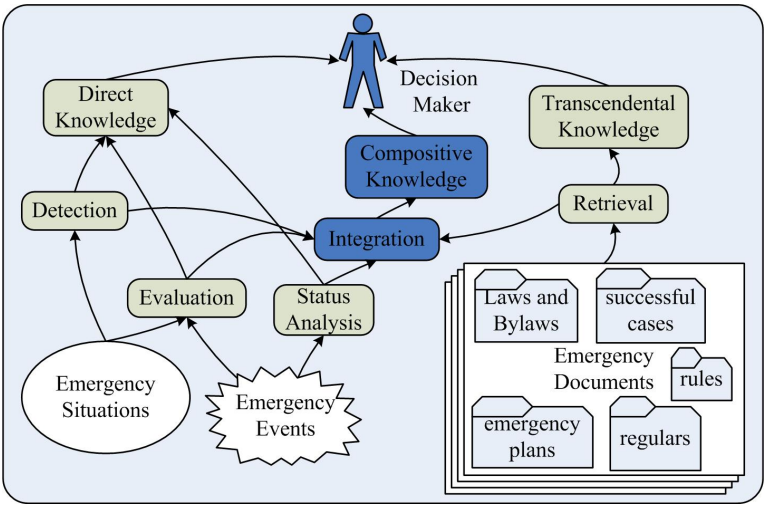


Fig. 1. Sources of Emergency Knowledge

Information materials usually come from emergency environmental monitoring and assessment, analysis and evaluation of emergent incidents, empirical information from emergency documentation, and the integration of them. They constitute respectively, direct knowledge, transcendental knowledge and integrated knowledge.

Direct knowledge and transcendental knowledge have a better application in the existing EDSS, but the most difficult thing is to integrate data, information and knowledge from various information sources [17]. Therefore, in knowledge-based emergency decision support we should not only provide the decision-makers with direct knowledge and transcendental knowledge, but also integrate them effectively to reduce the pressure on decision-makers for improved decision- making.

2.2 Requirements of Agility of Knowledge Support

"Knowledge should be presented understandably, simply, clearly"-- Z.T. Wang [18].

Knowledge in different applications requires different features. Such as generality, complexity and implicity are main features of knowledge in Knowledge Engineering.

In the process of quick emergency response, the knowledge decision-makers receive should be correct, complete, clear and simply described. In addition to these spatial characteristics emergency knowledge management also has characteristic temporal agility. Any loss of agility results in a loss of its original value to decision-makers. Emergency knowledge management agility minimizes time required for processing of emergency knowledge in acquisition, representation and reasoning.

We can regard agility as a feature of emergency knowledge for comparison. Suppose that knowledge provided is transcendental knowledge from emergency documentation. The correctness and completeness of it can be verified and evaluated before the emerging event. The relationships between agility and other characteristics of emergency knowledge can be illustrated roughly through a cube in Figure 2.

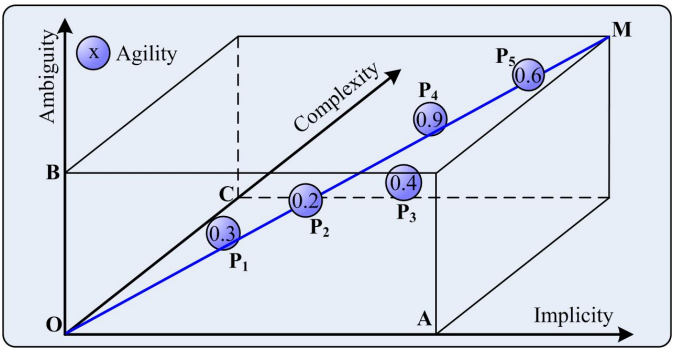


Fig. 2. Cube of Emergency Knowledge Features

With the exception of agility, the closer the points of knowledge are to *A* the more implicit the emergency knowledge is; the closer to *B*, the more ambiguous; to *C*, the more complex. By the same token, the closer they are to point *O*, they are more explicit, precise and simple. Point *O* can be expressed as (*Explicit*, *Precise*, *Simple*), and Point *M* as (*Implicit*, *Ambiguous*, *Complex*). In other words, efficient decision-making support should provide explicit, precise and simply described emergency knowledge. The closer they are to origin *O* in Figure 2 the more efficient they are.

However, in practical applications agility in decision-making support is usually the most critical factor, see figure 2. Among the knowledge points $\{P_1, P_2, P_3, P_4, P_5\}$, point P_4 (*Agility* = 0.9) has the highest probability of being accepted as the most agile.

2.3 Emergency Knowledge Features Required

"Knowledge is a product of the human spirit, mediated by language." -- A. Sigel [19].

In response to a certain emergency, the emergency knowledge available for decision-makers often comes from different departments, various emergency documentation, diverse environments, or experience from similar cases. The acquisition of emergency knowledge involves the process of knowledge decomposition, matching and integration.

Knowledge and products have their respective characteristics. For example, products are usually tangible expendable exclusive, and have value; but knowledge is intangible, unquantifiable value, re-usable and share-able, as it is shown in figure 3. However, in the knowledge management of emergency decision-making, there are notable similarities between product and knowledge, the intersection in Figure 3. They both are demand-driven, subject to supply and demand matching requirements, have their corresponding owners and transfer regularly from one to another, have their respective raw materials, structural characteristics and functions, can be decomposed and integrated, and are agile in generation and supply.

Therefore, in quick response emergency decision-making, the knowledge decision-makers receive and use is a kind of product, Emergency Knowledge Product (EKP). EKP is a product processed from dispersed emergency information, and it can be said that the process of EKP is a special approach to knowledge representation.

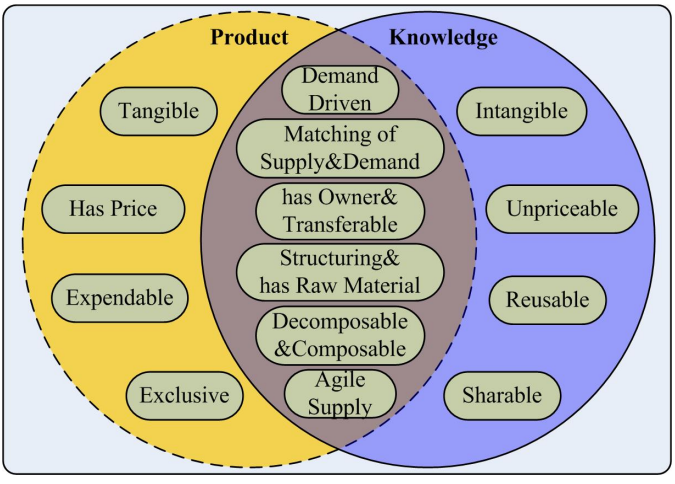


Fig. 3. Common Features of Knowledge and Products

Emergency decision-makers use EKP as reference to make appropriate decisions. Each decision is incorporated as an information source into subsequent decisions. Emergency decision-makers are either end-consumers or providers of raw materials.

3 Agile Knowledge Supply Chain

In the three knowledge requirements above we can see that emergency knowledge management is very similar to the ASC. The main characteristic of ASC which is distinct from the general Supply Chain is rapid reconstruction and readjustment with the formation and dissolution of a dynamic alliance. Agility of ASC is the ability of quick responses to survive, develop and enlarge the competitive advantages in an uncertain and rapid changing environment. Improved Supply Chain performance implies that a Supply Chain is capable of quickly responding to variations in customer demands with effective cost reduction, [20] such as lead time [21].

In emergency responses, alliances among various relevant organizations and departments (Emergency Alliances) are formed according to the different types of emergent incidents. Members of Emergency Alliances concentrate on their respective responsibilities, and establish organizational relationships based on integrative commands and cooperation. Emergency Alliances are special virtual enterprises and constitute Supply Chains based on knowledge management.

We construct Agile Knowledge Supply Chains (AKSC) to support emergency decision-making. They are directed toward the realization agility in decision-making support to reinforce emergency knowledge management. AKSC is a Supply Chain which takes knowledge as a product, establishes an Emergency Alliance, gives priority to its Command Center which directs Functional Departments, and has notable agility in emergency decision-making support. See the similarities between ASC and AKSC elements of reconstruction, information sharing and flow management in Table 1.

Table 1. Similarity between ASC and AKSC

Item	ASC	AKSC
Reconstruction	Rapid reconstruction and readjustment according to market demand	Rapid reconstruction and readjustment according to emergent incident
Information Sharing	Sharing information among enterprises in the Supply Chain to avoid a bullwhip effect	Sharing information among the emergency departments to avoid delays in knowledge acquisition
Organizational form	Virtual enterprise, a dynamic alliance based on the matching supply and demand of products	Emergency Alliance, a dynamic alliance based on the matching supply and demand of knowledge
Drive Mode	Customer order-driven	Decision-maker requirement-driven
Global Object	Reduce the business risks arising from uncertainty and variability	Achieve ultimate effectiveness of emergency decision implementation
Output	Goods have commercial value	All sorts of knowledge used in emergency decision-making
Flow Management	Integrated flows of materials, information, capital	Integrated flows of knowledge demands, supplies and production
Agility	Supply products to meet customer demands in the shortest possible time	Supply knowledge to meet decision-making demands in the shortest possible time
Network	Promote the circulation of product, capital and information	Facilitate knowledge re-use and sharing

4 Architecture of AKSC

Generally, Emergency Alliances consist of Command Centers and functional departments. Emergency Alliances are usually established by Command Centers which bring together various emergency functional departments around the requirements of emergent incidents. So an Emergency Alliance has only one leader, its Command Center, which is usually launched and directed by a specific level of government. The relationships among the members of Emergency Alliances are, under unified leadership, directive, rather than the collaborative equality and mutual trust characteristic of traditional Supply Chains. Because of changing targets, requirements and tasks of emergent incidents the structure of an Emergency Alliance is dynamic. Members of Emergency Alliances, and the knowledge transferred among them, constitute the dynamic Supply Chain of emergency knowledge management.

Core technologies of the AKSC are knowledge decomposition, matching and integration, i.e. the reorganization of knowledge. It extends the knowledge management mode for emergency management from the traditional knowledge database to flow-based knowledge, from rule driven to product driven. AKSC also has excellent extendibility to accommodate changes in incidents and organizational structure. AKSC can integrate technologies of ontology, multi-agent, network, knowledge representation, reasoning, etc. Figure 4 shows the architecture of AKSC which mainly concerns the Command Center in the Emergency Alliance.

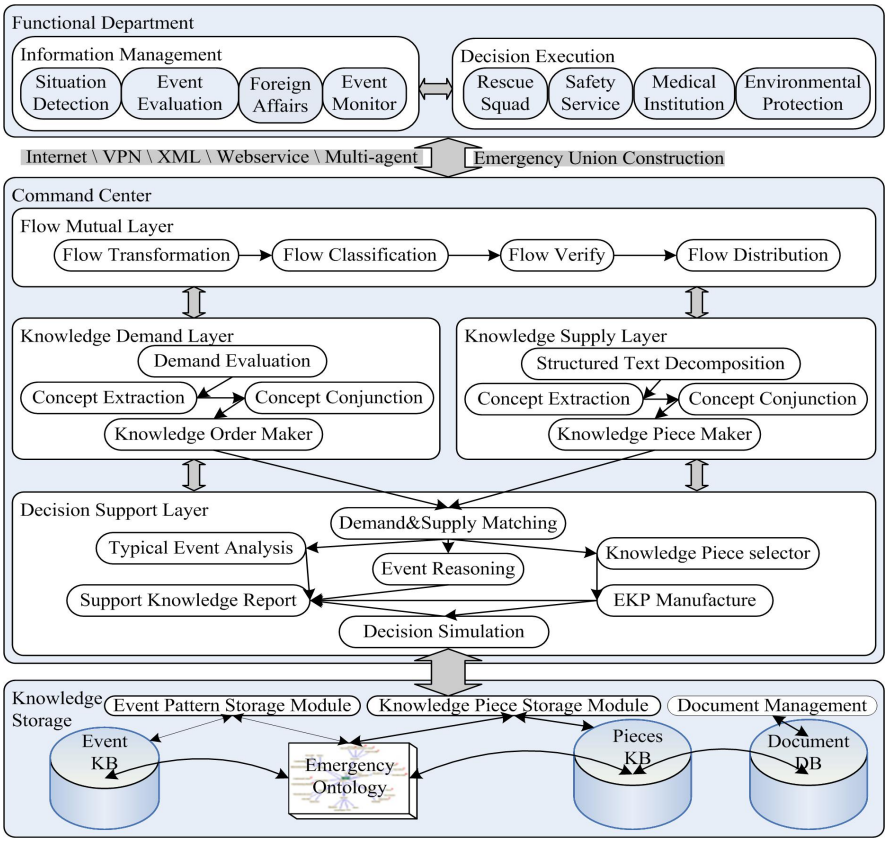


Fig. 4. Architecture of AKSC

5 Conclusion

This paper has analyzed three knowledge requirements for emergency decision-making: sources, agility and knowledge product features. We compared the knowledge management of emergency decision-making support with the Supply Chain management, a first in this paper, and regarded emergency knowledge as practical products. Using the operating mechanism of ASC for reference, we constructed the AKSC to support timely and effective decision-making through the quick acquisition of required knowledge. We briefly outline the definition, content and architecture of AKSC in this paper.

We believe the AKSC presented in this paper breaks new ground in knowledge management of the emergency decision-making support. However, some technical problems remain to be solved; such as methods of EKP design, approaches to demand description. We intend to dedicate ourselves to researching them.

Acknowledgement

This research is supported by the Natural Science Foundation of China (Grant No, 70571011, 70431001). The authors would like to express appreciation to the team of project EDKS for the sincere support.

References

1. Rong, L.L.: Reorganizing the Knowledge in Government Documents for Quick Response. KSS'2005. Laxenburg, Austria (2005)
2. Jia. X.N., Rong. L.L.: Classification Based Management Method for Government Documents in Emergency Response. KSS'2006. Beijing, China. (2006)
3. Hernandez, J.Z., Serrano, J.M.: Knowledge-based Models for Emergency Management Systems. Expert Systems with Applications, 20 (2001) 173-186
4. Mendonca, D., Rush, R., Wallace, W.A.: Timely Knowledge Elicitation from Geographically Separate, Mobile Experts during Emergency Response. Safety Science, 35 (2000) 193-208
5. Cortés, U., Sànchez-Marrè, M., Ceccaroni, L., R-Roda, I., Poch, M.: Artificial Intelligence and Environmental Decision Support Systems, Applied Intelligence, 13 (2000) 77-91
6. Yeh, T.H., Lo, J.G.: A Case Study of Knowledge-Based Expert System for Emergency Response of Chemical Spill and Fire. Environ. Inform. Arc. 2 (2004), 743-755
7. Liu, K.F.R.: Agent-based resource discovery architecture for environmental emergency management. Expert Systems with Applications, 27 (2004) 77–95
8. Viviane, B.D., Marcos, R.S.B., Jose, O.G., Jose, H.C.: Knowledge management support for collaborative emergency response. 9th CSCWD, Coventry, UK, (2005) 1188-1193
9. Lee, H.B., Kim, J.W., Park, S.J.: KWM: Knowledge-based Workflow Model for Agile Organization. Journal of Intelligent Information Systems, 13 (1999) 261-278
10. Melnik, G., Maurer, F.: Direct Verbal Communication as a Catalyst of Agile Knowledge Sharing. ADC'2004. Los Alamitos, CA. (2004)
11. Holz, H., Maus, H., Bernardi, A., Rostanin, O.: From Lightweight, Proactive Information Delivery to Business Process-Oriented Knowledge Management, Journal of Universal Knowledge Management 2 (2005) 101-127
12. Auer, S.: RapidOWL - an Agile Knowledge Engineering Methodology. STICA 2006, Manchester, UK. (2006)
13. Stojanovic, N., Mentza, G., Apostolou, D.: Semantic-enabled Agile Knowledge-based E-government. Http://www.imu.iccs.gr/sweg/papers/
14. Christopher, M.: The Agile Supply Chain: Competing in Volatile Markets. Industrial Marketing Management, 29 (2000) 37–44
15. Zhang, S.S., Gao, G.J.: Dynamic Allience and Agile Supply Chain. Computer Integration Manufacture System, 5 (1999) 1-5
16. Lou, P., Zhou, Z.D., Chen, Y.P., Ai, W.: Study on multi-agent-based Agile Supply Chain management. Int. J. Adv. Manuf. Tech. 23 (2004) 197-203
17. Stephanopoulos, G., Han, C.: Intelligent systems in process engineering: A review. Computers Chemical Engineering, 20 (1996) 743–791
18. Wang, Z.T.: Knowledge System Engineering. Beijing: Science Publishing Company, 2004
19. Sigel, A.: Topic Maps in Knowledge Organization. http://index.bonn.iz-soz.de/ ~sigel/
20. Agarwal, A., Shankar, R., Tiwari, M. K.: Modeling the Metrics of Lean, Agile and Leagile Supply Chain: An ANP-based Approach. Eur. J. Oper. Res., 173 (2006) 211–225
21. Mason-Jones, R., Towill, D.R.: Total Cycle Time Compression and the Agile Supply Chain. European Journal of Operational Research, 159 (2004) 379-392

Interactive Fuzzy Goal Programming Approach for Optimization of Extended Hub-and-Spoke Regional Port Transportation Networks

Chuanxu Wang[1] and Liangkui Jiang[2]

[1] School of Economy and Management,
Shanghai Maritime University, Shanghai, 200135, China
cxwang@shmtu.edu.cn
[2] Department of Basic Science, Shanghai Maritime University, Shanghai, 200135, China
lkjiang@dbc.shmtu.edu.cn

Abstract. Based on interactive fuzzy goal programming, a model is introduced for extended hub-and-spoke regional port transportation network optimization problem where one of the objective functions is considered as a non-linear function. The proposed model considers the imprecise nature of the input data and assumes that each objective function has a fuzzy goal, and aims at jointly minimizing the total costs including the sailing cost and handling cost as well as the total transit times consisting of sailing time and handling time. Meanwhile it optimizes the following factors: transportation quantities via a hub port from an original port to a destination port, transportation quantities directly from an original port to a destination port. The solution procedure to the proposed model is then presented. At last, a numerical example is given to demonstrate the effectiveness of the proposed model and evaluate the performance of the solution procedure.

Keywords: Fuzzy goal programming, Hub-and-spoke, Optimization, Port transportation network.

1 Introduction

In regional port transportation networks, the sizes and types of ships calling at every port are different because individual port's natural condition and capacity are different. The route selection is important in regional port transportation network optimization. The port condition and economy of scale in ship transportation should be considered to select the routes from original ports to destination ports. The ports are connected by direct transports or via a transshipment port. This has led to the formation of hub-and-spoke networks in regional port transportation industry. Cargoes from original ports are usually consolidated at hub port and shipped to different destination ports. Hub-and-spoke transportation can be classified into two types: pure and extended [1]. The pure hub-and-spoke transportation network is characterized by transshipment in which direct transport between ports is excluded and all cargoes have to be transported via a hub port. The extended hub-and–spoke transportation network consists of direct transportation between ports and transshipment via the hub port.

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 186–193, 2007.
© Springer-Verlag Berlin Heidelberg 2007

In this paper, based on extended hub-and-spoke port transportation system, a decision model is developed to determine the following factors: shipment volume via a hub port from an original port to a destination port, shipment volume directly from an original port to a destination port.

The decision problem of hub-and-spoke systems has received many attentions in academic literature. Some researchers solved it as the location –allocation problem that determines locations of hubs and assigns shipments to each route[2][3][4]. Another researchers examined it as the pure allocation problem that finds the optimal assignment of shipments under predetermined hub locations[5][6][7]. But the models and methods in these studies are applied in air or truck transportation networks and have been less focused on linking hub-and-spoke system to ports. Most of them only consider transportation (or travel) cost and don't include the handling costs occurring at nodes. Furthermore, these studies only consider one objective function. This paper introduces a model to examine the pure allocation problem in extended hub-and-spoke regional port transportation network. The model aims at jointly minimizing the total costs consisting of sailing cost and handling cost occurring at the ports as well as total time consisting of sailing time and handling time occurring at the ports, in which the time objective function is a non-linear function. In addition, the existing decision problem of hub-and-spoke system is a deterministic mathematical programming problem. However, in practice, the input data or decision parameters, such as transportation capacity, cost, time and objective function are often imprecise because some information is uncertain. Therefore, the values of these parameters are rarely constant. To deal with ambiguous parameters in the above decision problem, this paper uses an interactive fuzzy goal programming model to formulate the problem of extended hub-and-spoke regional port transportation network optimization, and propose a solution procedure for this model.

2 The Decision Problem of Extended Hub-and-Spoke Port Transportation Network

2.1 Notations

M **,** the number of ports excluding the hub port 0;

D_{ij}, transportation demand quantity from port i to port j;

C_{ij}, the transportation cost per unit cargo from port i to port j;

S_{ij}, the transportation capacity in the route from port i to port j;

T_{ij}^{k}, the sailing time from port i to port j;

C_i, the handling cost per unit cargo occurred at port i;

T_i the handling time per unit cargo occurred at port i;

x_{ij}**,** transportation quantity from port i to port j ($i \neq j$); (Decision variables)

2.2 Mathematical Model

The decision problem of extended hub-and-spoke port transportation system can be formulated as follows:

$$\text{(P1)}\ \min\ Z_1 = \sum_{i=0}^{M} \sum_{\substack{j=0 \\ i\neq j}}^{M} C_{ij} x_{ij} + \sum_{i=0}^{M} C_i \sum_{\substack{j=0 \\ j\neq i}}^{M} x_{ij} + \sum_{i=0}^{M} C_i \sum_{\substack{j=0 \\ j\neq i}}^{M} x_{ji} \tag{1}$$

$$\min\ Z_2 = \sum_{i=0}^{M} \sum_{\substack{j=0 \\ i\neq j}}^{M} T_{ij} x_{ij} + \sum_{i=0}^{M} \sum_{\substack{j=0 \\ i\neq j}}^{M} (T_i \sum_{\substack{j=0 \\ j\neq i}}^{M} x_{ij} + T_j \sum_{\substack{i=0 \\ i\neq j}}^{M} x_{ij}) x_{ij} \tag{2}$$

$$\text{s.t.}\ x_{i0} + \sum_{\substack{j=1 \\ j\neq i}}^{M} x_{ij} = \sum_{\substack{j=1 \\ j\neq i}}^{M} D_{ij},\ i = 1,2,...M, i \neq 0 \tag{3}$$

$$x_{0i} + \sum_{\substack{j=1 \\ j\neq i}}^{M} x_{ji} = \sum_{\substack{j=1 \\ j\neq i}}^{M} D_{ji}\ \ i = 1,2,...M, i \neq 0 \tag{4}$$

$$x_{ij} \le S_{ij}\ \ i = 0,1,...M,\ j = 0,1,...M, i \neq j \tag{5}$$

$$x_{ij} \ge 0\quad i = 0,1,...M,\ j = 0,1,...M, i \neq j \tag{6}$$

In real world, the input data or parameters of (P1) problem, such as transportation capacity and objective function are often imprecise because some information is unobtainable. Conventional mathematical programming can't capture such vagueness in the critical information. In this case, a fuzzy programming approach is commonly used to treat the information in a fuzzy environment. However, some researchers have presented the shortcomings of using fuzzy programming in solving some multi-objective decision problems. Abd El-Wahed (2001) proved that using fuzzy programming in solving such multi-objective transportation problem changes the standard form of the transportation problem [8]. In addition, Li and Lai (2000) proved that using the min-operator does not guarantee an efficient solution[9]. In this paper, we introduce an interactive fuzzy goal programming model for regional port transportation networks optimization, which is the combination of interactive programming, fuzzy programming and goal programming, and leverages the advantages of the three approaches as well as reduces some or all of the shortcomings of each individual approach [10].

3 Interactive Fuzzy Goal Programming Approach

3.1 IFGP Model

In our model, we capture the ambiguity about the fuzzy information related to the total transportation cost, total transportation time and transportation capacities by transforming the (P1) model into the following (P2) model.

(P2) $$\sum_{i=0}^{M}\sum_{\substack{j=0\\ i\neq j}}^{M} C_{ij}x_{ij} + \sum_{i=0}^{M} C_i \sum_{\substack{j=0\\ j\neq i}}^{M} x_{ij} + \sum_{i=0}^{M} C_i \sum_{\substack{j=0\\ j\neq i}}^{M} x_{ji} \underset{\sim}{<} \tilde{Z}_1 \tag{7}$$

$$\sum_{i=0}^{M}\sum_{\substack{j=0\\ i\neq j}}^{M} T_{ij}x_{ij} + \sum_{i=0}^{M}\sum_{\substack{j=0\\ i\neq j}}^{M} (T_i \sum_{\substack{j=0\\ j\neq i}}^{M} x_{ij} + T_j \sum_{\substack{i=0\\ i\neq j}}^{M} x_{ij}) x_{ij} \underset{\sim}{<} \tilde{Z}_2 \tag{8}$$

$$\text{s.t. } x_{ij} \underset{\sim}{<} \tilde{S}_{ij} \quad i = 0,1,...M,\ j = 0,1,...M, i \neq j, \tag{9}$$

as well as constraints given in (3), (4) and (6)

where the symbol " $\underset{\sim}{<}$ " indicates "essentially smaller than or equal to " and allows one reach some aspiration level, $\tilde{Z}_1, \tilde{Z}_2,$ and $\tilde{S}_{ij}$ denote fuzzy values.

In this paper, a linear membership function defined in Bellman and Zadeh(1970) has been considered for all fuzzy parameters in (P2) problem.

To formulate model (P2) as a goal programming model, we introducing the following positive and negative deviation variables:

$$Z_k(x) - d_k^+ + d_k^- = G_k, \qquad k = 1,2 \tag{10}$$

where G_i is the aspiration level of the objective function i.

By using the given membership functions and introducing an auxiliary variable L the fuzzy programming model (P2) can be formulated as the following equivalent programming model (P3) (Zimmermann, 1978):

(P3) Max L

$$\text{s.t. } L(Z_k^{\max} - Z_k^{\min}) + Z_k(x) \leq Z_k^{\max} \qquad k = 1,2 \tag{11}$$

$$L(S_{ij}^U - S_{ij}^L) + x_{ij} \leq S_{ij}^U \quad i = 0,1,...M,\ j = 0,1,...M, i \neq j \tag{12}$$

$$Z_k(x) - d_k^+ + d_k^- = G_k \qquad k = 1,2 \tag{13}$$

$$0 \leq L \leq 1, \tag{14}$$

$$d_k^+, d_k^- \geq 0, \qquad k = 1,2 \tag{15}$$

as well as constraints given in (3), (4) and (6).

3.2 The Solution Procedure

In order to obtain the solution to Model (P3), the following procedure can be applied.

Step1: Solve the model (P1) as a single objective problem to obtain the initial solutions for each objective function, i.e. $X^1 = \{x_{ij}^1\}$ and $X^2 = \{x_{ij}^2\}$, $i = 0,1,..M,\ j = 0,1,...,M,.\ i \neq j$. If $X^1 = X^2$, select one of them as an optimal solution and go to Step 6. Otherwise, go to Step 2.

Step 2: Determine the best lower bound ($Z_k^{\min}$) and the worst upper bound ($Z_k^{\max}$) for each objective function: $Z_1^{\min}=Z_1(X^1)$, $Z_1^{\max}=Z_1(X^2)$, $Z_2^{\min}=Z_2(X^2)$, $Z_2^{\max}=Z_2(X^1)$, $i=0,1,..M$, $j=0,1,...,M$, $i\neq j$.

Step 3: Define the membership functions of each objective function.

Step 4: Solve problem (P4) as a goal programming and obtain the solution to it. Compare the upper bound of each objective function with the new value of the objective function. If the new value of each objective function is equal to the upper bound, go to Step 6. Otherwise, go to Step 5.

Step 5: Update the upper bound of each objective function. If the new value of the objective function is lower than the upper bound, consider this as a new upper bound. Otherwise, keep the old one as is. Go to Step 3.

Step 6: Stop.

4 Numerical Example

We consider the following numerical example to demonstrate the application of IFGP. Assume that there are four ports and one hub in regional port transportation network. There are two types of ships employed in the route between two ports. The transportation capacity of ship for type 1 is 5000 Ton, whereas that for type 2 is 8000 ton. The handling cost per ton occurring at Port 1, Port 2, Port 3, Port 4 and Port 0 are 0.40,0.55,0.60,0.65 and 0.40 Yuan, respectively. The handling time per ton occurring at Port 1, Port 2, Port 3, Port 4 and Port 0 are 0.5,1.0,1.0,1.5 and 0.5 hour, respectively. The other data of the problem is given in Table1, Table2 and Table3.

Table 1. Transportation demand between ports (ton)

Original Ports	Destination ports				
	Port1	Port 2	Port 3	Port 4	Hub 0
Port 1	----	1790	1960	2680	750
Port 2	2500	----	3030	6830	1000
Port 3	2320	2670	----	9490	2100
Port 4	890	2330	3690	----	650
Hub 0	450	1500	2700	560	----

Table 2. Transportation costs (Yuan per Ton) and sailing time(Hours) in different routes

Original ports	Destination ports									
	Port 1		Port 2		Port 3		Port 4		Hub 0	
	cost	time	cost	time	cost	time	cost	time	cost	time
Port 1	----	----	3.0	10.0	4.0	8.0	4.0	8.0	1.5	5.0
Port 2	2.5	11.0	----	----	2.5	11.0	4.5	8.0	1.1	4.0
Port 3	3.0	10.0	2.8	9.0	----	----	4.0	8.0	1.5	5.0
Port 4	1.6	5.5	3.0	10.0	4.0	8.0	----	----	2.0	11.0
Hub 0	1.5	5.0	1.1	4.0	1.5	5.0	1.8	11.0	----	----

Table 3. Transportation capacity provided by carriers for different routes (Ton)

Original ports	Destination ports									
	Port 1		Port 2		Port 3		Port 4		Hub 0	
	Lower bound	Upper bound	Lower bound	Upper bound	Lower bound	Upper bound	Lower bound	Upper bound	Lower bound	Upper bound
Port 1	----	----	1850	2000	2350	2500	2850	3000	1570	1600
Port 2	2880	3000	----	----	2950	3000	6850	7000	1460	1500
Port 3	2450	2500	2850	3000	----	----	9350	9500	2350	2500
Port 4	850	900	2450	2500	3850	4000	----	----	1150	1200
Hub 0	560	600	1580	1600	2860	3000	560	600	----	----

4.1 Solution

The initial model (P1) of the numerical example can be obtained as

$$\text{Min } Z_1 = 2.3x_{01} + 2.05x_{02} + 2.5x_{03} + 2.85x_{04} + 3.95x_{12} + 5.0x_{13} + 5.05x_{14} + 2.3x_{10} + 3.45x_{21} + 3.65x_{23}$$
$$+ 5.70x_{24} + 2.05x_{20} + 4.0x_{31} + 3.95x_{32} + 5.25x_{34} + 2.5x_{30} + 2.65x_{41} + 4.2x_{42} + 5.25x_{43} + 3.05x_{40}$$
$$Z_2 = 5.0x_{01} + 4.0x_{02} + 5.0x_{03} + 11.0x_{04} + 10.0x_{12} + 8.0x_{13} + 8.0x_{14} + 5.0x_{10} + 11.0x_{21} + 11.0x_{23}$$
$$+ 8.0x_{24} + 4.0x_{20} + 10.0x_{31} + 9.0x_{32} + 8.0x_{34} + 5.0x_{30} + 5.5x_{41} + 10.0x_{42} + 8.0x_{43} + 11.0x_{40}$$
$$+ x_{01}^2 + x_{01}x_{02} + x_{01}x_{03} + x_{01}x_{04} + x_{01}x_{21} + x_{01}x_{31} + x_{01}x_{41} + 1.5x_{02}^2$$
$$+ x_{02}x_{03} + x_{02}x_{04} + 2.0x_{02}x_{12} + 2.0x_{02}x_{32} + 2.0x_{02}x_{42} + 1.5x_{03}^2 + 1.0x_{03}x_{04} + 2.0x_{03}x_{13}$$
$$+ 2.0x_{03}x_{23} + 2.0x_{03}x_{43} + 2.0x_{04}^2 + 3.0x_{04}x_{14} + 3.0x_{04}x_{24} + 3.0x_{04}x_{34} + x_{10}^2 + x_{10}x_{12}$$
$$+ x_{10}x_{13} + x_{10}x_{14} + x_{10}x_{20} + x_{10}x_{30} + x_{10}x_{40} + 1.5x_{12}^2 + x_{12}x_{13} + x_{12}x_{14} + 2.0x_{12}x_{32}$$
$$+ 2.0x_{12}x_{42} + 1.5x_{13}^2 + x_{13}x_{14} + 2.0x_{13}x_{23} + 2.0x_{13}x_{43} + 2.0x_{14}^2 + 3.0x_{14}x_{24} + 3.0x_{14}x_{34}$$
$$1.5x_{20}^2 + 2.0x_{20}x_{21} + 2.0x_{20}x_{23} + 2.0x_{20}x_{24} + x_{20}x_{30} + x_{20}x_{40} + 1.5x_{21}^2 + 2.0x_{21}x_{23}$$
$$+ 2.0x_{21}x_{24} + x_{21}x_{31} + x_{21}x_{41} + 2.0x_{23}^2 + 2.0x_{23}x_{24} + 2.0x_{23}x_{43} + 2.5x_{24}^2 + 3.0x_{24}x_{34}$$
$$+ 2.0x_{30}x_{34} + 2.0x_{31}x_{34} + 2.0x_{32}x_{34} + 2.5x_{34}^2 + 2.0x_{30}x_{32} + 2.0x_{31}x_{32} + 2.0x_{32}^2 + 2.0x_{32}x_{42}$$
$$+ 2.0x_{30}x_{31} + 1.5x_{31}^2 + 1.5x_{30}^2 + 1.0x_{31}x_{41} + 1.0x_{30}x_{40} + 2.0x_{40}^2 + 3.0x_{40}x_{41} + 3.0x_{40}x_{42} + 3.0x_{40}x_{43}$$
$$+ 2.0x_{41}^2 + 3.0x_{41}x_{42} + 3.0x_{41}x_{43} + 2.5x_{42}^2 + 3.0x_{42}x_{43} + 2.5x_{43}^2$$

$$\text{s.t. } x_{10} + x_{12} + x_{13} + x_{14} = 7180,\ x_{20} + x_{21} + x_{23} + x_{24} = 13360$$
$$x_{30} + x_{31} + x_{32} + x_{34} = 16580,\ x_{40} + x_{41} + x_{42} + x_{43} = 7560$$
$$x_{01} + x_{21} + x_{31} + x_{41} = 6160,\ x_{02} + x_{12} + x_{32} + x_{42} = 8290$$
$$x_{03} + x_{13} + x_{23} + x_{43} = 11380,\ x_{04} + x_{14} + x_{24} + x_{34} = 19560$$
$$x_{10} \le 1600,\ x_{12} \le 2000,\ x_{13} \le 2500,\ x_{14} \le 3000,\ x_{20} \le 1500,\ x_{21} \le 3000$$
$$x_{23} \le 3000, x_{24} \le 7000,\ x_{30} \le 2500,\ x_{31} \le 2500, x_{32} \le 3000,\ x_{34} \le 9500$$
$$x_{40} \le 1200, x_{41} \le 900,\ x_{42} \le 2500,\ x_{43} \le 4000, x_{01} \le 600,\ x_{02} \le 1600$$
$$x_{03} \le 3000,\ x_{04} \le 600,\quad x_{ij} \ge 0 \quad i = 0,1,...4,\ j = 0,1,...4, i \ne j$$

By using the above-mentioned solution procedure, the set of solutions (Z^1, Z^2) can be obtained after 17 iterations. The values of objective function are converged at (212509.0, 1366308000), the corresponding optimal solution is as follows:

$$x_{01}=0, x_{02}=790.1, x_{03}=2200.0, x_{04}=507.0, x_{12}=2000.0, x_{13}=2180.1, x_{14}=3000.0, x_{10}=0.0,$$
$$x_{21}=3000.0, x_{23}=3000.0, x_{24}=6553.1, x_{20}=807.0, x_{31}=2260.0, x_{32}=3000.0, x_{34}=9500.0, x_{30}=1820.0$$
$$x_{41}=900.0, x_{42}=2500.0, x_{43}=4000.0, x_{40}=160.0, d_1^+=234.0, d_1^-=0.0, d_2^+=2614425.0, d_2^-=0.0$$

4.2 Performance Evaluation

To evaluate the performance of the proposed model, the solution to the illustrative example by using different methods is considered. The fuzzy programming approach gives the following results: $Z^1=212729$, $Z^2=1368880000$. The interactive fuzzy goal programming approach provides the following results: $Z^1=212505$, $Z^2=1366640000$. To determine the degree of closeness of the IFGP model results to the ideal solution, the family of distance functions presented in [10][11] is considered.

$$D_p(\lambda, K)=[\sum_{k=1}^{K}\lambda_k^p(1-d_k)^p]^{1/p}, \tag{16}$$

where d_k represents the degree of closeness of the solution to the ideal optimal solution with respect to the k th objective function, $d_k=$ the ideal optimal value of Z^k / the solution of Z^k. λ_k is the weight of the k th objective and $\sum_{k=1}^{K}\lambda_k=1$. The power p represents a distance parameter $1\le p\le\infty$.

Thus, we can state the approach is better than others if $Min\ D_p(\lambda, K)$ is achieved by its solution with respect to some p. Assuming $\lambda_1=\lambda_2=0.5$, $p=1,2$ and ∞, the results of the different approaches are given in Table 4.

Table 4. Comparison of results by different approaches

	Fuzzy programming	Interactive fuzzy programming	Interactive fuzzy goal programming	Optimal solution
(Z^1,Z^2)	(212729, 1368880000)	(212505, 1366640000)	(212509.0, 1366308000)	(212274.5, 1363694000)
D_1	0.0030	0.0016	0.0015	--
D_2	0.0022	0.0012	0.0011	--
D_∞	0.0019	0.0011	0.0010	--

It is clear from Table 2 that the solution to the proposed approach is better than the solution by the other approaches for all the distance functions.

5 Conclusions

This paper proposes an interactive fuzzy goal programming model of the extended hub-and-spoke regional port transportation network optimization problem, where transportation capacity and objective function are considered as o imprecise/fuzzy, and aims at jointly minimizing the total costs including the sailing cost between ports and handling cost occurring at all ports as well as the total transit times consisting of sailing time between ports and handling time in all ports. To obtain the solution of the above-mentioned problem, the solution procedure for the model is illustrated by updating both the membership values and the aspiration levels. At last, a numerical example is given to demonstrate the effectiveness of the proposed model and evaluate the performance of the solution procedure by comparing its results with those of the fuzzy programming approach and interactive fuzzy programming approach.

Acknowledgments. The work in this paper was supported by National Natural Science Foundation of China (No.70573068), Shanghai Shuguang Program (No.06SG48) and Shanghai Pujiang Program.

References

1. Zapfel, G., Wasner, M., Planning and optimization of hub-and-spoke transportation networks of cooperative third party logistics providers, *International Journal of Production Economics,* (2002), 78 :207-220.
2. Aykin,T.,Networking policies for hub-and-spoke systems with applications to the air transportation system, *Transportation Science*, (1995),26:201-221.
3. O'Kelly, M.E., A quadratic integer program for location of interacting hub facilities, *European Journal of Operational Research*, 1987, 32: 393-404.
4. Pirkul,H.,Schilling,D.A.,An efficient procedure for designing single allocation hub and spoke systems, *Management Science*, (1998), 44: S235-242.
5. Abdinnourhelm,S.,Venkataramanan,M.A., Solution approaches to hub location problems, *Annals of Operations Research*, (1998), 78:31-50.
6. Bertazzi,L.,Speranza,M.G.,Ukovich,W.,Minimization of logistics costs with given frequencies, *Transportation Research* B, (1997), 31: 327-340.
7. Liu, J.,Li,C.L.,Chan,C.Y.,Mixed truck delivery systems with both hub-and-spoke and direct shipment. *Transportation Research* E, (2003), 39:325-339.
8. Abd El-Wahed WF, A multi-objective transportation problem under fuzziness. *Fuzzy Sets and Systems*, (2001), 117:27-33
9. Li,L.,Lai KK., A fuzzy approach to the multi-objective transportation problem, *Computers & Operational Research*, (2000), 27: 43-57
10. Abd El-Wahed WF, Lee, S.M., Interactive fuzzy goal programming for multi-objective transportation problems, *Omega*, (2006), 34: 158-166
11. Steuer, R., Multiple criteria optimization: theory, computation, and application. New York: Wiley, (1986).

A Pseudo-Boolean Optimization for Multiple Criteria Decision Making in Complex Systems

Bahram Alidaee[1], Haibo Wang[2,⋆], and Yaquan Xu[3]

[1] University of Mississippi, University, MS 38677, USA
balidaee@bus.olemiss.edu
[2] Texas A& M International University, Laredo, TX 78041, USA
hwang@tamiu.edu
[3] Viginia State University, Petersburg, VA 23806, USA
yxu@vsu.edu

Abstract. In complex system problems, a Decision Maker (DM) is often faced with choosing a subset of alternatives from a bigger set. This process is known as multiple criteria decision making (MCDM). Examples of MCDM include decision making in human resource management, water resource management, environmental management and site selection, energy policy issues, portfolio selection, transportation and routing selection, student admission. In general, there are several criteria that need to be satisfied; however, they are usually at least partly conflicting. In this study, we propose a pseudo-boolean approach for multiple criteria decision making on complex system problems. The computational results illustrate both robustness and attractiveness of this solution approach.

Keywords: Pseudo-Boolean Optimization, Multiple Criteria Decision Making.

1 Introduction

Complex system problems are often hard to model and some of them can be represented by nonlinear objective functions with interconnectedness among multiple dimensional spaces. Conventional mathematical programming approaches are sometimes unable to solve it directly. It is conventional wisdom to apply the combination of artificial intelligent, evolutionary computation, system dynamics and heuristic such as genetic algorithm and ant colony to solve these problems [1,2,3,4,5,6]. Other approaches using chaos theory and morphological analysis are reported in the literature [7,8]. In this study, we apply multiple criteria decision making model to complex systems which can be solved by a pseudo-boolean optimization approach.

pseudo-boolean formulations are widely known for their ability to represent a rich variety of important discrete problems, which are often NP-hard. Hammer and Rudeanu [9] introduced pseudo-boolean optimization in their early work and presented a dynamic programming procedure for solving certain problems.

⋆ Corresponding author.

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 194–201, 2007.
© Springer-Verlag Berlin Heidelberg 2007

In their paper, they gave a definition of a *first-order derivative* and indicated its use in solving discrete pseudo-boolean optimization problems. Detail surveys of pseudo-boolean optimization are given by Boris and Hammer [10] and Crama and Hammer [11].

In this paper we extend this notion of first order derivatives by defining *higher-order derivatives* for discrete pseudo-boolean optimization. Moreover, in the context of changing r elements of x at a time (the so-called *r-flip* moves for $r=$ 2 and 3), we present closed-form formulas that allow '*efficient*' implementation of such compound moves. Then, for the important special cases of quadratic and cubic optimization, we define a general *r-flip* move that allows efficient implementation of multi-exchange neighborhood search process for solving such problems. Finally, we illustrate the use of such moves by applying variants of simple search processes based on *r-flip* moves (with $r = 1$ and 2) to a test bed of complex system problems. The paper then concludes with summary and a look ahead to future research.

2 Multiple Criteria Decision Making

In complex system problems, a Decision Maker (DM) is often faced with choosing a subset of alternatives from a bigger set. This process is known as multiple criteria decision making (MCDM). Examples include decision making in human resource management, water resource management, environmental management and site selection, energy policy issues, portfolio selection, transportation and routing selection, student admission. In complex system, there are several criteria that need to be satisfied; however, they are usually at least partly conflicting. If an alternative is selected in conjunction with other alternatives it has different (positive or negative) effects on a specific criterion. Thus, selected alternatives have *interaction* effects on different criteria. Often a DM in practices uses a simple weighted linear function to rank alternatives then select top few. Not considering *interdependencies* among alternatives can lead to an undesirable outcome. For an illustration consider the following example on Waste Disposal Location (WDL) problem taken from Rajabi et al. [12]. In WDL problem, the objective is to identify the two best among five potential sites of equal capacity. Three criteria for success has been identified, (1) *proximity to population*, (2) *infrastructure requirements*, and (3) *environmental risk*. The WDL problem is characterized by certain interdependencies among alternatives. Building a new road near sites 4 and 5 could serve both sites; if both sites are selected, a saving in infrastructure investment will be obtained. If both sites, 4 and 5, are selected there is an increased of 10% on the infrastructure investment criterion. Similarly, if sites 1 and 2 are simultaneously selected then a single power plant facility may be built for both, taking advantage of economy of scale. In that case, a positive energy of 30% is estimated when it comes to infrastructure investment criterion. Finally, when it comes to environmental risk criterion, selection of site 4 and 5 together has a negative effect of 30% . The scenario is illustrated in the Figure 1, and data is presented in Table 1.

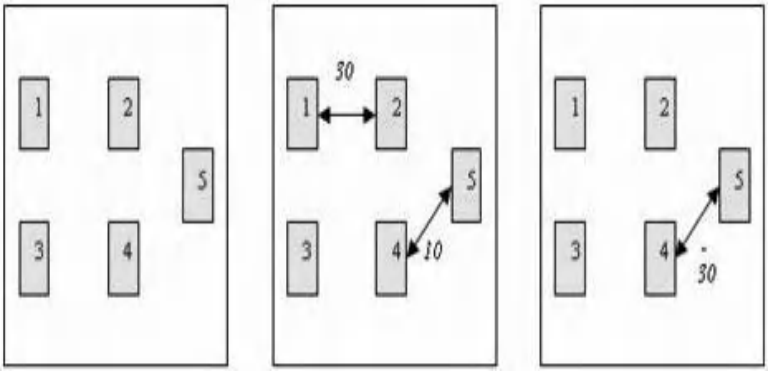


Fig. 1. WDL example: Interdependence actions and synergy levels. (Left: Proximity to population, Middle: Infrastructure requirements, Right: Environmental risk).

Table 1. Normalized consequences of five feasible sites

	Weights	Criteria	Alternatives				
			a^1	a^2	a^3	a^4	a^5
1	0.23	Population	0.45	0.45	1.00	0.55	0.84
2	0.39	Infrastructure	0.80	0.70	0.75	0.83	0.83
3	0.38	Environmental Risk	0.60	0.87	0.50	0.75	0.60
Additive Value			0.644	0.707	0.713	0.735	0.745

The following notations are adopted in the paper:
$N = \{a^1, \ldots, a^n\}$ is set of alternatives.
$a^i, (i = 1, \ldots, n)$ is alternative i.
$P = \{1, \ldots, n\}$ is set of criteria.
c_k^i is effect of alternative i on criterion k.
S_k is a set of alternatives that if selected together have some positive or negative effect on criterion k.
$\gamma(S_k)$ is the amount of effect (positive or negative) of an *interacting* set S_k on criterion k.
w_k is weight associated with criterion k.
$v(S)$ is total payoff of a subset of alternatives $S \subseteq N$.
x_i is equal to 1 if an alternative a^i is selected and 0 otherwise.

To illustrate the notations, consider Table 1, we have 5 alternatives $N = \{1, 2, 3, 4, 5\}$, and 3 criteria $P = \{1, 2, 3\}$. Alternative A^2 has effects of $c_1^2 = 0.45, c_2^2 = 0.70$, and $c_3^2 = 0.87$ on criterion, 1, 2, and 3, respectively. Weights of criteria are $w = (w_1, w_2, w_3) = (0.23, 0.39, 0.38)$. Considering criterion 2 there are two interacting alternatives, $S_2 = \{a^1, a^2\}$ and $S_2 = \{a^4, a^5\}$, with interaction effects (positive energy) of $\gamma(\{a^1, a^2\}) = 0.30$ and $\gamma(\{a^4, a^5\}) = 0.10$. If no

interdependencies are considered among the alternatives then the payoff value is defined as a simple additive function, for example, for $S = \{a^i, a^j\}$ we have

$$v(a^i, a^j) = v(a^i) + v(a^j) = \sum_{k=1}^{3} w_k c_k^i + \sum_{k=1}^{3} w_k c_k^j.$$

In this case, considering all possible set of 2 alternatives and selecting a subset with the largest payoff we have optimal solution as $v(a^4, a^5) = v(a^4) + v(a^5) = 0.735 + 0.745 = 1.48$. However, this is a simplistic view of the situation. If we consider *interdependencies* between alternatives and taking into accounts positive and negative energies we have a *dynamic weight* for each criterion when applied to an alternative. The dynamic weight of an alternative depends on its interaction with other alternatives in the selected subset.

3 Model and Solution Methodology

In the following, we will present a pseudo-boolean formulation of the problem that can overcome both of these deficiencies if appropriate solution procedure is available. The pseudo-boolean optimization for the problem can be presented as a maximization problem given below [10].

$$\begin{aligned} \text{Max} \quad & \sum_{i \in N} x_i [\sum_{k \in P} w_k c_k^i] + \sum_{k \in P} \sum_{s_k \subseteq N} \left(w_k [\gamma(S_k)] [\sum_{i \in S_k} c_k^i] \right) \prod_{i \in s_k} x_i \\ \text{s.t.} \quad & \sum_{i \in N} x_i = m, \\ & x_i \in \{0, 1\}. \end{aligned}$$

In this problem, if we want to find any subset of alternatives with largest payoff, then we remove the equality constraint and solve the problem.

To illustrate the optimization problem we use the WDL example presented below.

$$\begin{aligned} \text{Max} \quad & .644x_1 + .707x_2 + .713x_3 + .735x_4 + .745x_5 + .39(.30)[.80 + .70]x_1x_2 + \\ & .39(.10)[.83 + .83]x_4x_5 - .39(.30)[.75 + .60]x_4x_5 \\ \text{s.t.} \quad & x_1 + x_2 + x_3 + x_4 + x_5 = 2, \\ & x_i \in \{0, 1\}. \end{aligned}$$

Simplifying the above formulation we get

$$\begin{aligned} \text{Max} \quad & .644x_1 + .707x_2 + .713x_3 + .735x_4 + .745x_5 + .1755x_1x_2 - .8916x_4x_5 \\ \text{s.t.} \quad & x_1 + x_2 + x_3 + x_4 + x_5 = 2, \\ & x_i \in \{0, 1\}. \end{aligned}$$

Glover [13] used an efficient *1-flip* search implementation and tested it in the context of Tabu Search and Scatter Search.

Definition 1. *Let $x = (x_1, \ldots, x_i, \ldots, x_n)$ be a solution to pseudo-boolean and let $x = (x_1, \ldots, \bar{x}_i, \ldots, x_n)$ be a solution obtained from x by complementing x_i.*

Since Q is upper triangular matrix then (ij)th element,q_{ij},is equal to the second derivative of $f(x)$, i.e.,$\delta_{ij}(x) = q_{ij}$, with respect to x_i and x_j.

Theorem 1 (Glover et al.[14]). *For x and x' as given in Definition 1, the change in the value of the objective function is*

$f(x') - f(x) = (x'_i - x_i)\Delta_i(x)$

Moreover, in any local optimal solution of the pseudo-boolean with respect to 1-flip search we have

either $(x_i = 0 \iff \Delta_i(x) \geqslant 0)$ *or* $(x_i = 1 \iff \Delta_i(x) \leqslant 0)$, $\forall i = 1, \ldots, n$.

Furthermore, after changing x_i to $\bar{x}_i$ the update for all $\Delta_i(x)$ can be calculated as follows

$$\begin{aligned} &\forall j < i, && \Delta_j(x) = \Delta_j(x) + q_{ji}(x'_i - x_i), \\ &\forall j > i, && \Delta_j(x) = \Delta_j(x) + q_{ij}(x'_i - x_i), \\ &j = i, && \Delta_j(x) = \Delta_i(x). \end{aligned}$$

In Theorem 2, we extend the above result to general case for *r-flip* search for pseudo-boolean.

Theorem 2. *Let x be a given solution of pseudo-boolean and x' obtained from x by an r-flip move where $S \subseteq V$, and $|S| = r$. Now, the change in the value of the objective function is*

$f(x') - f(x) = \sum_{i \in S}(\bar{x}_i - x_x)\Delta_i(x) + \sum_{\forall i,j \in S}(\bar{x}_i - x_i)(\bar{x}_j - x_j)\Delta_{ij}(x)$.

Furthermore, after changing x to x' the update for all $\Delta_i(x)$ can be calculated as follows,

$$\begin{aligned} &\forall j \in V \setminus S && \Delta_j(x) = \Delta_j(x) + \sum_{i \in S}(\bar{x}_i - x_i)\Delta_{ij}(x), \text{and} \\ &\forall j \in S && \Delta_j(x) = \Delta_j(x) + \sum_{i \in S \setminus \{j\}}(\bar{x}_i - x_i)\Delta_{ij}(x). \end{aligned}$$

Proof. For two vectors x' and x where component x_i for all $i \in S$ are complemented we have $f(x)$ and $f(x')$ as follows:

$f(x) = \sum_{i \in S} x_i[q_{ij} + \sum_{j \notin S} x_j q_{ij}] + \sum_{ij \in S} x_i x_j q_{ij} + \ldots$

$f(x) = \sum_{i \in S} x_i[\Delta_i(x) - \sum_{j \notin S \setminus \{i\}} x_j q_{ij}] + \sum_{ij \in S} x_i x_j q_{ij} + \ldots$

$f(x') = \sum_{i \in S} \bar{x}_i[q_{ij} + \sum_{j \notin S} x_j q_{ij}] + \sum_{ij \in S} \bar{x}_i \bar{x}_j q_{ij} + \ldots$

$f(x') = \sum_{i \in S} \bar{x}_i[\Delta_i(x) - \sum_{j \notin S \setminus \{i\}} x_j q_{ij}] + \sum_{ij \in S} \bar{x}_i \bar{x}_j q_{ij} + \ldots$

$f(x') - f(x) = \sum_{i \in S}(\bar{x}_i - x_i)[\Delta_i(x) - \sum_{j \notin S \setminus \{i\}} x_j q_{ij}] + \sum_{ij \in S}(\bar{x}_i \bar{x}_j - x_i x_j)q_{ij}$

$f(x') - f(x) = \sum_{i \in S}(\bar{x}_i - x_i)\Delta_i(x) - \sum_{j \in S} \sum_{j \notin S \setminus \{i\}}(\bar{x}_i - x_i)x_i q_{ij} + \sum_{ij \in S}(\bar{x}_i \bar{x}_j - x_i x_j)q_{ij}$

For each pair of elements i, j in the second and the third terms we have
$\bar{x}_i\bar{x}_j + x_ix_j - \bar{x}_ix_j - x_i\bar{x}_j = (\bar{x}_i - x_i)(\bar{x}_j - x_j)$
From this we get
$f(x') - f(x) = \sum_{i\in S}(\bar{x}_i - x_i)\Delta_i(x) + \sum_{\forall i,j\in S}(\bar{x}_i - x_i)(\bar{x}_j - x_j)\Delta_{ij}(x)$
which proves the first part of the theorem.

Note that we can substitute $\Delta_{ij}(x)$ for q_{ij} . From this the desired result follows immediately. We propose several methods to solve the transformed pseudo-boolean problem.

Filtration and Sequential Fan(F&F) (see Glover[13]) is a search process utilizing compound move strategies that have proven to be effective for a variety of combinatorial problems. Here we employ a simple version of F&F utilizing a combination of *r-flip* moves. In the paragraph below we give a brief overview of our implementation.

F&F organizes an aggressive search process utilizing a variety of different *r-flip* moves. The version we tested here starts with a set of locally optimal solutions found using *1-flip* search. Then, for each solution in this set a second round of *1-flip* moves is executed. If an improving move is found we accept that move and initiate a complete new round of *1-flip* moves. If an improving second flip move can not be found, we initiate a series of *3-flip* moves starting with the best local optimal solutions found so far. This continues until a pre-set stopping criterion is satisfied or until no further improvement is realized.

4 Computational Results

Table 2 presents all subset of 2 alternatives and associated weights of each criterion, and payoff value of the selected subset of alternatives. In that, obviously subset (a^4, a^5) is not any more the optimal solution as we have $v(a^4, a^5) = 1.391$. However, the subset (a^1, a^2) has the largest payoff of $v(a^1, a^2) = 1.526$ which is optimal. Table 3 illustrates calculation of weights for each criterion for all subsets of 4 alternatives solved via Method 4. The dynamic nature of weights is clear from these tables. Obviously, when the interdependencies among

Table 2. Calculation of payoff when interdependencies are considered

2 alternatives (a^i, a^j)	w_1	w_2	w_3	$v(a^i, a^j)$
1,2	0.23	0.39(1+0.30)=0.507	0.38	1.526
1,3	0.23	0.39	0.38	1.356
1,4	0.23	0.39	0.38	1.379
1,5	0.23	0.39	0.38	1.388
2,3	0.23	0.39	0.38	1.420
2,4	0.23	0.39	0.38	1.442
2,5	0.23	0.39	0.38	1.452
3,4	0.23	0.39	0.38	1.448
3,5	0.23	0.39	0.38	1.457
4,5	0.23	0.39(1+0.10)=0.429	0.38(1-0.30)=0.266	1.391

Table 3. Calculation of weights for subset of 4 alternatives when interdependencies are considered

	a^1	a^2	a^3	a^4	a^5
w_1	0.23	0.23	0.23	0.23	
w_2	0.39(1+0.30)=0.507	0.39(1+0.30)=0.507	0.39	0.39	
w_3	0.38	0.38	0.38	0.38	
w_1	0.23	0.23	0.23		0.23
w_2	0.39(1+0.30)=0.507	0.39(1+0.30)=0.507	0.39		0.39
w_3	0.38	0.38	0.38		0.38
w_1	0.23	0.23		0.23	0.23
w_2	0.39(1+0.30)=0.507	0.39(1+0.30)=0.507		0.39(1+0.10)=0.429	0.39(1+0.10)=0.429
w_3	0.38	0.38		0.38(1-0.30)=0.266	0.38(1-0.30)=0.266
w_1	0.23		0.23	0.23	0.23
w_2	0.39		0.39	0.39(1+0.10)=0.429	0.39(1+0.10)=0.429
w_3	0.38		0.38	0.38(1-0.30)=0.266	0.38(1-0.30)=0.266
w_1		0.23	0.23	0.23	0.23
w_2		0.39	0.39	0.39(1+0.10)=0.429	0.39(1+0.10)=0.429
w_3		0.38	0.38	0.38(1-0.30)=0.266	0.38(1-0.30)=0.266

alternatives on different criteria are high the level of *dynamism* for each weight is also high. While considering the interdependencies among alternatives represents the practical nature of the problem, however calculating the dynamic weights is very difficult. If we need to select m alternatives with highest total payoff we have to consider calculation of weights in all possible m set of alternatives. And, in case, the problem is to select the best subset of alternatives (with no pre-specified number of elements) we have to calculate weights for all subset of alternatives. In both cases, the number of subsets is exponential.

5 Summary and Conclusion

In this paper we reported on modeling complex system problems with the pseudo-boolean formulation and proposed several solution approaches. The complex systems with very high-dimensional spaces often associate with more complicated variables. The pseudo-boolean formulation can reduce the variable dimension without loss important information. The preliminary results illustrate the attractiveness of this modeling method in term of generating the best alternative. In our continuing work, we intend to carry out further testing of pseudo-boolean formulation on cubic, quartic, even quintic variables.

References

1. Carpenter, G.A., Grossberg, S. and Rosen, D.B.: Fuzzy ART: Fast stable learning and categorization of analog patterns by an adaptive resonance system, Neural Networks 4 (1991) 759–771
2. Axelrod, R.: The Complexity of Cooperation: agent-based models of competition and collaboration. Princeton: Princeton University Press (1997)

3. Döner D.: Heuristics and Cognition in Complex Systems, in: Methods of Heuristics, Groner R., Groner M. Bischof W.F. ed., Erlbaum, Hillsdale NJ(1983), 89–107
4. Holland J.H.: Studies of the spontaneous emergence of self-replicating systems using cellular automata and formal grammars. In A. Lindenmayer and G. Rozenberg, editors, Automata, Languages, Development, North-Holland, New York, (1976) 385–404
5. Stuart A.K.:The origins of order: self-organization and selection in evolution. Oxford University Press, Oxford, (1993)
6. Kaplan, D., Glass, L.: Understanding Nonlinear Dynamics, Springer-Verlag, NY (1995)
7. Ritchey, T.: Problem Structuring using Computer-Aided Morphological Analysis, Journal of the Operational Research Society, 57(2006), 792-801
8. Hayles, N. K. : Introduction: Complex Dynamics in Literature and Science, in Chaos and Order: Complex Dynamics in Literature and Science, (N. K. Hayles, ed.), Univ. Chicago Press, Chicago (1991) 1–36
9. Hammer, P., Rudeanu, S.: Boolean Methods in Operations Research and Related Areas, Springer, Berlin, Heidelberg, New York, (1968)
10. Boros, E., Hammer, P.: Pseudo-Boolean Optimization, Discrete Applied Mathematics, 123(2002), 155–225
11. Crama, Y., Hammer, P.: Boolean Functions: Theory, Algorithms and Applications (to be published in 2006)
12. Rajabi, S., Hipel,K.W., Kilgour, D.M.: Multiple criteria decision making under interdependence of actions - Systems, Man and Cybernetics, IEEE International Conference on Systems, Man and Cybernetics –Intelligent Systems for the 21st Century (1995)3, 2365–2370
13. Glover, F.: A Template for Scatter Search and Path Relinking, Working Paper, School of Business Administration, The University of Colorado. (1998)
14. Glover, F., Kochenberger, G., Alidaee, B.: Adaptive Memory Search for Binary Quadratic Programs, Management Science, 44(1998), 336–345

The Study of Mission Reliability of QRMS Based on the Multistage Markov Process

Liang Liang and Bo Guo

School of Information System and Management, National University of Defence Technology, Changsha, Hunan Province, 410073, P.R. China
{Liang Liang,Bo Guo,doulfin}@gmail.com

Abstract. The iteration is a fundamental characteristic of the complex developing process. The developing mission of new product, during the quick response manufacturing process, will be iterated step by step following certain probability rules at the work nodes. Based on the analysis of the development mission process, the method of the multi-stages, finite state Markov process is presented to describe the system process. The models of the mission reliability for both task stages and system have then been built. Finally, the quantitative methods of mission reliability calculation were explained in details through the analysis of an actual developing process. The methods can also be used to evaluate the development system.

Keywords: Quick response manufacturing system(QRMS), Design iteration, Mission reliability.

1 Introduction

The Quick Response Manufacturing System (QRMS) is based on the quick response concept with the purpose to quick response to the demand for developing new products and carry out the design and manufacturing process in a very timely manner. The frequent changing of market demand requires higher competence of new product development. On one hand, companies need to analyze its current development & manufacturing system to estimate, from the perspective of mission reliability, system's ability to finish the task in the required time period. On the other hand, companies may also need to improve on those weaknesses that pertain to the system therefore increase its competence of development.

Iteration and conflict of resource are the most two important factors in an estimation of the mission reliability for a QRMS. Iteration is a basic characteristic of the complex development process[1, 2]. The iteration processes means rework and improve on previous work and that always leads to risk of scheduling and repeated planning and eventually make the whole developing process delayed. The fast changing demand of market gets the QRMS overloaded and that means a task always has to queue in the process due to the lack of resource. Hence, the task cannot be finished on time and companies will suffer loss.

Currently, Design structure Matrix[3] (DSM) are widely used for the time evaluation of product development process. DSM has been extended to the numerical DSM,

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 202–209, 2007.
© Springer-Verlag Berlin Heidelberg 2007

which can further describe the iteration relation among different developing processes. In DSM, probability of rework indicates the possible rework of the whole developing process. However, the scenario derived from DSM that the probability of rework is a constant does not reflect the actual operating rule. Therefore, this paper extends the probability of rework scenario by taking the changing probability of rework during different task developing stages into account.

The conflict of recourse in the developing process can be better reflected using the model of queuing network. Meanwhile, the transferring matrix of queuing network can be related to the matrix derived from DSM. Hence, the queuing network model has greater applicability as during the developing process. The paper analyzes the task developing process of quick response manufacturing system and describes the dynamic probability of rework by adopting multi-stages, finite states Markov process. The paper finally figures out the analytical model and arithmetic for task reliability calculation by analyzing and modelling those probabilities using queuing network. The following cases further show how arithmetic works.

2 Problem Formulation

Suppose there are M nodes in a QRMS and each division has one workgroup. The service time for each workgroup follows exponential distributions with parameter μ_i $(i=1,2,3...N)$. Additional predetermined assumptions are:

1. Those tasks that have been undertaken by the system are homogeneous. The arrival process is Poisson with arrival rate λ. The tasks arrive at the first node from outside and finish all the service work in the order of the nodes. When task arrives, the idle workgroup reacts immediately while others keep waiting on a queue. Services at all nodes are independent and the number of queues at each node is limitless.

2. In a task developing process, the reworks exist and are allocated in each node following certain probability. When a task is first finished at node i, it can only be returned to node $(1,2,\cdots,i-1)$ if any rework occurred.

3. Followed by the proceeding of a task among nodes, the probability of rework will change because new information arrives.

3 System Modelling and Analysis

3.1 Model Definition

Definition 1:

Set $A=\left\{t_i^{(1)}, i=1,2,3...N\right\}$, $t_i^{(1)}$ is referred to as the point of time when the task first arrives (be proceeded) at node i. Set $t_1^{(1)}$=0, and $t_{N+1}^{(1)}$ is referred to as the point of time when task has been finished.

From the hypothesis 1, if the points of time $t_1^{(1)}$ $t_2^{(1)}$ ... $t_{N+1}^{(1)}$ all exist, then:

$$t_1^{(1)} \le t_2^{(1)} \le ... \le t_{N+1}^{(1)} \tag{1}$$

Definition 2:

The stage I of a task is referred to as the process of a task from node i to node $i+1$, which can defined as the period $\left\{\zeta_i \left| t_i^{(1)} \le \zeta_i < t_{i+1}^{(1)}, i = 1,2,3...N \right.\right\}$.

Definition 3:

Set $X_I = t_{i+1}^{(1)} - t_i^{(1)}, i = 1,2,\cdots,N$, T_I indicates the given time to finish the task at stage I. The mission reliability of stage I can be defined as follows:

$$R_I\left(X_I\right) = P\left[X_I \le T_I\right]$$

Definition 4:

Set $X = t_{N+1}^{(1)} = \sum_{I=1}^{N} X_I$, T indicates the given time to finish the task. The system's mission reliability can be defined as:

$$R\left(X\right) = P\left[X \le T\right]$$

Definition 5:

The task will enter the second node directly after reaching the first node at first time and being finished at this node. That is, there is no rework in the first stage, symbolized as $\mathbf{P}_1 = [0]$. While in the stage $I\left(I = 2,\cdots,M\right)$, the rework at different nodes can be illustrated by the following matrix:

$$\mathbf{P}_I = \begin{bmatrix} r_{11}(I) & \cdots & r_{1i}(I) \\ \vdots & \ddots & \vdots \\ r_{i1}(I) & \cdots & r_{ii}(I) \end{bmatrix} \tag{2}$$

$r_{ij}(I)$: The probability of task being reworked from the first node i to node j $\left(0 < i, j \le I\right)$ when task has been processed in stage I.

3.2 The Single Task Process Modelling

The development process is shown in Fig. 1. It is known from the definition that under the system discussed in this paper, the rework of a task can only happen among the nodes which have been serviced already. Further, the probability of rework is changing upon the task being processed forward. This paper then puts forward a measure called Multistage, Finite-state, absorbing Markov chains for system modelling and analyzing.

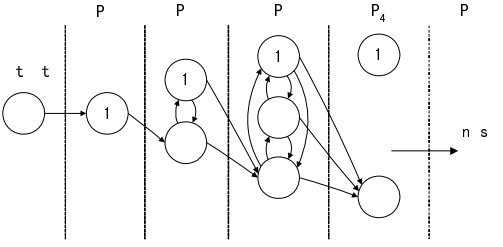


Fig. 1. The development process based on dynamic probability of rework

For single task, when it has been processed to stage I , at the observation time $t_i^{(1)}$, task reaches node i at first time. After that, a virtual absorbing node $(i+1)'$ will be constructed, that is $r_{(i+1)'j}=0$, $j=1,2,\cdots,i$, set:

$$\begin{cases} r_{j(i+1)'} = r_{j(i+1)} \\ t^{(1)}_{(i+1)'} = t^{(1)}_{(i+1)} \end{cases} \tag{3}$$

given that the matrix $\boldsymbol{\beta}_I=(\beta_{ij})=(\mathbf{P}_I-\mathbf{e})$ is invertible square matrix. (Throughout this paper, $\mathbf{e}$ will denote the column vector with all components equal to one, whose length is determined by the context in which it appears.), Without regard to the waiting time taken in the queue at nodes, it can be shown from Eq. (3) that the finite-state, absorbing Markov Chains has been constructed during the process of the task from node i to note $(i+1)'$. The initial state of the process is $(\boldsymbol{\alpha}_I,\alpha_{I+1})$, where $\boldsymbol{\alpha}_I=(0,\cdots,0,1)_{1\times I}$. The set of states is $\left(1,...,i,(i+1)'\right)$. Among that the state $(1,2,\cdots,i)$ are all transient, state $(i+1)'$ is absorbing. Before the process absorbed to the state $(i+1)'$, the mean visit times to each node can be described as the vector[4]:

$$\mathbf{K}_I=(k_I(1),\cdots,k_I(j),\cdots,k_I(i))=\boldsymbol{\alpha}_I\int_0^\infty \exp\left[(\mathbf{P}_I-\mathbf{e})x\right]dx=-\boldsymbol{\alpha}_I\boldsymbol{\beta}_I^{-1} \tag{4}$$

3.3 Steady-State System Development Process Model

3.3.1 Steady-State Mean Arrival Rate of Each Node

From assumption we can know that the arrivals of tasks follow independent, exponentially distributed interarrival times with rate λ. The task service time at each node also follows exponentially distributed interarrival times with rate μ_i. Under steady state, every node in the queuing network can be viewed as a $M/M/1$ queuing system. Thus, the arrival rate of a task entering quick response manufacturing system, under steady state, must equalize its departure rate. Assume that total mean arrival

rate (the sum of the arrival rate from outside and those from other nodes to node i) is Λ_i at node i, then

Theorem 1. under steady state, the total mean arrival rate of each node is

$$\Lambda_j = \lambda \sum_{I=1}^{N} k_I(j) \qquad (j = 1\ 2, \cdots, N) \tag{5}$$

Proof: Under steady state, the jobs at node j come from every stage in the whole task process. From Eq.(4), in the task stage I, the mean times for task visiting node $j(j = 1,2,\cdots,i)$ is $k_I(j)$, therefore, under steady state, the sum of the mean times for task to come from every stage visiting node j is $\sum_{I=1}^{N} k_I(j)$.

From model assume: the task stage is the sequence, the task from outside to the first node in the system follows the Poisson process with parameters λ, and the other nodes haven't got any input from outside. Then the steady-state overall arrival rate at node $j(j = 1,2,\cdots,N)$ can be written as:

$$\Lambda_j = \lambda \sum_{I=1}^{N} k_I(j) \qquad (j = 1,2,\cdots,N)$$

The theorem has then been proved.

3.3.2 Steady-State Response Time Distribution $W_i(t)$ of Node i

When the utilization ρ_i of the node i meets:

$$\rho_i = \frac{\Lambda_i}{u_i} < 1 \qquad (i = 1,2,\cdots,N) \tag{6}$$

queueing network will leave the steady-state distribution, which the marginal queue length distribution of every node at the queueing networks is the same as that of the $M/M/1$ queueing system[5]. Then the distribution of steady-state response time (the sum of the waiting time and the server time) at node $i(i = 1,2,\cdots,N)$ is:

$$W_i(t) = P[V_i \le t] = \mathrm{E}\left[P[V_i \le t] \middle| N = n_i\right] = 1 - e^{-(\mu_i - \Lambda_i)t}, \quad t \ge 0 \tag{7}$$

From Eq.(7), the steady-state distribution at node i is exponentially distribution with parameters:

$$\lambda_i = \mu_i - \Lambda_i \tag{8}$$

3.4 The System Modelling for Mission Reliability

3.4.1 Mission Reliability Model of the Task Stage

Theorem 2. set $\boldsymbol{\lambda}_i = (\lambda_1, \lambda_2, \cdots \lambda_i)_{1\times i}$, $i = 1,2,\cdots,M$ under steady state, mission reliability at task stage I:

$$R_I(t) = 1 - \boldsymbol{\alpha}_I \exp[\mathbf{T}_I t]\mathbf{e},\ t \geq 0 \tag{9}$$

Where, $\boldsymbol{\alpha}_I = (0,\cdots,0,1)_{1\times I}$, $\mathbf{T}_I = (T_{kj}) = \{\lambda_i \beta_{ij}\}$, β_{ij} is the matrix element defined by the Eq.(4).

Proof: From Eq.(8), the steady-state distribution at every node is exponentially distribution. Therefore, in stage I, the process constructed in the 2.2 section was a finite-state Markov process with generator:

$$\mathrm{Q}_I = \begin{pmatrix} \mathbf{T}_I & \mathbf{T}_I^0 \\ \mathbf{0} & 0 \end{pmatrix}$$

The matrix $\mathbf{T}_I = (T_{kj}) = \{\lambda_i \beta_{ij}\}$ of order I is non-singular. It has negative diagonal elements and nonnegative off-diagonal elements and satisfies $\mathbf{T}_I\mathbf{e} + \mathbf{T}_I^0 = \mathbf{0}$. The initial state of the process is $(\boldsymbol{\alpha}_I, \alpha_{I+1})$, where $\boldsymbol{\alpha}_I = (0,\cdots,0,1)_{1\times I}$. Then the absorbing time distribution of Markov process follows the Phase-Type(PH) distribution. The distribution function is

$$F_I(t) = P[t \leq T_I] = 1 - \boldsymbol{\alpha}_I \exp[\mathbf{T}_I t]\mathbf{e},\ t \geq 0$$

Where, T_I was the given time to finish the task at stage I.

From Definition 3, it is proved that $R_I(t) = F_I(t)$.

3.4.2 Mission Reliability Model of the System

From Definition 3, $X_I\ (i = 1,2,\cdots,N)$ was defined as the actual task serving time at stage I, and based on the system hypothesis, X_I is non-negative independent random variables, and the total time of a task $X = \sum_{I=1}^{N} X_I$, then the total service time distribution of a task is the convolution of the service time distribution of each task stage.

The convolution of the limited number of PH distribution is also the PH distribution[4]. It is obviously that it is the finite process for a task in QRMS. The total service time distribution of a task also follows the PH distribution denoted as $(\boldsymbol{\gamma}, \mathbf{L})$, then the mission reliability of the system

$$R(t) = P[t \leq T] = 1 - \boldsymbol{\gamma} \exp[\mathbf{L}t]\mathbf{e},\ t \geq 0 \tag{10}$$

Where T is the given time to finish the task

4 Case Study

The key accessory design for an aero engine makes up of five working nodes: unit design, finite analysis, artwork design, mould design, and quick-to-mockup. The task

departs the system after going through above five nodes one by one. The arrivals of tasks follow the exponential distribution with parameters $\lambda = 0.007$.

Table 1. The parameter of the nodes

Nodes_No.	1	2	3	4	5
μ_i	0.0240	0.0310	0.0320	0.0252	0.0350

The rework probability matrixes for each stage are as follows:

$$P_1 = [0] \qquad P_2 = \begin{bmatrix} 0 & 0.95 \\ 0.15 & 0 \end{bmatrix} \qquad P_3 = \begin{bmatrix} 0 & 0.90 & 0 \\ 0.12 & 0 & 0.8 \\ 0.3 & 0.2 & 0 \end{bmatrix}$$

$$P_4 = \begin{bmatrix} 0 & 0.85 & 0 & 0 \\ 0.1 & 0 & 0.7 & 0 \\ 0.25 & 0.15 & 0 & 0.5 \\ 0.1 & 0 & 0 & 0 \end{bmatrix} \quad P_5 = \begin{bmatrix} 0 & 0.7 & 0 & 0 & 0 \\ 0.05 & 0 & 0.6 & 0 & 0 \\ 0.2 & 0.1 & 0 & 0.4 & 0 \\ 0.05 & 0 & 0 & 0 & 0.8 \\ 0 & 0 & 0 & 0.1 & 0 \end{bmatrix}$$

Detail calculation result Based on Eq. (5) and Eq. (8), the total mean arrival rate Λ_i and the mean response time $1/\lambda_i$ can be shown in Table 2. Based on Eq. (9) & (10), the mission reliability curves for each stage and the whole system can be showed in Fig. 2 .

Table 2. Steady-state arrival rate Λ_i and mean serving time $1/\lambda_i$

node_no	1	2	3	4	5
Λ_i	0.0137	0.0155	0.0128	0.0081	0.0076
$1/\lambda_i$	96.674	64.575	52.029	58.477	36.517

From Fig. 2, it is known that under the same requirement for mission reliability, the task in the third stage takes the longest time to be finished, that is, it takes the longest time for a task being processed from node 3 to node 4. Therefore, the third stage is the bottleneck for the whole mission. Table 2 shows the average waiting time at node 3 is 52.029h. Such long time being taken is mainly due to the reworks caused by design iteration. Hence the designer can consider some measures to reduce the design failure and improve the level of system's mission reliability level. Possible measures are: the increase of design quality control; more training for designing staff, and so on.

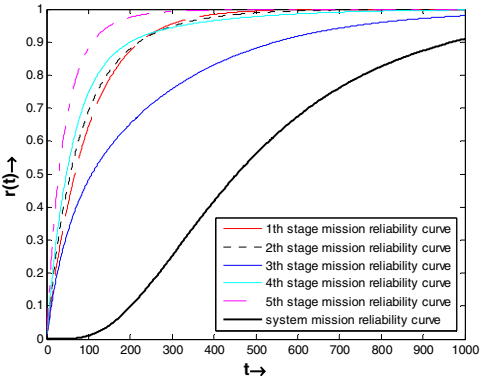


Fig. 2. Mission reliability curve of each stage and system

5 Conclusion

By analyzing the mission reliability for the QRMS, a model has been built to reflect such mission reliability under dynamic iteration and conflict of resource. Consequently, a quantitative solution has been given to calculate those mission reliabilities for each stage and for the whole system. Both the model and solution give an effective tool to evaluate the mission reliability of quick response manufacturing. They also facilitate the discovery of system's weaknesses and give great help to improve those weaknesses.

References

1. Eppinger, S., Whitney, D., Smith, R., Gebala, D. A Model-Based Method for Organizing Tasks in Product Development[J]. Research in Engineering Design. 1994.6: pp.1-13.
2. Cho, S.-H., Eppinger, S. D. A simulation-based process model for managing complex design projects[J]. IEEE Transactions on Engineering Management. 2005(52).3: pp.316-328.
3. Peter, B., LuhFeng, L., Bryan, M. Scheduling of Design Projects with Uncertain Number of Iterations[J]. European Journal of Operational Research. 1999(113).3: pp.575--592.
4. Naishuo, T., Dequan, Y. Quasi Birth-and-death Process and Matrix--Geometric Solutions(in Chinese)[M]. Beijing: Science Press. 2002.
5. Bolch, G., Greiner, S., Meer, H. d., Trivedi, K. S. Queueing Networks and Markov Chains--Modeling and Performance Evaluation with Computer Science Applications[M]. Hoboken: John Wiley & Sons, Inc. 2006.

Performance Analysis and Evaluation of Digital Connection Oriented Internet Service Systems

Shunfu Jin[1] and Wuyi Yue[2]

[1] College of Information Science and Engineering
Yanshan University, Qinhuangdao 066004 China
jsf@ysu.edu.cn
[2] Department of Information Science and Systems Engineering
Konan University, Kobe 658-8501 Japan
yue@konan-u.ac.jp

Abstract. In this paper, we propose a discrete-time connection oriented Internet service system with a release delay, and present performance analysis and evaluation of the system. We build a batch arrival Geom*/G/1 queueing model with a setup/close-delay/close-down strategy to characterize the system operation, and suppose that the batch size is a random variable following Pareto(c, α) distribution. We describe the performance measures for setup ratio and utility of connection based on numerical results.

Keywords: Connection oriented service, burst, queueing system.

1 Introduction

Advancement in Internet services is urgently needed to satisfy service-specific Quality of Service (QoS) requirements. For example, different service classes offering packet traffic considered in terms of the required packet delay, response time and system throughput need to meet the real time and high transmission rate. For design and tuning of the advanced Internet services, system performance must be mathematically analyzed and numerically evaluated [1]. Queueing theory and Markov chains are used for the performance and reliability evaluation of communication networks.

In recent years, there have been many achievements in the research and application of continuous-time vacation queueing, especially in the performance evaluation of communication [2]. However, it is indicated that it would be more accurate and efficient to use discrete-time queue models rather than continuous counterparts when analyzing and designing digital transmitting systems [3]. The generally accepted view is that discrete-time systems can be more complex to analyze than equivalent continuous-time systems. The classical discrete-time queuing analyses can be found in [3], [4]. Analyses of discrete-time queue models with server vacation or setup strategy can be found in [5], [6].

Taking into account the memoryless character of a user initiated connection oriented session in a switched virtual channel, a delayed vacation Geom/G/1

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 210–217, 2007.
© Springer-Verlag Berlin Heidelberg 2007

queue model with setup was built. Some performance measures were then calculated in [7], [8].

In this paper, we propose a discrete-time connection oriented Internet service system with a release delay set before a release process, and present performance analysis and evaluation of the system. We also introduce an upper limit length of a release delay called timer length T, as a system parameter to control the length of the release delay. To give performance evaluation for this system, we build a batch arrival Geom*/G/1 queue model with a setup/close-delay/close-down strategy to characterize the system operation. We derive performance measures such as the setup ratio and utility of connection. In numerical results, we show that the setting of the timer length T is significant for improving system performance.

2 System Model

We assume the time axis to be divided into slots of equal length and batch arrivals to follow a Bernoulli process with a batch size following Pareto(c, α) distribution. Packets in batches arrived in the buffer having infinite capacity are transmitted with a common channel in a First Come First Serve (FCFS) discipline. The system works as follows. And this process will be repeated.

(1) When a batch arrives in the system, a setup period called setup period U will be started, where the setup period U corresponds to a time period for setting up a new connection using a three-handshake signaling procedure.
(2) After the setup period U finishes, a busy period called busy period Θ will begin. The busy period Θ is a time period in which packets are transmitted continuously until the buffer becomes empty.
(3) When there are not any packets in the buffer to be transmitted, the system will enter a close-delay period called close-delay period D. The close-delay period D corresponds to a release delay time with an upper limit timer length T in slots $(0 \leq D \leq T)$. During the close-delay period D, the connection is reserved in anticipation of more packets being transmitted using the same connection before the system goes into close-down phase. A close-delay period will finish either T is over or a batch arrives within T.
(4) If there is a batch arrival within the close-delay period D, a new busy period Θ will be started immediately without a setup period U. Otherwise, the system will enter a close-down period called close-down period C when the timer length T is over. C corresponds to the time period required to release the connection using another three-handshake signaling procedure.
(5) If there is a batch arrival during the close-down period C, after C finishing, a setup period U will begin. If this does not occur, the system will enter an idle period called idle period I. A batch arriving during the idle period I will make the system to enter a new setup period U.

We define a transmission period B as being the time period in slots taken to transmit a single packet with generally distributed.

The setup period U, close-down period C and transmission period B are independent and identical discrete-time random variables in slots, and are assumed to be generally distributed with Probability Generation Functions (P.G.Fs.) $U(z)$, $C(z)$ and $B(z)$, respectively. Then $U(z)$, $C(z)$ and $B(z)$ are defined as follows:

$$u_k = P\{U = k\},\ U(z) = \sum_{k=1}^{\infty} u_k z^k, \quad c_k = P\{C = k\},\ C(z) = \sum_{k=1}^{\infty} c_k z^k,$$
$$b_k = P\{B = k\},\ B(z) = \sum_{k=1}^{\infty} b_k z^k. \tag{1}$$

Let $E[U]$, $E[C]$ and $E[B]$ be the means of U, C and B in slots, we have

$$E[U] = \sum_{k=1}^{\infty} k u_k, \quad E[C] = \sum_{k=1}^{\infty} k c_k, \quad E[B] = \sum_{k=1}^{\infty} k b_k.$$

To consider the batch arrivals in the system, we denote by Λ the number of packets in a batch called batch size Λ (packets/batch). Λ in this system is supposed to be a random variable. Let $E[\Lambda]$ be the mean of the batch size Λ. We can obtain the probability distribution, the P.G.F. $\Lambda(z)$ and $E[\Lambda]$ of Λ as

$$\lambda_k = P\{\Lambda = k\}, \quad k \geq 0, \quad \Lambda(z) = \sum_{k=0}^{\infty} \lambda_k z^k, \quad E[\Lambda] = \sum_{k=0}^{\infty} k\lambda_k \tag{2}$$

where λ_k is the probability that there are k packets in a batch per slot. Specially, $\lambda_0 = P\{\Lambda = 0\}$ is the probability that there is no batch ($\Lambda = 0$) arrival in a slot. We define the probability of no batch arrival during the timer length T to be $T(\lambda_0)$, where $T(\lambda_0) = \lambda_0^T$. From Eq. (1), we also know that the probability of no batch arrival during the close-down period C is $C(\lambda_0) = \lambda_0^C$ and the probability of no batch arrival during the transmission period B is $B(\lambda_0) = \lambda_0^B$. The ergodic condition is $\rho = E[\Lambda]E[B] < 1$.

Let A_U, A_C and A_B be random variables representing the numbers of packets arriving during U, C and B. We also define $\Lambda(B(z))$ to be the P.G.F. of the transmission time of a batch in slots. Then we can give $\Lambda(B(z))$ and the P.G.Fs. $A_U(z)$, $A_C(z)$ and $A_B(z)$ of A_U, A_C and A_B as follows:

$$\Lambda(B(z)) = \sum_{k=0}^{\infty} \lambda_k (B(z))^k, \quad A_U(z) = \sum_{k=1}^{\infty} u_k (\Lambda(z))^k = U(\Lambda(z)),$$
$$A_C(z) = \sum_{k=1}^{\infty} c_k (\Lambda(z))^k = C(\Lambda(z)), \quad A_B(z) = \sum_{k=1}^{\infty} b_k (\Lambda(z))^k = B(\Lambda(z)). \tag{3}$$

3 Performance Analysis

3.1 Queue Length

We assume that packet arrival and packet departure occur only at the boundary of a slot. Let $Q_n = Q(\tau_n^+)$ be the number of packets in the system immediately

after the nth packet departure. Then $\{Q_n, n \geq 1\}$ forms an imbedded Markov chain. We define the state of the system by the number Q of packets in the system at the imbedded Markov points as follows:

$$Q_{n+1} = \begin{cases} Q_n - 1 + A_B^{(n+1)}, & Q_n \geq 1 \\ \eta, & Q_n = 0 \end{cases} \tag{4}$$

where $A_B^{(n+1)}$ is the number of packets arriving during the transmission time of the $(n+1)$th packet. η is the number of packets left in the system after the departure of the first packet in a busy period Θ.

A busy period Θ begins with one of the following three cases:

(1) If there is a batch arrival within the timer length T, the batch will trigger a busy period Θ immediately. The P.G.F. $\eta_1(z)$ for this case is given as follows:

$$\eta_1(z) = \frac{1}{z}\frac{\Lambda(z) - \lambda_0}{1 - \lambda_0} A_B(z) = \frac{1}{z}\frac{\Lambda(z) - \lambda_0}{1 - \lambda_0} B(\Lambda(z)).$$

(2) If there is no batch arrival within either the timer length T or the close-down period C, the batch arriving during the idle period I will trigger a setup period U, and then a busy period Θ will begin. The P.G.F. $\eta_2(z)$ for this case is given as follows:

$$\eta_2(z) = \frac{1}{z}\frac{\Lambda(z) - \lambda_0}{1 - \lambda_0} A_U(z) A_B(z) = \frac{1}{z}\frac{\Lambda(z) - \lambda_0}{1 - \lambda_0} U(\Lambda(z)) B(\Lambda(z)).$$

(3) If there is no batch arrival within the timer length T, but there is at least one batch arrival within the close-down period C, after the close-down period C finished, the system will directly enter a setup period U and then a busy period Θ will begin. The P.G.F. $\eta_3(z)$ for this case is given as follows:

$$\eta_3(z) = \frac{1}{z}\frac{A_C(z) - C(\lambda_0)}{1 - C(\lambda_0)} A_U(z) A_B(z) = \frac{1}{z}\frac{A_C(z) - C(\lambda_0)}{1 - C(\lambda_0)} U(\Lambda(z)) B(\Lambda(z)).$$

With these three cases, we can give the P.G.F. $\eta(z)$ of η as follows:

$$\eta(z) = \frac{1}{z} B(\Lambda(z)) \Bigg(\frac{\Lambda(z) - \lambda_0}{1 - \lambda_0} (1 - T(\lambda_0) + C(\lambda_0) T(\lambda_0) U(\Lambda(z))) + T(\lambda_0) U(\Lambda(z)) (C(\Lambda(z)) - C(\lambda_0)) \Bigg). \tag{5}$$

From Eq. (4), we can obtain the P.G.F. $Q(z)$ of Q as follows:

$$Q(z) = P\{Q \geq 1\} E\left[z^{Q-1+A_B} | Q \geq 1\right] + P\{Q = 0\} \eta(z). \tag{6}$$

Substituting Eq. (5) to Eq. (6), we can give that

$$Q(z) = P\{Q = 0\} \frac{B(\Lambda(z))}{B(\Lambda(z)) - z} \Bigg(1 - \Bigg(\frac{\Lambda(z) - \lambda_0}{1 - \lambda_0} (1 - T(\lambda_0) + C(\lambda_0) T(\lambda_0) U(\Lambda(z))) + T(\lambda_0) U(\Lambda(z)) \left(C(\Lambda(z)) - C(\lambda_0)\right) \Bigg) \Bigg). \tag{7}$$

Using the normalization condition and L'Hospital principle in Eq. (7), we have

$$P\{Q=0\} = \frac{1-\rho}{K} \tag{8}$$

where K is given as follows:

$$K = \frac{E[\Lambda]}{1-\lambda_0}\left(1 - T(\lambda_0) + T(\lambda_0)C(\lambda_0) + E[\Lambda]T(\lambda_0)\left(E[U]+E[C]\right)\right). \tag{9}$$

Substituting Eq. (8) to Eq. (7), then the P.G.F. $Q(z)$ and the mean $E[Q]$ of Q can be obtained as follows:

$$\begin{aligned} Q(z) &= \frac{(1-\rho)(1-\Lambda(z))B(\Lambda(z))}{E[\Lambda](B(\Lambda(z))-z)}\left(\frac{E[\Lambda]}{(1-\lambda_0)K}(1-T(\lambda_0))\frac{E[\Lambda]T(\lambda_0)C(\lambda_0)}{(1-\lambda_0)K}\right. \\ &\quad \left.\cdot U(\Lambda(z)) + \frac{E[\Lambda](E[U]+E[C])T(\lambda_0)}{K}\cdot\frac{1-U(\Lambda(z))C(\Lambda(z))}{(E[U]+E[C])(1-\Lambda(z))}\right), \\ E[Q] &= \rho + \frac{\Lambda^{(2)} + (E[\Lambda])^3 B^{(2)}}{2(1-\rho)E[\Lambda]} + \frac{E[\Lambda]T(\lambda_0)C(\lambda_0)}{(1-\lambda_0)K}E[U] \\ &\quad + \frac{1}{2K}\left((E[\Lambda])^2 T(\lambda_0)\left(U^{(2)} + C^{(2)} + 2E[U]E[C]\right)\right). \end{aligned} \tag{10}$$

where $U^{(2)}$, $C^{(2)}$, $B^{(2)}$ and $\Lambda^{(2)}$ are the second factorial moments of the P.G.Fs. $U(z)$, $C(z)$ and $B(z)$ of the setup period U, close-down period C and transmission period B, and $\Lambda(z)$ of the batch size Λ by differentiating the P.G.Fs. presented in Eqs. (1) and (2) with respect to z and evaluating the result at $z=1$.

3.2 Waiting Time

We focus on an arbitrary packet in the system called tagged packet M. We note that the waiting time W of the tagged packet M can be divided into two parts as follows. One is the waiting time W_b of the batch that the tagged packet M belongs to. The other is the total transmission time J of the packets before the tagged packet M in the same batch. W_b and J are independent random variables, so we have the P.G.F. $W(z)$ of the waiting time W of the tagged packet M as $W(z) = W_b(z)J(z)$. Where $W_b(z)$ and $J(z)$ are P.G.Fs. of W_b and J.

Applying the analysis of delayed vacation Geom/G/1 queueing model with setup in [7] and referencing [3], with $\Lambda(B(z))$ given in Eq. (3), we have

$$\begin{aligned} W_b(z) &= \frac{(1-\rho)(1-z)}{\Lambda(B(z))-z}\left(\frac{E[\Lambda](1-T(\lambda_0))}{(1-\lambda_0)K} + \frac{E[\Lambda]T(\lambda_0)C(\lambda_0)}{(1-\lambda_0)K}U(z)\right. \\ &\quad \left. + \frac{E[\Lambda](E[U]+E[C])T(\lambda_0)}{K}\cdot\frac{1-U(z)C(z)}{(E[U]+E[C])(1-z)}\right), \end{aligned} \tag{11}$$

$$J(z) = \frac{1-\Lambda(B(z))}{E[\Lambda](1-B(z))}. \tag{12}$$

Concluding Eqs. (11) and (12), the P.G.F. $W(z)$ and the mean $E[W]$ of W can be obtained as follows:

$$\begin{aligned} W(z) &= \frac{1-\Lambda(B(z))}{E[\Lambda](1-B(z))} \cdot \frac{(1-\rho)(1-z)}{\Lambda(B(z))-z} \bigg(\frac{E[\Lambda]}{(1-\lambda_0)K}(1-T(\lambda_0)) \\ &\quad + \frac{E[\Lambda]T(\lambda_0)C(\lambda_0)}{(1-\lambda_0)K}U(z) + \frac{E[\Lambda]T(\lambda_0)(1-U(z)C(z))}{K(1-z)} \bigg), \\ E[W] &= \frac{\Lambda^{(2)}\left(E[B]\right)^2 + E[\Lambda]B^{(2)}}{2(1-\rho)} + \frac{E[\Lambda]T(\lambda_0)C(\lambda_0)}{(1-\lambda_0)K}E[U] \\ &\quad + \frac{E[\Lambda]T(\lambda_0)\left(U^{(2)} + C^{(2)} + 2E[U]E[C]\right)}{2K} + \frac{\Lambda^{(2)}E[B]}{2E[\Lambda]}. \end{aligned} \tag{13}$$

3.3 Busy Period

We define a busy cycle R as a time period from the instant in which a busy period Θ completes to the instant in which the next busy period Θ completes.

If there is a batch arrival within the timer length T, R is composed of the close-delay period D and Θ. If there are no batches arriving within T and there is at least one batch arriving within the close-down period C, R is composed of T, C, U, Θ. If there are not any batch arrivals either within T or within C, R is composed of T, C, U, Θ and the idle period I.

Denote by T_U, T_D, T_C, T_Θ, T_I and T_R the actual lengths of U, D, C, Θ, I and R in slots. It is obvious that

$$T_R = T_U + T_D + T_C + T_\Theta + T_I. \tag{14}$$

The event in which T_D equals the timer length T occurs with the probability $T(\lambda_0)$, event that T_D equals a conditional interval time occurring with the probability $1 - T(\lambda_0)$. So the P.G.F. $T_D(z)$ and the mean $E[T_D]$ of T_D are

$$T_D(z) = \frac{T(\lambda_0 z)(1-z) + (1-\lambda_0)z}{1-\lambda_0 z}, \quad E[T_D] = \frac{1-T(\lambda_0)}{1-\lambda_0}. \tag{15}$$

We can also give the means $E[T_U]$, $E[T_C]$ and $E[T_I]$ as follows:

$$E[T_U] = T(\lambda_0)E[U], \quad E[T_C] = T(\lambda_0)E[C], \quad E[T_I] = \frac{1}{1-\lambda_0}T(\lambda_0)C(\lambda_0). \tag{16}$$

Each packet of batches at the beginning of a busy period Θ will introduce a sub-busy period θ. All the sub-busy periods brought by the packets at the beginning of a busy period combine to make a busy period Θ in the system.

The P.G.F. $T_\Theta(z)$ and the mean $E[T_\Theta]$ of T_Θ are given as follows:

$$\begin{aligned} T_\Theta(z) &= (1-T(\lambda_0))\frac{\Lambda(\theta(z))-\lambda_0}{1-\lambda_0} + T(\lambda_0)C(\lambda_0)\frac{\Lambda(\theta(z))-\lambda_0}{1-\lambda_0}U(\Lambda(\theta(z))) \\ &\quad + T(\lambda_0)(C(\Lambda(\theta(z))) - C(\lambda_0))U(\Lambda(\theta(z))), \\ E[T_\Theta] &= K\frac{E[B]}{1-\rho}. \end{aligned} \tag{17}$$

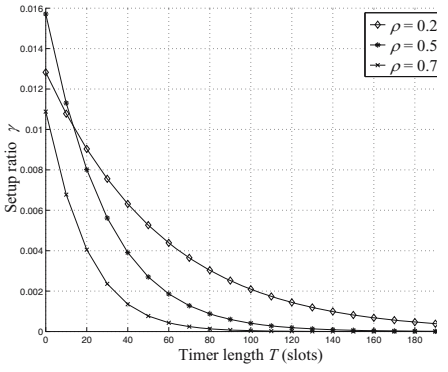


Fig. 1. Setup ratio for different offered load

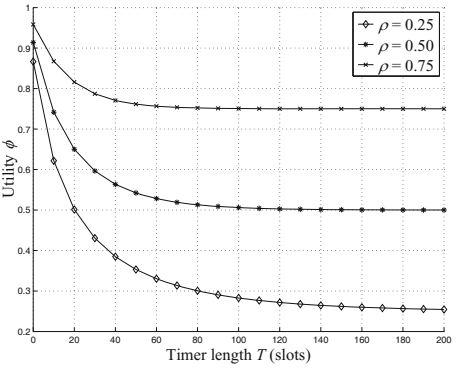


Fig. 2. Utility for different offered load

Concluding Eqs. (16) and (17), we can obtain the mean $E[T_R]$ as $E[T_R] = K/E[\Lambda](1-\rho)$, where K is given in Eq. (9) and $E[\Lambda]$ is given in Eq. (2).

4 Performance Measures and Numerical Results

Setup ratio γ is defined as how many times the system goes to setup period U per slot, which is a measurement of the processing overhead in a connection oriented service. Obviously, the setup ratio γ is given by

$$\gamma = \frac{T(\lambda_0)}{E[T_R]} = \frac{E[\Lambda](1-\rho)T(\lambda_0)}{K}.$$

Utility ϕ of connections is defined as the ratio of the time during which there are packets being transmitted to the time during which the network resource is occupied. Namely, the system is at one of the states of the busy period Θ, close-delay period D, or close-down period C. The utility ϕ is useful for the optimal design of the discrete-time connection oriented service system with a release delay. Clearly, the utility ϕ can be obtained as follows:

$$\phi = \frac{E[T_\Theta]}{E[T_\Theta] + E[T_D] + E[T_C]} = \frac{\rho K}{\rho K + (1-\rho)(1 - T(\lambda_0) + E[\Lambda]T(\lambda_0)E[C])}.$$

In numerical results, we consider that $E[U] = 5$ (slots), $E[C] = 2$ (slots) and $E[B] = 5$ (slots). Considering the burst data shown in Internet traffic, we suppose the batch size Λ to be Pareto(c, α) distributed with $\lambda_k = ck^{-(\alpha+1)}, k = 0, 1, ...,$ where c is a normalization factor for $\sum_{k=1}^{\infty} \lambda_k = 1$, the parameter α is related to the Hurst factor H by $H = (3-\alpha)/2, 0.5 < H < 1, 1 < \alpha < 2$. The smaller the result of α is, the more the burst shows in Internet traffic. Here, we suppose that $H = 0.85$, then $\alpha = 1.3$.

We show numerical results in the Figs. 1 and 2, and discuss the influence of the timer length T on the performance measures for the setup ratio γ and utility

ϕ of connection. In Figs. 1 and 2, the cases with the timer length $T = 0$ show the cases without considering the release delay in the systems as in [4]-[6].

Fig. 1 shows γ as a function of T with three different offered loads $\rho = 0.25, 0.50, 0.75$. We can find that when T increases, γ will decrease quickly to a low level (nearly zero). We also note that the larger the offered load ρ is, the larger the decrease of the setup ratio γ will be.

Fig. 2 shows the utility ϕ as a function of T with three different offered loads $\rho = 0.25, 0.50, 0.75$. When T increases, the utility ϕ will decrease quickly to another fixed value. It can also be found that, for the same timer length T, the larger the offer load ρ is, the larger the utility ϕ will be.

5 Conclusions

One way to reduce the cost of connection oriented service is to reduce the number of connection setups. We proposed a discrete-time connection oriented service system with a release delay in this paper to reduce the cost. We built a batch arrival Geom*/G/1 queue model with a setup/close-delay/close-down strategy to characterize the system operation. We supposed the batch size to be Pareto(c, α) distributed to describe the burst data in Internet traffic. We presented performance measures such as the setup ratio and the utility of connection. We showed that the choice of the timer length T is significant in improving performance.

This paper has potential applications in network design, network maintenance and network management for the next generation Internet.

Acknowledgments. This work was supported in part by MEXT.ORC (2004-2008), Japan and in part by NSFC (No. 10671170) and MADIS, China.

References

1. Yue W. and Matsumoto Y.: Performance Analysis of Multi-Channel and Multi-Traffic on Wireless Communication Networks. Kluwer Academic Publishers (2002)
2. Niu Z. and Takahashi Y.: A Finite Capacity Queue with Exhaustive Vacation/Close-Down/Setup Times and Markovian Arrival Processed. Questa. **31** (1999) 1-23
3. Takagi H.: Queueing Analysis. Vol. 3: Discrete-Time Systems. North-Holland (1993)
4. Tian N. and Zhang G.: Vacation Queueing Models-Theory and Applications. Springer-Verlag (2006)
5. Tian N. and Zhang G.: The Discrete Time GI/Geo/1 Queue With Multiple Vacations. Queueing Systems. **40** (2002) 283-294
6. Attahiru A.: Vacation Models in Discrete Time. Queueing System. **44** (2003) 5-30
7. Jin S. and Tian N.: Performance Evaluation of Virtual Channel Switching System Based on Discrete Time Queue. Journal of China Institute of Communications. **25** (2004) 58-68 (in Chinese)
8. Jin S., Yue W. and Liu M.: Queue Model and Performance Analysis for Discrete Time Switch Virtual Channels Systems. Lecture Notes in Operation Research and its Applications (2005) 26-36

A Knowledge-Based Model Representation and On-Line Solution Method for Dynamic Vehicle Routing Problem

Lijun Sun, Xiangpei Hu, Zheng Wang, and Minfang Huang

School of Management, Dalian University of Technology,
Dalian 116023, P. R. China
{Lijun Sun,yuqier2008}@163.com

Abstract. We propose a knowledge-based model representation and on-line solution method for dynamic vehicle routing problem (DVRP) in the paper in order to realize on-line modeling and solution process of the problem. This knowledge-based model representation is composed of six components—*B* (*B*asic Data collector), *R* (*R*estrictions), *I* (*I*nitial state generator), *S* (*S*tate operator), *G* (*G*oal state), and *C* (*C*ontroller). We term the representation *BRISGC six-component model representation*. Based on the representation, an on-line solution approach to DVRP is presented. And a real-world DVRP in e-Commerce is solved by the representation and solution method in case study. The result proves that the proposed approach is effective for online and real-time vehicle routing.

Keywords: model representation, dynamic vehicle routing problem (DVRP), knowledge, algorithm.

1 Introduction

Vehicle routing problem (VRP) is a NP-hard and key problem in logistics distribution systems. It includes vehicle scheduling and routing. A distribution center that operates routing and scheduling scientifically could promote vehicle's utility, decrease travel distance and travel times, accordingly responds to the clients' demands promptly and increases customer satisfaction. VRP in practice frequently changes because of emergencies and environments, which needs a quick response mechanism to adapt routing process to new states. In order to adapt VRP to a dynamic environment, the model for it should own knowledge-based reasoning ability. Moreover, the final goal of modeling is to facilitate solution process. Therefore, the model should be represented based on knowledge and solution process's characteristics of VRP. So after reviewing literatures from two aspects: *categories of Dynamic Vehicle Routing Problem (DVRP) and their solution methods, knowledge representations for models*, we present an innovative method that combines reasoning knowledge with heuristic solution knowledge to represent a specified DVRP, which contains multiple dynamic changes consisting of three discrete events—new order arrival, old order withdrawnness, and demand change.

The definition of VRP and algorithms for this kind of problems can be found in reference [1]. Here, we unnecessarily waste pages to introduce them. We'll focus our

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 218–226, 2007.
© Springer-Verlag Berlin Heidelberg 2007

discussion on dynamic VRP, which is defined by Larsen [2] as: not all information relevant to the planning of the routes is known by the planner when the routing process begins, and information can change after the initial routes have been constructed. In recent years, with the development of web technology and information technology, dynamic VRP has drawn more attention.

Haghani and Jung have reviewed in reference [3] the existing literatures on dynamic vehicle routing problem. They found that most of the existing researches are concerning problems similar to DVRP, like dynamic traveling salesman problem, dynamic traveling repairman problem, dynamic dial-a-ride problem, and dynamic vehicle allocation problem. Existing researches focusing on pure dynamic vehicle routing can be categorized into three categories according to the different dynamic parameters. One is the *DVRP with Dynamic Travel Times*, whose dynamic parameter is the vehicle's travel time [3]~[6]. Another category is the *DVRP with Dynamic Customer Order Arrival*, in which new orders arrive when existing orders have been scheduled, dynamically changing the optimization problem. Reference [7]~[9] mainly concerns this kind of problems. The last category is the *DVRP with Stochastic Customer Demands*, which is also called stochastic vehicle routing problems. Secomandi [10] studied the applicability of neuro-dynamic programming algorithms to the single-vehicle routing problem with stochastic demands.

The above mentioned DVRP and solution methods for them are all exclusively suitable for a special defined problem which usually contains one parameter's one kind of change. However, in practice, multiple changes usually concur in a same case. How to handle this kind of problem is the mainly concerned topic of the paper.

Knowledge representation for models was firstly brought out by scholars in the field of model management in DSS in 1980s. As DSS is developed to intelligent decision support systems (IDSS), knowledge representation for models becomes a concerned issue in this field. Fay [11] presented a Fuzzy Petri Net notion that combines the graphical power of Petri Nets and the capabilities of Fuzzy Sets to model rule-based expert knowledge to represent models in a train traffic control decision support system. Oldenburg and Marquardt [12] proposed flatness and higher order differential model representations. This method improved the performance of dynamic optimization algorithms employed in on-line applications, and improved the efficiency of on-line decision support. Biswas and Narahari [13], Yeh and Qiao [14] used object-oriented method to represent models. The research of model representation in China started in late 1980s and has become flourishing in recent years. Many representations have been brought out. Hu et al. [15] brought out a six tuple structure M = (I, G, O, T, B, S) to represent the model of discrete dynamic programming. Zhang et al. [16] came up with a model case representation method. Wang and Cai [17] presented an object-oriented reasoning model and its representation method.

Analyzing the development status of model representation, we can find that influenced by the structure of the multi-base system, conventional model representations in DSS usually separate data, model, inference and knowledge from each other and set up their bases and management systems respectively. In fact, in the process of decision making, all the acquired things that can provide help for decision makers, including facts, rules, methods and inference, are all the essential knowledge. Regarding all of

them as knowledge and managing them as a whole can facilitate the concision of the system structure and the consistent processing procedure, which has become the main research trend of knowledge representation in IDSS. We'll adopt the idea in the paper.

We present an innovative method that combines reasoning knowledge with heuristic solution knowledge to represent a specified DVRP, which contains multiple dynamic changes consisting of three discrete events—new order arrival, old order withdrawnness, and demand change. And a solution framework based on the knowledge representation is presented, too.

The remaining paper is organized as follows. Section 2 presents the knowledge representation for DVRP model. Section 3 proposes an on-line solution approach to DVRP based on the presented representation. Section 4 studies a case to demonstrate the approach's effect. The last section draws our conclusions.

2 Knowledge Representation for DVRP Model

2.1 Mathematical Model of VRP

The mathematical model of a VRP is usually an integer model as follows.

$$
\begin{cases}
\min \quad z = \sum_{i=0}^{n} \sum_{j=0}^{n} \sum_{k=1}^{m} c_{ij} x_{ijk} \\
\sum_{i=1}^{n} g_i y_{ki} \le Q \quad k = 1, \cdots, m \\
\sum_{k=1}^{m} y_{ki} = 1 \qquad i = 1, \cdots, n \\
\sum_{i=0}^{n} x_{ijk} = y_{kj} \qquad j = 0,1, \cdots, n;\ k = 1, \cdots, m \\
\sum_{j=0}^{n} x_{ijk} = y_{ki} \qquad i = 0,1, \cdots, n;\ k = 1, \cdots, m \\
X = (x_{ijk}) \in S \qquad (S \text{ is the constraint for the removal of branches}) \\
x_{ijk} = \begin{cases} 1 & \text{vehicle } k \text{ travels from } i \text{ to } j \\ 0 & \text{otherwise} \end{cases} \qquad y_{ki} = \begin{cases} 1 & \text{vehicle } k \text{ visits node } i \\ 0 & \text{otherwise} \end{cases} \qquad y_{k0} = 1
\end{cases}
$$

This mathematical model is complicated and very difficult to be represented in a computer. And its fixed structures cannot be changed easily with the change of problems. What's worse, the pure mathematical symbols cannot support reasoning processes that is needed in a dynamic environment.

2.2 BRISGC Six-Component Representation

Generally, the solution process of heuristic algorithms for VRP needs four elements: initial state, neighborhood function, evaluation function, and goal state. An initial state is the necessity of the solution process of VRP, since every improving algorithm and meta-heuristic algorithm must start work on an initial solution. Every algorithm has its own neighborhood function, which is defined by specific algorithm, guiding the solution process to generate new neighborhood solutions of the current solutions. An evaluation function is also defined by specific algorithm. Different algorithms have different evaluation functions guiding the solution process to select better solutions in

current solution's neighborhood. A goal state, where the solution process must stop at, is preset by a real-world problem or founded by an algorithm.

Besides the four elements, there is another important element relevant to the working process of a heuristic algorithm for VRP. When a heuristic algorithm constructs or selects a state, it will test if the restrictions can be satisfied. Thus, restrictions should be separately represented. Moreover, an exclusive component is needed to dynamically collect basic information because of the unstable customer cluster in the on-line e-commerce. Finally, a coordinator is needed. Accordant with Du et al.'s conclusions [18], we believe that different phase of a solution process for DVRP should use different kinds of algorithms. A coordinator is needed to make these algorithms well cooperate with each other.

From the above analysis, we represent model of DVRP as follows:

M_{DVRP} = (B, R, I, S, G, C). *B* is *B*asic data collector, a processor behind the man-machine interface. It is in charge of select route information associated with current planned customers from the digital map, and collect customers' demands. After collecting needed information, this component will transfer the information into a database. *R* is *R*estrictions that mainly contain the resource limit of the third logistics center, like vehicle's capacity limit, drivers' schedule limit, etc. *I* is *I*nitial state generator, which contains several constructive heuristic algorithms. *S* is *S*tate operator, which consists of neighborhood function and evaluation function. This component mainly concerns how to generate new states and which state should be chosen as the next solution. Since the two functions are defined by specific improving heuristic algorithm and vice versa, the component accommodates different kind of improving heuristic algorithms. *G* is *G*oal state defining the stopping rules for improving algorithms' repeat process. *C* is *C*ontroller, a cooperative coordinator of the above five component. It has two functions. One is to connect the five components. The other is to dynamically select appropriate algorithm according to different problem goal and different phases of solution process. This component is realized by rules and it is the brain of solution process for DVRP. Because of the page limit, the detailed rules will be narrated in another paper. We term the representation *BRISGC Six-component representation* in the paper.

3 On-Line Solution Method for Dynamic Vehicle Routing Problem

3.1 Structure of the Knowledgebase

The relation among BRISGC six components and the structure of the knowledgebase composed of them is shown in figure 1.

In the knowledgebase, fact base stores the information from orders, including demand, customer information, and rout information associated with the customer, and information from the distribution center, like vehicle information, schedule information etc.. The rule base stores the modeling and solution rules for problem-solving process, including restrictions that the process should obey, constructive rules for initial routes,

rules for improving routes and process stopping rules. The dynamic knowledgebase is a special store to deposit interim solutions and temporally to store the final solutions.

3.2 Problem-Solving Process Based on BRISGC Six-Component Knowledge Representation

The solution approach to DVRP based on BRISGC Six-component representation is shown in figure 2.

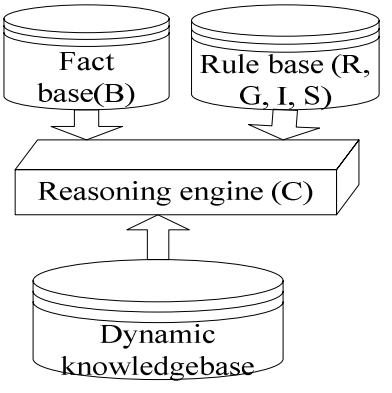


Fig. 1. Structure of the knowledgebase for BRISGC model representation of DVRP

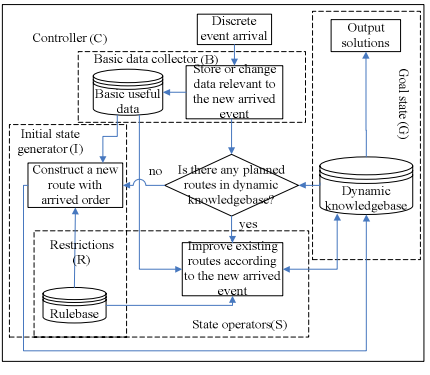


Fig. 2. Problem-solving process of DVRP based on BRISGC Six-component knowledge representation

The flowchart in real line rectangle indicates the process that the BRISGC six-component representation solves the on-line DVRP. The five dashed rectangles represent the working realm of the other five components respectively. The whole process is realized under the control of Component C. The module "discrete event arrival" in the figure contains three kind of event—new order arrival, old order withdrawnness, and demand change. Component C can adopt different rules to adjust current solutions. For example, if suddenly a customer in one of routes in planned distribution scheme withdraws its order, component C will use "Delete Demand" rules, these rules are associated with how to rearrange customers in routes in existing distribution scheme after delete some customers. Details of some other rules, including "Add Demand", "Change Demand", aren't given in the paper. We will exclusively draft another paper on the controller to narrate all rules. The arrows in the figure indicate the sequence when these components work and their transferring relations. The next part demonstrates the solution process in detail in a real-world case.

4 Case Study

A gas sale company in Dalian city sells several types of gas, including 97#, 93#, and 90#. According to historical sales experience, the company prepares vehicles with

different capacities respectively for delivering different kinds of gas, 8t capacity vehicle in charge of 97# gas, 10t in charge of 93#, and 16t in charge of 90#. Most of gas stations in the city adopt just-in-time policy, so they place orders through the gas company's website before 2-3 hours of their gas being sold up. In order to fulfill orders timely, the company process orders once an hour. However, after a batch of orders within an hour has been processed, some orders might be withdrawn because of unexpected market, or certain ordering quantity is changed. Therefore, a dynamic routing system which can promptly adjust distribution schemes to these emergencies is needed. Orders received on 11/5/2006 from 8:00a.m. to 9:00a.m. are shown in table 1.

Working process of the six components is as follows:

B component: the component will firstly abstract route information associated with the above mentioned 13 gas stations and the company's distribution center (results shown in table 2, the distribution center is represented by 0). And then, it aggregates demand information, including demand quantity of every type of gas, desired delivery time, etc. and transfers this information to basic database.

Table 1. Orders Received from Website of the Gas Company on 11/5/2006 from 8:00am. to 9:00am

Gas Station Id	Gas Station Name	Gas Type	Quantity	[illegible]
1	Yinghua	90#	3	10:00
1	Yinghua	97#	6	10:00
3	Hanlin	90#	9	10:30
4	Qingming	90#	8	10:30
4	Huangming	93#	8	10:30
5	Taishang	90#	7	10:30
5	Taishang	97#	2	10:30
9	Suiyuan	90#	9	10:00
9	Suiyuan	97#	5	10:00
7	Lvshan	90#	9	10:30
8	Dafu	97#	2	10:00
9	Xinghaiwan	90#	4	10:00
9	Xinghaiwan	93#	6	10:00
10	Huaxia	93#	8	10:00
11	Tianying	90#	3	10:30
11	Tianying	93#	4	10:30
12	Tianlu	90#	9	10:00
13	Heishijiao	90#	3	10:30
15	Tchai	93#	3	10:00

Table 2. Route information (symmetrical routes)

nodes \ nodes	0	1	3	4	5	6	7	8	9	10	11	12	13	15
0	0	9	14	9	16	6	14	7	8	8	17	9	6	15
1		0	7	13	12	7	8	6	12	9	10	15	6	13
3			0	5	7	7	12	17	9	9	7	8	8	6
4				0	7	13	10	11	8	7	5	13	16	15
5					0	5	13	10	16	13	15	14	11	5
6						0	11	13	6	14	11	15	14	14
7							0	16	16	16	16	17	15	15
8								0	8	9	16	15	15	5
9									0	10	11	9	17	7
10										0	9	7	14	14
11											0	7	13	7
12												0	11	13
13													0	15
15														0

R component: the content of the component in this case consists of capacities of different kinds of vehicle.

I component: after C component examines tradeoffs between basic information, it will command I component to use an appropriate constructive heuristic algorithm to generate initial routes under the constraint of R component. In this case, the component will use the Time-Oriented, Nearest-Neighbor heuristic algorithm to generate routes (consult reference [19] for details of the algorithm) according to commands from C component. The initial routes are as follows:

For delivering 90# gas:

Route 90#1: $0 \xrightarrow{6} 13 \xrightarrow{8} 3 \xrightarrow{7} 11 \xrightarrow{17} 0$, total distance (or travel time) : 38, total load : 15

Route 90#2: $0 \xrightarrow{6} 6 \xrightarrow{13} 4 \xrightarrow{9} 0$, total distance (or travel time) : 28, total load : 15

Route 90#3: $0 \xrightarrow{8} 9 \xrightarrow{9} 12 \xrightarrow{15} 1 \xrightarrow{9} 0$, total distance (or travel time) : 41, total load : 16

Route 90#4: $0 \xrightarrow{14} 7 \xrightarrow{13} 5 \xrightarrow{16} 0$, total distance (or travel time) : 43, total load : 16

For delivering 93# gas:

Route 93#1: $0 \xrightarrow{8} 9 \xrightarrow{11} 11 \xrightarrow{17} 0$, total distance (or travel time) : 36, total load : 10

Route 93#2: $0 \xrightarrow{8} 10 \xrightarrow{14} 15 \xrightarrow{15} 0$, total distance (or travel time) : 37, total load : 9

Route 93#3: $0 \xrightarrow{9} 4 \xrightarrow{9} 0$, total distance (or travel time) : 18, total load : 8

For delivering 97# gas:

Route 97#1: $0 \xrightarrow{6} 6 \xrightarrow{5} 5 \xrightarrow{16} 0$, total distance (or travel time) : 27, total load : 7

Route 97#2: $0 \xrightarrow{7} 8 \xrightarrow{6} 1 \xrightarrow{9} 0$, total distance (or travel time) : 22, total load : 8

G component: the component will use time as the stop rule of improving process of S component according to the command from C component. In this case, the time is set to 30 seconds. That is the total work time of S component is 30 seconds.

S component: the 2-swap and 2-exchange algorithms in the component respectively improve initial routes within and between routes within 30 seconds under the constraint of R component. After 30 seconds, the initial routes are improved. Route 90#3 has been changed to Route 90#3':

Route 90#3': $0 \xrightarrow{9} 1 \xrightarrow{12} 9 \xrightarrow{9} 12 \xrightarrow{9} 0$, total distance (or travel time) : 39, total load : 16

and other routes maintain original state. And after the goal is achieved, G component transfers the result to C component.

C component: the first function of the component is examining characteristics of the problem in order to command other components as we have mentioned in I and G components' work processes. The other function is commanding dynamic knowledgebase to hold the result according to G component for 5 minutes (the time is preset according to system characteristics). If within the 5 minutes there aren't changes, C component will command the knowledgebase to output the result. Otherwise, if some events occur within the 5 minutes, then C component will take appropriate actions to adjust the result.

For example, in the case, if gas station 3 additionally places an order at 9:02 a.m., which order 1t 93# gas, the component will trigger "Add Demand" rule module. The module will select the most appropriate insertion algorithm (in the case, it selected C-W savings algorithm) to insert the demand into the most suitable route:

Route 93#2: $0 \xrightarrow{8} 10 \xrightarrow{14} 15 \xrightarrow{15} 0$, total distance (or travel time) : 37, total load : 9

and get Route 93#2':

Route 93#2': $0 \xrightarrow{8} 10 \xrightarrow{14} 15 \xrightarrow{6} 3 \xrightarrow{14} 0$, total distance (or travel time) : 42, total load : 10.

After that, if there are no other emergencies within the 5 minutes, G component will output final schemes consist of Route 90#1, Route 90#2, Route 90#3', Route 90#4, Route 93#1, Route 93#2', Route 93#3, Route 97#1, Route 97#2.

5 Conclusions

In this paper, we propose a knowledge-based model representation—BRISGC six-component model representation for DVRP and a problem-solving approach to it based on deep analysis of existing solutions to VRP. The structure of the representation is flexible enough to support on-line and real-time modeling and solution process. Moreover, the representation can make fully use of conventional algorithms and extend their application field from static VRP to dynamic VRP, and support on-line and real-time modeling and solution process.

Future work includes how to adapt the representation structure to integrate meta-heuristic algorithms and to solve DVRP with more parameters' more changes.

Acknowledgement. This work is supported by National Natural Science Foundation of China (Grant No. 70571009, 70171040 and 70031020 (key project)), Key Project of the Chinese Ministry of Education (No. 03052), and Ph.D. Program Foundation of Ministry of Education of China (No. 20060141013).

References

1. Sun, L. J., Hu, X. P., Wang, Z.: Reviews on Vehicle Routing Problem and Its Solution Methods. Systems Engineering, no.11 (2006) 35-41
2. Larsen, A.: The Dynamic Vehicle Routing Problem. PhD Thesis, IMM-PHD-2000-73, ISSN 0909-3192, LYNGBY (2001) 5
3. Haghani, A. and Jung, S.: A dynamic vehicle routing problem with time-dependent travel times. Computers & Operations Research 32 (2005) 2959-2986
4. Taniguchi, E. and Shimamoto, H.: Intelligent transportation system based dynamic vehicle routing and scheduling with variable travel times, Transportation Research Part C 12 (2004) 235-250
5. Ichoua, S., Gendreau, M., Potvin, JY.: Vehicle dispatching with time dependent travel times. European Journal of Operations Research 144 (2003) 379–96
6. Potvina, J. Y., Xua, Y., Benyahia, I.: A dynamic vehicle routing problem with time-dependent travel times. Computers & Operations Research 32 (2005) 2959-2986
7. Fleischmann, B., Gnutzmann, S., Sandvoß, E.: Dynamic Vehicle Routing Based on Online Traffic Information, Transportation science, Vol. 38, No. 4 (2004) 420-433
8. Branke, J., Middendorf, M., Noeth, G., Dessouky, M.: Waiting Strategies for Dynamic Vehicle Routing, Transportation science, Vol. 39, No. 3 (2005) 298-312
9. Montemanni, R., Gambardella, L. M., Rizzoli, A. E., Donati, A. V.: Ant Colony System for a Dynamic Vehicle Routing Problem, Journal of Combinatorial Optimization, Vol. 10, No. 4 (2005) 327–343
10. Secomandi, N.: Comparing neuro-dynamic programming algorithms for the vehicle routing problem with stochastic demands. Computers & Operations Research 27 (2000) 1201-1225
11. Fay, A.: A fuzzy knowledge-based system for railway traffic control. Engineering Applications of Artificial Intelligence 13 (2000) 719-729
12. Oldenburg, J., and Marquardt, W.: Flatness and higher order differential model representations in dynamic optimization. Computers and Chemical Engineering 26 (2002) 385-400

13. Biswas, S., and Narahari, Y: Object oriented modeling and decision support for supply chains. European Journal of Operational Research 153 (2004) 704-726
14. Yeh, A. G., and Qiao, J. J.: ModelObjects—a model management component for the development of planning support systems. Computers, Environment and Urban Systems 29 (2005) 133-157
15. Hu, X. P., Xu, Z. C., and Yang, D. L.: Intelligent operations research and real-time optimal control for dynamic systems, Journal of Management Sciences In China, vol. 5, no. 4 (2002) 13-21
16. Zhang, J., Pan, Q. S., and Hu, Y. Q.: Model Case Representation. Journal of Management Science in China, vol. 3, no. 2 (2000) 90-94
17. Wang, H. Y. and Cai, R. Y.: An Object-oriented Inference Model and Its Knowledge Representation. Journal of Nanjing University of Technology, vol. 24, no. 3 (2002) 20-24
18. Du, T. C., Li, E. Y., and Chou, D: Dynamic vehicle routing for online B2C delivery. Omega 33 (2005) 33-45
19. Solomon, M. M.: Algorithms for the vehicle routing and scheduling problems with time window constrains. Operations Research 35 (1987) 254-265

Numerical Simulation of Static Noise Margin for a Six-Transistor Static Random Access Memory Cell with 32nm Fin-Typed Field Effect Transistors

Yiming Li[1], Chih-Hong Hwang[1], and Shao-Ming Yu[2]

[1] Department of Communication Engineering, National Chiao Tung University, Hsinchu 300, Taiwan
ymli@faculty.nctu.edu.tw, chhwang@mail.ymlab.org
[2] Department of Computer Science, National Chiao Tung University, Hsinchu 300, Taiwan
smyu@mail.ymlab.org

Abstract. We in this paper for the first time explore the static noise margin (SNM) of a six-transistor (6T) static random access memory (SRAM) cell with nanoscale silicon-on-insulator (SOI) fin-typed field effect transistors (FinFETs). The SNM is calculated with respect to the supply voltage, operating temperature, and cell ratio by performing a three-dimensional mixed-mode simulation. To include the quantum mechanical effect, the density-gradient equation is simultaneously solved in the coupled device and circuit equations. The standard deviation (σ_{SNM}) of SNM versus device's channel length is computed, based upon the design of experiment and response surface methodology. Compared with the result of SNM for SRAM with 32nm planar metal-oxide-semiconductor field effect transistors, SRAM with SOI FinFETs quantitatively exhibits higher SNM and lower σ_{SNM}. Improvement of characteristics resulting from good channel controllability implies that SRAM cells fabricated with FinFETs continuously maintains cell stability in sub-32nm technology nodes.

Keywords: FinFET, SRAM, modeling and simulation, computational statistics.

1 Introduction

Silicon-based metal-oxide-semiconductor field effect transistors (MOSFETs) have been building blocks for SRAM cells [1]. For 45nm fabrication technologies, planar MOSFETs encounter significant challenges to device performance and circuit stability [2,3,4]. The SRAM with diverse device structures, such as the thin-buried-oxide SOI MOSFETs and SOI FinFETs have been of great interest [5,6,7,8] due to good suppression of short-channel effects and high area's efficiency [9,10,11,12,13,14]. Study of SNM for SRAM with the 32nm planar MOSFETs and SOI FinFETs helps technological development.

In this paper, SNM of a 6T SRAM cell, shown in Fig. 1a, is examined with a mixed-mode simulation. The three-dimensional (3D) device simulation is performed to calculate device current-voltage ($I-V$) characteristics by solving a set of density-gradient drift-diffusion equations [13, 14] in the solution of coupled device and circuit equations [15,16,17]. For SRAM with the 32nm planar MOSFETs and SOI FinFETs, shown in the insets of Fig.1b, SNM is explored and compared. SNM is the minimum DC-voltage disturbance necessary to upset the state of SRAM cell [2]. It is quantified by the

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 227–234, 2007.
© Springer-Verlag Berlin Heidelberg 2007

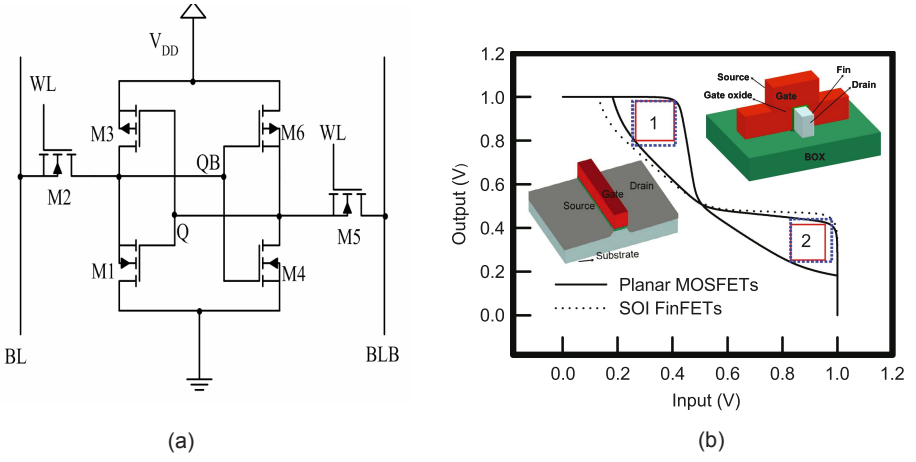


Fig. 1. (a) A circuit diagram of the simulated 6T SRAM, where the N- and P-typed devices are with the 32nm conventional planar MOSFETs and SOI FinFETs.(b) An illustration of the transfer characteristics of the SRAM with planar MOSFETs (the left inset plot) and SOI FinFETs (the right inset one). The edge of the maximum square is right the SNM. The blue squares are from SOI FinFETs and the red squares are from planar MOSFETs.

length of the side of the maximum square that can fit inside the butterfly curves formed by the cross-coupled inverters. The dependence of SNM, during the modes of hold and read, on the supply voltage, operating temperature, and cell ratio is examined. To explore the sensitivity of SNM, the standard deviation with respect to the device's channel length for the SRAM under the read mode is computed. The calculation is based on the computational statistics methodology [18], which integrates the design of experiment (DOE), a second-order response surface model (RSM), and the simulation of mixed-mode.

Simulation results imply that the SRAM with the 32nm SOI FinFETs achieves dramatic improvements in the cell stability. For the mode of read, where the word line (WL) is biased at V_{DD}, the SNM of SRAM with SOI FinFETs has 20% increase, compared with the SRAM with planar MOSFETs. Improvement is evident when the cell ratio increases. The cell ratio means that the ratio of the widths of the pull-down transistor to the access transistor, shown in Fig. 1a. For the model of hold, the WL is equal to 0V, and a 6% improvement of SNM is observed for the SRAM with SOI FinFETs. Increase of the operating temperature leads to a decrease of SNM due to a reduction of the threshold voltage [1, 2], but the SRAM with SOI FinFET exhibits that the SNM is less temperature dependence, compared with the SRAM using planar MOSFETs. Considering the Gaussian distribution to the variation of the device channel length, more than 2-times reduction of the standard deviation of SNM for the SRAM with SOI FinFETs is achieved, compared with the SRAM using planar MOSFETs.

This paper is organized as follows. In Sec. 2, we briefly state the simulation methodology. In Sec. 3, we report and discuss the simulation results. Finally, we draw the conclusions and suggest future work.

2 Simulation Methodology

To assess the effect of device structure on the SNM of a 6T SRAM cell, a mixed-mode simulation is directly adopted due to lack of industrial standard equivalent circuit model [19] for nanoscale SOI FinFETs. The coupled semiconductor device equations and circuit equations are iteratively solved for device electrical characteristics and SRAM's transfer characteristics. The results are used in the calculation of SNM for the SRAM with different devices. To consider the quantum mechanical effects in the mixed-mode simulation, 3D density-gradient is solved together with the drift-diffusion transport model for electrical characteristics of planar MOSFETs and SOI FinFETs [13, 14, 15, 17, 20]. The computed device characteristics are connected to the SRAM's circuit simulation. The formulation of SRAM's circuit equations is mainly based upon the Kirchhoff's current conservation law [21], shown in Fig. 1a. The ordinary differential equations are solved to estimate the nodal voltage and loop current, and then the DC transfer characteristics of SRAM are systematically computed with respect to different biasing condition, cell ratio, and operating temperature.

With the simulation of mixed-mode, the computational statistics methodology [18,22] is advanced to investigate the sensitivity of the SNM versus the device's channel length of the 6T SRAM cell. According to a face-centered-cube DOE, the simulation of mixed-mode is performed, and the simulation results are used in construction of the second-order RSM model of the SNM. The constructed RSM allows us to quickly analyze the sensitivity of SNM by assuming a proper distribution, such as a Gaussian distribution to the device's channel length.

3 Results and Discussion

For the 32nm planar MOSFETs, we assume that the device width is equal to 50nm, the thickness of gate oxide is 1.5nm, the substrate doping concentration is $10^{18}cm^{-3}$ uniformly, and the source/drain doping is $10^{20}cm^{-3}$. To adjust device performance, compared with the 32nm SOI FinFET, metal gate material with 4.7eV workfunction is selected in the 32nm planar MOSFET. For the 32nm SOI FinFETs, the fin height is 20nm, and the device width is 10nm. The source/drain doping is the same with the one of the planar device. To consider a lightly doping channel, the substrate doping concentration of SOI FinFET is $3 \times 10^{16}cm^{-3}$. To eliminate the short channel effects, we note that the setting on SOI FinFET's silicon fin thickness and the channel length has resulted in an optimal geometry aspect ratio (fin thickness/channel length) [6, 7, 9, 10, 11, 12, 13, 14]. A nearly undoped channel will reduce the effect of random dopant on the device performance [23]. Our 3D device simulation shows that SOI FinFETs not only have high ratio of the on- and off-state currents, but also possess a lower drain induced barrier height lowering (DIBL=$39.88mV$) and stable subthreshold swing (SS=$70.73mV/Dec$), compared with the characteristics of planar MOSFET (DIBL=$95.19mV$ and SS=$80.85mV/Dec$).

By mirroring the inverter characteristics of transfer curves of input and output voltages, butterfly curves of the 6T SRAM operated under the read mode (the WL is biased at V_{DD}.), shown in Fig. 1b, are numerically computed. In this calculation, SRAM is with two different devices, where the $V_{DD} = 1.0V$, the operating temperature = 300K, and the

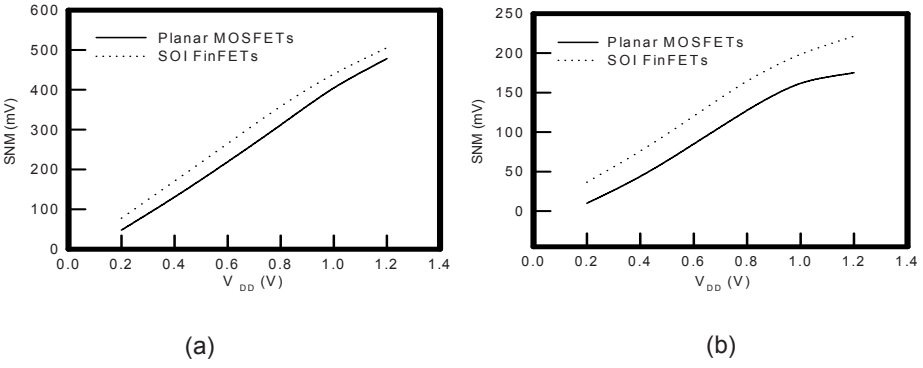


Fig. 2. SNM versus V_{DD}, under the (a) hold and (b) read modes, for SRAM with two device structures, where the cell ratio = 1

cell ratio = 1. We calculate the SNM with butterfly curves by finding an edge of a maximum square. The maximum square is given by the relation, Min{square 1, square 2}, where the square 1 and the square 2 are the maximum squares locating in the upper and lower parts of the butterfly curves, respectively. The red and blue squares, shown in Fig. 1b, are two maximum squares for the SRAM using different devices. The SRAM with SOI FinFETs achieves an improvement in SNM, compared with the result of SRAM with planar MOSFETs. Similarly, we have computed SNM for the hold mode, where the WL is 0V. To explore the characteristics of SRAM's SNM, variations of SNM with respect to different parameters are further examined. We note that increase of the operating temperature and reduction of V_{DD} will induce a significant degradation of SNM in the SRAM cell. Figures 2a and 2b show SNM varies with V_{DD} for both the hold and read modes, where the cell ratio = 1. Comparison between the results of 6T SRAM with 32nm planar MOSFETs and SOI FinFETs shows that SRAM with the latter building blocks possesses much better stability of SNM; in particular, for high supply voltage.

Under the hold mode, the SNM versus supply voltage for the SRAM with the planar MOSFETs and SOI FinFETs, is shown in Figs. 3a and 4a, respectively. The SNM slightly is reduced when the cell ratio is increased. However, the change is insignificant due to an off-state operation of the transistors M2 and M5. The turned-off M2 and M5 keep the symmetry properties of the upper and lower squares in the butterfly curve and then maintain the original states at the nodes Q and QB. Therefore, an increase of the cell ratio does not significantly alter SNM for the SRAM under the hold mode. Under the read mode, for the SRAM with the planar MOSFETs and SOI FinFETs, SNM is depicted in Figs. 3b and 4b, respectively. The SNM is increased when the cell ratio is increased under the read mode. It is a direct result due to high current resulting from the high cell ratio in the transistor M1, where the node Q is assumed at the logic "1" and the node QB is at the logic "0". The difference of SNM increases when the cell ratio is increased, for example, when the $V_{DD} = 1.0V$, the difference of SNM is about 100mV for the SRAM using SOI FinFETs and planar MOSFETs, respectively. The large difference comes from the nature of the device with triple-gate structure; in particular, for the driving transistors of SRAM cell.

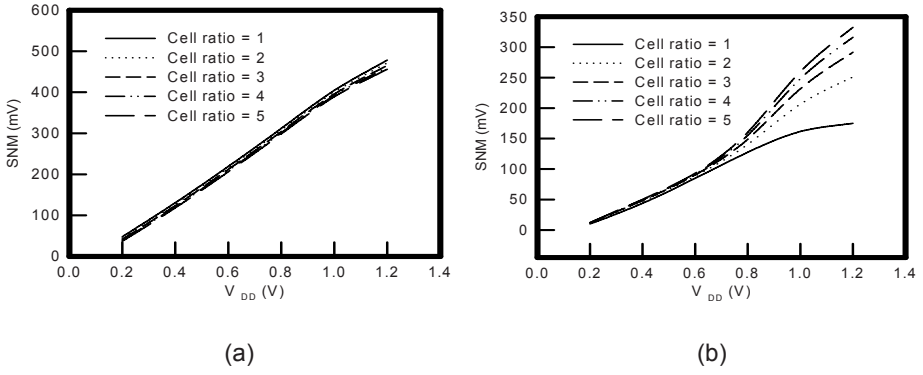


Fig. 3. The SNM versus V_{DD} with respect to different cell ratios, under the (a) hold and (b) read modes, for the SRAM with 32nm conventional planar MOSFETs

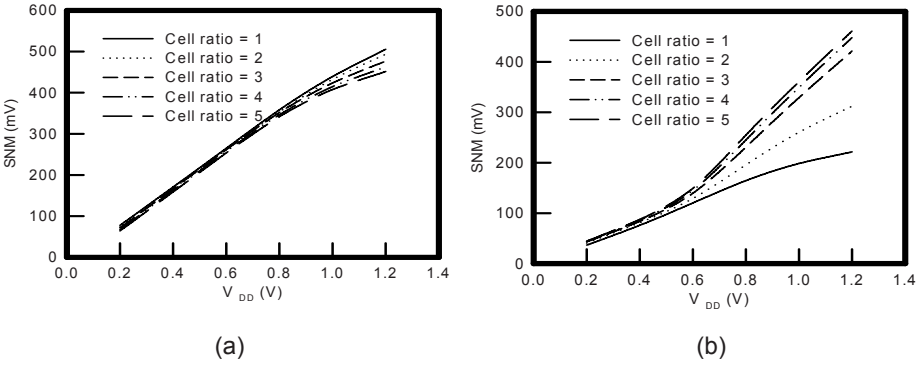


Fig. 4. The SNM versus V_{DD} with respect to different cell ratios, under the (a) hold and (b) read modes, for the SRAM with 32nm SOI FinFETs

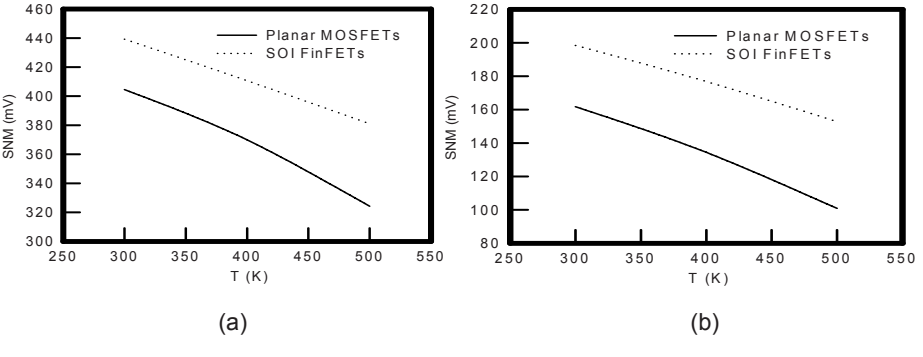


Fig. 5. The SNM versus the temperature for the SRAM with two device structures under the (a) hold and (b) read modes, where the supply voltage is equal to 1.0V

The SNM versus different operating temperature for the SRAM under the modes of hold and read are shown in Figs. 5a and 5b, respectively. The SNM, under the modes of hold and read, decreases when the temperature increases; it is attributed to decrease of the threshold voltage when the temperature increases. According to our calculation, the variation of threshold voltage versus temperature is almost the same between the planar MOSFETs and SOI FinFETs; nevertheless, the SNM of SRAM with 32nm SOI FinFETs is still less dependent upon the temperature (dotted-line in Fig. 5b) due to good channel controllability and improved short channel effects, compared with the SRAM using planar MOSFETs, in particular, for the SRAM under the read mode.

The sensitivity of SNM versus the channel length of transistors in the 6T SRAM cell is further investigated. We first model a second-order RSM for the SNM of 6T SRAM cell during the read mode, where the supply voltage is 1.0V at the room temperature. Without loss of generality, we consider only the six devices' channel lengths. The RSM of SNM is in terms of the channel lengths of transistors M1, M4, and others including M2, M3, M5, and M6. We note the node Q is assumed to have a logic "0" and the node QB is with a logic "1", shown in Fig. 1a. The constructed RSM for the 6T SRAM using 32nm conventional planar MOSFETs and SOI FinFETs are given by:

$$SNM = 162.21 + 26.91A - 17.89B + 13.79C - 8.77A^2 - 3.98B^2 + 0.15C^2 - 0.095AC, \quad (1)$$

and

$$SNM = 198.64 + 14.45A - 9.61B + 7.41C - 4.71A^2 - 2.14B^2 + 0.081C^2 - 0.051AC, \quad (2)$$

respectively, where A, B, and C are the channel lengths of M1, M4, and others, respectively. For the same model accuracy of the SNM of the 6T SRAM cell, the approach of RSM is computationally cost-effective, compared with an empirically fitted model. The Eqs. (1) and (2) suggest that the M1's channel length will result in a large impact on SNM. Therefore, we preliminarily explore the sensitivity of SNM versus the channel length of M1, shown in Fig. 6. Simultaneously assuming Gaussian distributions

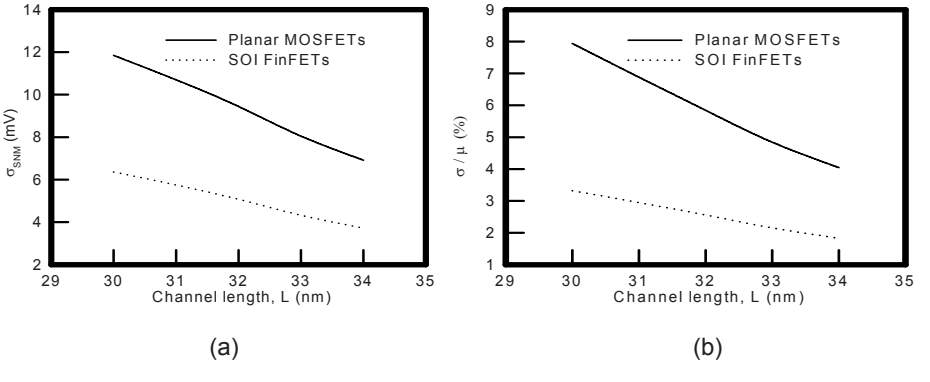


Fig. 6. (a) The standard deviation of SNM versus the M1's channel length for the SRAM using two device structures. (b) The normalized standard deviation of SNM versus the M1's channel length, where μ is the nominal value of SNM.

on A, B, and C, where the 3-sigma = 3nm is assumed for each nominal value. The standard deviations of the SNM versus the channel length for the SRAM with 32nm conventional planar MOSFETs (solid-line) and SOI FinFETs (dotted-line) are shown in Fig. 6a, where the cell ratio is = 1. For the SRAM with conventional planar MOSFETs, σ_{SNM} is larger than that of SOI FinFETs due to a serious short channel effects in planar MOSFETs. The plot of σ_{SNM}/mu_{SNM} versus the M1's channel length is depicted in Fig. 6b. We notice that the difference of the ratio (σ/μ) becomes large if a uniform distribution for the device's channel length is assumed.

4 Conclusions

We have explored the SNM and its sensitivity for a 6T SRAM cell with the 32nm SOI FinFETs. Compared with the SRAM with conventional planar MOSFETs, the SRAM with SOI FinFETs possesses much better SNM and more stable variation of SNM under the modes of read and hold. The improvement of the cell stability attributes to good channel controllability and suppression of the short channel effects when the SRAM with SOI FinFETs. For more accurate estimation of the temperature effect on the SNM, the density-gradient equation could be solved with hydrodynamic model or Boltzmann transport equation instead of the drift-diffusion model. The SNM and its sensitivity can be subject to further investigation for the SRAM with other device structures, such as the bulk FinFET and nanowire FinFET. Impact of device parameters including line edge roughness and random doping on the fluctuation of SNM should also be considered in future works. We are currently designing device for the SRAM fabrication and the SNM comparison between theoretical estimation and experimental measurement.

Acknowledgments

This work was supported in part by Taiwan National Science Council (NSC) under Contract NSC-95-2221-E-009-336 and Contract NSC-95-2752-E-009-003-PAE, by MoE ATU Program, Taiwan, under a 2006-2007 grant, and by the Taiwan Semiconductor Manufacturing Company under a 2006-2007 grant.

References

1. Sze, S. M.: Physics of Semiconductor Devices. New York:Wiley. (1981)
2. Seevinck, E., List, F.J., and Lohstroh, J.: Static-noise margin analysis of MOS SRAM cells. IEEE J. Solid-State Circuits. **22**(5) (1987) 748–754
3. Wakabayashi, H., Yamagami, S., Ikezawa, N., Ogura A., Narihiro, M., Arai, K., Ochiai, Y., Takeuchi, K., Yamamoto, T., and Mogami, T.:Sub-10-nm planar-bulk-CMOS devices using lateral junction control. Int. Electron Devices Meeting Tech. Dig. (2003) 20.7.1–20.7.3
4. Fischetti, M. V.: Scaling MOSFET's to the limit: A physicists's perspective. J. Comput. Electronics. **2**(2-4) (2003) 73–79
5. Ananthan, H., Bansal, A., and Roy, K.: FinFET SRAM - device and circuit design considerations. Proc. 5th Int. Symp. Quality Electronic Design. (2004) 511–516

6. Park, T., Cho, H.J., Choe, J.D., Han, S.Y., Jung, S.-M., Jeong, J.H., Nam, B.Y., Kwon, O.I., Han, J.N., Kang, H.S., Chae, M.C., Yeo, G.S., Lee, S.W., Lee, D.Y., Park, D., Kim, K., Yoon, E., and Lee, J.H.: Static noise margin of the full DG-CMOS SRAM cell using bulk FinFETs (Omega MOSFETs). Int. Electron Devices Meeting Tech. Dig. (2003) 2.2.1–2.2.4
7. Chen, H.-Y., Chang C.-Y., Huang, C.-C., Chung, T.-X., Liu, S.-D., Hwang, J.-R., Liu, Y.-H., Chou, Y.-J., Wu, H.-J., Shu, K.-C., Huang, C.-K., You, J.-W., Shin, J.-J., Chen, C.-K., Lin, C.-H., Hsu, J.-W., Perng, B.-C., Tsai, P.-Y., Chen, C.-C., Shieh, J.-H., Tao, H.-J., Chen, S.-C., Gau, T.-S., and Yang F.-L.: Novel 20nm hybrid SOI/bulk CMOS technology with 0.183 m2 6T-SRAM cell by immersion lithography. VLSI Technology Symp. Tech. Dig.. (2005) 16–17
8. Samsudin, K., Cheng, B., Brown, A. R., Roy, S., and Asenov, A.: UTB SOI SRAM Cell Stability Under the Influence of Intrinsic Parameter Fluctuation. Proc. European Solid-State Device Research Conf. (2005) 553–556
9. Xiong, S. and Bokor, J.: Sensitivity of double-gate and FinFET Devices to process variations. IEEE Trans. Electron Devices. **50**(11) (2003) 2255–2261
10. Park, J.-T. and Colinge, J.-P.: Multiple-gate SOI MOSFETs: device design guidelines. IEEE Trans. Electron Devices. **49**(12) (2002) 2222–2229
11. Yang, F.-L., Chen, H.-Y., Chen, F.-C., Huang, C.-C., Chang, C.-Y., Chiu, H.-K., Lee, C.-C., Chen, C.-C., Huang, H.-T., Chen, C.-J., Tao, H.-J., Yeo, Y.-C. , Liang, M.-S., and Hu, C.: 25 nm CMOS Omega FETs. J Int. Electron Devices Meeting Tech. Dig. (2002) 255–258
12. Li, Y. and Yu, S.-M.:Comparison of Threshold Voltage Fluctuations in Sub-45 nm Planar MOSFET and Thin-Buried-Oxide SOI Devices. Extended Abstracts Int. Solid State Devices and Materials Conf. (2006) 370–371
13. Li, Y., Chou, H.-M., and Lee, J.-W.: Investigation of Electrical Characteristics on Surrounding-Gate and Omega-Shaped-Gate Nanowire FinFETs. IEEE Trans. Nanotechnology. **4**(5) (2005) 510–516
14. Li, Y.and Chou, H.-M.: AA Comparative Study of Electrical Characteristic on Sub-10 nm Double Gate MOSFETs. IEEE Trans. Nanotechnology. **4**(5) (2005) 645–647
15. Palankovski, V., Belova, N., Grasser, T., Puchner, H., Aronowitz, S., and Selberherr, S.: A methodology for deep sub-0.25 gm CMOS technology prediction. IEEE Trans. Electron Devices. **48**(10) (2001) 2331–2336
16. Grasser, T. and Selberherr, S.: Mixed-mode device simulation. Proc. 22nd Int. Conf. Microelectronics. **1** (2000) 35–42
17. Binder, T., Heitzinger, C., and Selberherr, S.: A study on global and local optimization techniques for TCAD analysis tasks. IEEE Trans. Computer-Aided Design of Integrated Circuits and Systems. **23**(6) (2004) 814–822
18. Li, Y. and Chou, Y.-S.:A Novel Statistical Methodology for Sub-100 nm MOSFET Fabrication Optimization and Sensitivity Analysis. Extended Abstracts Int. Solid State Devices and Materials Conf. (2005) 622–623
19. Li, Y. and Cho, Y.-Y.: Intelligent BSIM4 Model Parameter Extraction for Sub-100 nm MOSFETs era. Jpn. J. Appl. Phys. **43**(4B) (2004) 1717–1722
20. Li, Y., Sze, S. M., and Chao, T.-S.: A Practical Implementation of Parallel Dynamic Load Balancing for Adaptive Computing in VLSI Device Simulation. Engineering with Computers. **18**(2) (2004) 124–137
21. Huang, K.-Y., Li, Y., and Lee, C.-P.: A Time Domain Approach to Simulation and Characterization of RF HBT Two-Tone Intermodulation Distortion. IEEE Trans. Microwave Theory and Techniques. **51**(10) (2003) 2055–2062
22. Myers, R. H. and Montgomery, D. C.: Response surface methodology: process and product optimization using designed experiments. New York:Wiley. (2002)
23. Li, Y. and Yu, S.-M.: A Study of Threshold Voltage Fluctuations of Nanoscale Double Gate Metal-Oxide-Semiconductor Field Effect Transistors Using Quantum Correction Simulation. J. Comput. Electronics. **5** (2006) 125–129

Numerical Solution to Maxwell's Equations in Singular Waveguides

Franck Assous[1] and Patrick Ciarlet Jr.[2]

[1] Bar-Ilan University, 52900 Ramat-Gan, Israel
and College of Judea&Samaria, Ariel, Israel
franckassous@netscape.net
[2] ENSTA,32 bvd Victor, 75739, Paris Cedex 15
ciarlet@ensta.fr

Abstract. This paper is devoted to the numerical solution of the instationary Maxwell equations in *singular* waveguides. The geometry is called *singular*, as its boundary includes reentrant corners or edges, which generate, in their neighborhood, strong electromagnetic fields. We have built a method which allows to compute the time-dependent electromagnetic field, based on a splitting of the spaces of solutions: First, the subspace of regular fields, which coincides with the whole space of solutions, in the case of convex or smooth boundary; Second, a singular subspace, defined and characterized *via* the singularities of the Laplace operator. Numerical results illustrate the influence of frequency of the ingoing electromagnetic waves in a L-shaped waveguide.

1 Introduction

Many practical problems require the computation of electromagnetic fields. They are usually based on Maxwell equations. Within this framework, we developed a numerical method for solving the instationary Maxwell equations (see [5]), with continuous approximations of the electromagnetic field. However, in practical examples, the boundary of the computational domain includes reentrant corners and/or edges, called *geometrical singularities* because they generate strong fields.

We developed a method, the so-called *Singular Complement Method*, which consists in splitting the space of electromagnetic fields into a two-term, *direct*, possibly *orthogonal* sum. The first subspace is made of regular fields, the second one is called the subspace of *singular electromagnetic fields*. One compute the regular part of the solution with the help of an *ad hoc* – classical – method [5]. The singular part is computed with the help of specifically designed methods.

The present paper is a continuation of the *Singular Complement Method*, developed for Maxwell equations in 2D [4], and for the Vlasov-Maxwell equations [1]. We first recall Maxwell's equations, together with the functional framework, which is then used to describe the *Singular Complement Method*. Section 3 is devoted to the numerical algorithms. In particular, the computation of singular basis functions is described, together with the discretization of the variational formulations. Numerical experiments are presented in the last Section.

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 235–242, 2007.
© Springer-Verlag Berlin Heidelberg 2007

2 Mathematical Analysis of the Problem

Let Ω be a bounded, open, polyhedral subset of $\mathbb{R}^3$. We denote by Γ its boundary and by $\mathbf{n}$ the unit outward normal to Γ. If we let c, ε_0 and μ_0 be respectively the light velocity, the dielectric permittivity and the magnetic permeability ($\varepsilon_0\mu_0c^2 = 1$), Maxwell's equations in vacuum read,

$$\frac{\partial \mathcal{E}}{\partial t} - c^2\,\mathbf{curl}\,\mathcal{B} = -\frac{1}{\varepsilon_0}\mathcal{J}, \qquad \operatorname{div}\mathcal{E} = \frac{\rho}{\varepsilon_0},$$
$$\frac{\partial \mathcal{B}}{\partial t} + \mathbf{curl}\,\mathcal{E} = 0, \qquad \operatorname{div}\mathcal{B} = 0,$$

where $\mathcal{E}$ and $\mathcal{B}$ are the electric and magnetic fields, ρ and $\mathcal{J}$ the charge and current densities, which depend on the space variable $\mathbf{x}$ and on the time variable t. As it is well known, ρ and $\mathcal{J}$ have to verify the charge conservation

$$\frac{\partial \rho}{\partial t} + \operatorname{div}\mathcal{J} = 0.$$

These equations are supplemented with appropriate boundary conditions. For the sake of simplicity, we will only consider first perfectly conducting boundary. The case of Silver-Müller boundary condition will be then introduced. For the time being, we suppose that

$$\mathcal{E} \times \mathbf{n} = 0 \text{ and } \mathcal{B} \cdot \mathbf{n} = 0 \text{ on } \Gamma.$$

Finally, one adds initial conditions, set at time $t = 0$, $\mathcal{E}(0) = \mathcal{E}_0, \mathcal{B}(0) = \mathcal{B}_0$. We explain below how in our formulation the electric and the magnetic fields can be handled separately. Even if the principles of analysis are the same, the results and the mathematical tools are different (see theoretical details in [2]). The electric case generally appears in a non divergence-free modelling, typically the Vlasov-Maxwell equations, and has been exposed in [1]. In this paper, we will focus on the magnetic field formulation. Let us recall the definitions of the following spaces

$$\mathbf{H}(\mathbf{curl}\,,\Omega) = \{\mathbf{u} \in \mathbf{L}^2(\Omega), \mathbf{curl}\,\mathbf{u} \in \mathbf{L}^2(\Omega)\}\,,$$
$$\mathbf{H}(\operatorname{div}\,,\Omega) = \{\mathbf{u} \in \mathbf{L}^2(\Omega), \operatorname{div}\mathbf{u} \in L^2(\Omega)\}\,,$$
$$\mathbf{H}^1(\Omega) = \{\mathbf{u} \in \mathbf{L}^2(\Omega), \mathbf{grad}\,\mathbf{u} \in \mathbf{L}^2(\Omega)\}\,.$$

We define the space of magnetic fields $\mathcal{B}$, called $\mathbf{Y}$,

$$\mathbf{Y} = \{\mathbf{y} \in \mathbf{H}(\mathbf{curl}\,,\Omega) \cap \mathbf{H}(\operatorname{div}\,,\Omega) \ : \ \mathbf{y} \cdot \mathbf{n}_{|\Gamma} = 0\}\,.$$

In what follows, we use the notation $(\cdot,\cdot)_0$ for the usual scalar product in $L^2(\Omega)$ or $\mathbf{L}^2(\Omega)$, and $(\cdot,\cdot)_{\mathbf{Y}} = (\mathbf{curl}\,\cdot, \mathbf{curl}\,\cdot)_0 + (\operatorname{div}\cdot, \operatorname{div}\cdot)_0$ for the one in $\mathbf{Y}$.

When the domain is convex (or with a smooth boundary), the space of magnetic fields $\mathbf{Y}$ is included in $\mathbf{H}^1(\Omega)$. That is not the case anymore in a singular

domain (see for instance [10]). One thus introduces the regular subspace for magnetic fields (indexed with $_R$)

$$\mathbf{Y}_R = \mathbf{Y} \cap \mathbf{H}^1(\Omega),$$

which is actually closed in $\mathbf{Y}$ [9]. Hence, one can consider its orthogonal subspace (called singular subspace and indexed with $_S$), and then define a two-part, direct, and orthogonal sum of the space as

$$\mathbf{Y} = \mathbf{Y}_R \overset{\perp_{\mathbf{Y}}}{\oplus} \mathbf{Y}_S \,.$$

As a consequence, one can split an element $\mathbf{y}$ into an orthogonal sum of a *regular* part and of a *singular* one, namely $\mathbf{y} = \mathbf{y}_R + \mathbf{y}_S$. We have now to characterize the singular magnetic fields. Following [9], elements $\mathbf{y}_S \in \mathbf{Y}_S$ satisfy

$$\begin{aligned} \boldsymbol{\Delta}\, \mathbf{y}_S &= 0 \quad \text{in } \Omega \,, \\ \mathbf{y}_S \cdot \mathbf{n}_{|\Gamma} &= 0 \,. \end{aligned}$$

Now, we suppose that a part Γ_C of the boundary Γ behaves as a perfect conductor, namely $\mathcal{B} \cdot \mathbf{n}_{|\Gamma} = 0$. On the other part $\Gamma_A = \Gamma \setminus \Gamma_C$, we have to model the electromagnetic interactions between the domain Ω and the exterior. One has

$$(\mathcal{E} - c\mathcal{B} \times \mathbf{n}) \times \mathbf{n} = \mathbf{e}^\star \times \mathbf{n} \text{ on } \Gamma_A \,, \tag{1}$$

where the surface field $\mathbf{e}^\star$ is given. These conditions are known as the Silver-Müller boundary conditions. Moreover, the artificial boundary Γ_A is often splitted into Γ_A^i and Γ_A^a. On Γ_A^i, we model incoming plane waves by a non-vanishing function $\mathbf{e}^\star$, whereas we impose on Γ_A^a an absorbing boundary condition by choosing $\mathbf{e}^\star = 0$. Without loss of generality, one can choose the location of the artificial boundary Γ_A, in such a way that it does not intersect with a geometrical singularity. Moreover, one can also choose a regular shape for Γ_A.
Now, one could consider the space of solutions

$$\mathbf{Y}^{\Gamma_A} = \{\mathbf{y} \in \mathbf{H}(\mathbf{curl}\,, \Omega) \cap \mathbf{H}(\mathrm{div}\,, \Omega) \ : \ \mathbf{y} \cdot \mathbf{n}_{|\Gamma_C} = 0\} \,.$$

Then introduce the regular subspace $\mathbf{Y}_R^{\Gamma_A} = \mathbf{Y}^{\Gamma_A} \cap \mathbf{H}^1(\Omega)$, and construct the ad hoc orthogonal splitting, in which appears a singular space, say $\mathbf{Y}_S^{\Gamma_A}$. Nevertheless, it is more interesting from a numerical point of view, to consider the (non-orthogonal) splitting

$$\mathbf{Y}^{\Gamma_A} = \mathbf{Y}_R^{\Gamma_A} \oplus \mathbf{Y}_S \,.$$

First, since the subspace of singular magnetic fields is $\mathbf{Y}_S$, as before. Second, modelling incoming plane waves, or imposing an absorbing boundary condition has no impact, as far as the singular subspace is concerned. It will be sufficient, as soon as Γ_A is not empty, to add in the variational formulation, integral terms on Γ_A as for a regular domain Ω.

3 Numerical Algorithms

The numerical method consists in computing first the basis of the singular subspace. Then we solve the problem by coupling a classical method (to compute the regular part of the solution) to the linear system, which allows to compute the singular part of the solution.

To compute $\mathbf{y}_S \in \mathbf{Y}_S$, it is convenient to introduce its divergence- and curl-free parts $\mathbf{w}_S$ and $\mathbf{m}_S$, which verify the following Helmholtz decomposition

$$\mathbf{y}_S = \mathbf{w}_S + \mathbf{m}_S \,. \tag{2}$$

From now on, we assume that the singular subspace $\mathbf{Y}_S$ is finite-dimensional. This is actually the case for a two-dimensional domain Ω, where the dimension of the singular subspace is equal to the number of reentrant corners (cf. [4]). The three-dimensional case can be written formally as below, but some technical mathematical tools are needed (unclassical spaces, weak trace properties, etc.). We refer the reader to [2] for more theoretical details.

We shall need s_N and s_D, the *non-vanishing*, *singular*, harmonic functions, with Neumann and Dirichlet homogeneous boundary condition respectively, solutions to

$$\begin{aligned} &\Delta s_N = 0, && \Delta s_D = 0\,, && \text{in } \Omega \\ &\frac{\partial s_N}{\partial \boldsymbol{\nu}} = 0, && s_D = 0\,, && \text{on } \Gamma\,. \end{aligned}$$

Remark that s_N and s_D are not equal to zero since we are looking for a *singular* solution, namely with a too poor regularity to be a variational solution. One then introduce ϕ_S and ψ_S respectively solution to

$$-\Delta\phi_S = s_N, \qquad -\Delta\psi_S = s_D \qquad \text{in } \Omega, \tag{3}$$

still with Neumann and Dirichlet homogeneous boundary condition. Next, the singular basis functions $\mathbf{y}_S \in \mathbf{Y}_S$ can be obtained (see [3]) with the relation

$$\mathbf{w}_S = \mathbf{curl}\,\psi_S, \qquad \mathbf{m}_S = \mathbf{grad}\,\phi_S\,, \tag{4}$$

together with relation (2). Hence, the keypoint is to compute s_N and s_D.

Consider, for simplicity reasons, a domain with one reentrant corner. To compute s_N and s_D, we have chosen to use the *Principal Part Method*. Let us describe it on s_D. It consists in splitting s_D in a regular part $\tilde{s}_D$ (which belongs to $H^1(\Omega)$) and a known singular part s_D^P

$$s_D = s_D^P + \tilde{s}_D\,. \tag{5}$$

It is common knowledge that $s_D^P = r^{-\alpha}\sin(\alpha\theta)$, where (r,θ) denote the polar coordinates centered on the reentrant corner of angle π/α. Above, s_D^P is singular since it belongs to $L^2(\Omega)$ but not to $H^1(\Omega)$, and verifies $\Delta s_D^P = 0$. One thus computes with a P^1 finite element method, $\tilde{s}_D$ by solving

$$\begin{aligned} &\Delta\tilde{s}_D = 0 \text{ in } \Omega, \\ &\tilde{s}_D = -s_D^P \text{ on } \Gamma\,. \end{aligned}$$

Next, one proceeds similarly for the function $\psi_S \in H_0^1(\Omega)$ solution to

$$-\Delta\psi_S = s_D \text{ in } \Omega,$$
$$\psi_S = 0 \text{ on } \Gamma\,.$$

Again, splitting ψ_S in a regular part $\tilde{\psi}_S$ (which belongs to $H^2(\Omega)$) and a singular one ψ_S^P,

$$\psi_S = \tilde{\psi}_S + C_\psi \psi_S^P\,, \tag{6}$$

where C_ψ is a constant which can be determined with an integration by parts formula (cf. [3]). One needs the expression of ψ_S^P in polar coordinates, $\psi_S^P = r^\alpha \sin(\alpha\theta)$. The regular part $\tilde{\psi}_S$ is then computed, by solving a standard variational formulation. The singular function s_N and ϕ_S are obtained in the same way.

With the help of singular mappings (see [2]), one can compute the singular electromagnetic basis functions. We get the basis $\mathbf{w}_S$ (resp. $\mathbf{m}_S$) by simply taking the curl of ψ_S (resp. the gradient of ϕ_S)

$$\mathbf{w}_S = \mathbf{curl}\,\tilde{\psi}_S + C_\psi \mathbf{curl}\,\psi_S^P\,, \tag{7}$$
$$\mathbf{m}_S = \nabla\tilde{\phi}_S + C_\phi \nabla\phi_S^P\,, \tag{8}$$

and $\mathbf{y}_S$ is easily obtained with (2).

We recall now the *Variational Formulation*, or *VF*, which have been developed to solve the problem. We also introduce the discretization of this VF. First, Ampère and Faraday's laws are written equivalently as two second-order in time equations, plus suitable initial and boundary conditions. Then, the electric and magnetic fields are decoupled (up to the initial conditions). Next, following [5], we enforce the divergence constraints on the electromagnetic field by introducing two Lagrange multipliers, which dualize Coulomb's and absence of free magnetic monopole's laws. This approach gives a Mixed VF of Maxwell's equations, which is well-posed, if the well-known *inf-sup* (or Babuska-Brezzi [6,7]) condition holds. In addition, we use an Augmented VF, by adding to the bilinear form $(\mathbf{curl}\,\cdot, \mathbf{curl}\,\cdot)_0$, the term $(\mathrm{div}\,\cdot, \mathrm{div}\,\cdot)_0$. This results in a Mixed, Augmented VF, or *MAVF*. In our case, the magnetic field belongs to $\mathbf{Y}$. Then, the correct Lagrange multiplier space is $L_0^2(\Omega)$. Denote by $p(t)$ the Lagrange multiplier, this formulation reads

Find $(\mathcal{B}(t), p(t)) \in \mathbf{Y} \times L_0^2$ *such that*

$$\frac{d^2}{dt^2}(\mathcal{B}(t), \mathbf{y})_0 + c^2(\mathcal{B}(t), \mathbf{y})_\mathbf{Y} + (p(t), \mathrm{div}\,\mathbf{y})_0 = \frac{1}{\varepsilon_0}(\mathcal{J}(t), \mathbf{curl}\,\mathbf{y})_0\,, \quad \forall \mathbf{y} \in \mathbf{Y},$$
$$(\mathrm{div}\,\mathcal{B}(t), q)_0 = 0 \qquad \forall q \in L_0^2(\Omega)\,.$$

One has to include the regular/singular splitting in this formulation. The magnetic field $\mathcal{B}$ being decomposed into $\mathcal{B}(t) = \mathcal{B}_R(t) + \mathcal{B}_S(t)$, and the same for the test functions, the variational formulation reads now:
Find $(\mathcal{B}_R(t), \mathcal{B}_S(t), p(t)) \in \mathbf{Y}_R \times \mathbf{Y}_S \times L^2_0(\Omega)$ *such that*

$$\frac{d^2}{dt^2}(\mathcal{B}_R(t), \mathbf{y}_R)_0 + c^2(\mathcal{B}_R(t), \mathbf{y}_R)_{\mathbf{Y}} + (p(t), \operatorname{div} \mathbf{y}_R)_0 = \frac{1}{\varepsilon_0}(\mathcal{J}(t), \mathbf{curl}\, \mathbf{y}_R)_0 - \frac{d^2}{dt^2}(\mathcal{B}_S(t), \mathbf{y}_R)_0, \qquad \forall \mathbf{y}_R \in \mathbf{Y}_R, \tag{9}$$

$$\frac{d^2}{dt^2}(\mathcal{B}_S(t), \mathbf{y}_S)_0 + c^2(\mathcal{B}_S(t), \mathbf{y}_S)_{\mathbf{Y}} + (p(t), \operatorname{div} \mathbf{y}_S)_0 = \frac{1}{\varepsilon_0}(\mathcal{J}(t), \mathbf{curl}\, \mathbf{y}_S)_0 - \frac{d^2}{dt^2}(\mathcal{B}_R(t), \mathbf{y}_S)_0, \qquad \forall \mathbf{y}_S \in \mathbf{Y}_S, \tag{10}$$

$$(\operatorname{div} \mathcal{B}_R(t), q)_0 + (\operatorname{div} \mathcal{B}_S(t), q)_0 = 0, \qquad \forall q \in L^2_0(\Omega). \tag{11}$$

Remark 3.1. *In the case of a non-orthogonal splitting, as for instance with a non-empty absorbing boundary* Γ_A*, one has to add* $-c^2(\mathcal{B}_S(t), \mathbf{y}_R)_{\mathbf{Y}}$ *to equation (9), and the term* $-c^2(\mathcal{B}_R(t), \mathbf{y}_S)_{\mathbf{Y}}$ *to equation (10) (generated by the loss of orthogonality).*

To derive a finite-element approximation of this formulation, we have now to choose discrete fields and test-functions, which verify a uniform, discrete *inf-sup* condition. The Taylor-Hood, P_2-iso-P_1 Finite Element retains our attention here, first because it fulfill this condition (cf. [8]). Moreover, it allows to build diagonal mass matrices, when suitable quadrature formulas are used (cf. [5]). Thus, the solution to the linear system, which involves the mass matrix, is straightforward. Since there is one singularity, we can write $\mathcal{B}_S(t) = \kappa(t)\mathbf{y}_S$, where $(\mathbf{y}_S)$ denotes the basis of the discrete singular space, and κ is a continous time-dependent function. Next, discretizing in time this formulation with the help of the well-known leap-frog scheme, this results in the following fully discretized scheme:

$$\mathbb{M}_\Omega \boldsymbol{B}_R^{n+1} + \mathbb{M}_{RS}\boldsymbol{\kappa}^{n+1} + \mathbb{L}_\Omega \boldsymbol{p}^{n+1} = \boldsymbol{F}^n\,, \tag{12}$$

$$\mathbb{M}_{RS}^T \boldsymbol{B}_R^{n+1} + \mathbb{M}_S\boldsymbol{\kappa}^{n+1} + \mathbb{L}_S \boldsymbol{p}^{n+1} = \boldsymbol{G}^n\,, \tag{13}$$

$$\mathbb{L}_\Omega^T \boldsymbol{B}_R^{n+1} + \mathbb{L}_S^T\boldsymbol{\kappa}^{n+1} = 0\,. \tag{14}$$

Above $\mathbb{M}_\Omega$ denotes the usual mass matrix, and $\mathbb{L}_\Omega$ corresponds to the divergence term involving $\mathbf{y}_R^h$ and $p_h(t)$. Then, $\mathbb{M}_{RS}$ is a rectangular matrix, which is obtained by taking $\mathbf{L}^2$ scalar products between regular and singular basis functions, $\mathbb{M}_S$ is the "singular" mass matrix, and finally, $\mathbb{L}_S$ corresponds to the divergence term involving $\mathbf{y}_S^i$ and $p_h(t)$.

One can solve this system by removing the unknown $\boldsymbol{\kappa}^{n+1}$. To that aim, replace equation (12) by (12)$-\mathbb{M}_{RS}\mathbb{M}_S^{-1}$(13), and equation (14) by (14)$-\mathbb{L}_S^T\mathbb{M}_S^{-1}$(13).

In this modified system, only the unknowns $(\boldsymbol{B}_R^{n+1}, \boldsymbol{p}^{n+1})$ appear. If one lets $\widetilde{\ }$ stand for the modified matrices and right-hand sides, it reads

$$\begin{aligned} \widetilde{\mathbb{M}}\boldsymbol{B}_R^{n+1} \ + \ \widetilde{\mathbb{L}}\boldsymbol{p}^{n+1} \ &= \widetilde{\boldsymbol{F}^n} \ , \\ \widetilde{\mathbb{L}}^T\boldsymbol{B}_R^{n+1} \ - \ \mathbb{L}_S^T\mathbb{M}_S^{-1}\mathbb{L}_S\boldsymbol{p}^{n+1} \ &= \widetilde{\boldsymbol{H}^n} \ . \end{aligned}$$

Its solution can be obtained with the help of a Uzawa-type algorithm. Finally, one concludes the time-stepping scheme by computing $\boldsymbol{\kappa}^{n+1}$ with the help of (13).

4 Numerical Experiments

We study now the influence of the frequency of the incoming signal, on the localization in space of the singular effects. This is of importance, since not taking into account the singular part of the electromagnetic field can result in a computed solution, which is wrong over the whole domain.

We consider an L-shaped domain Ω, with a boundary Γ split into two parts, $\Gamma = \Gamma_C \cup \Gamma_A$. On Γ_C (top and bottom parts), a perfect conducting boundary condition is imposed. An incident wave enters the waveguide through the boundary Γ_A^i (left side), and exists through Γ_A^a (rigth side). This behavior is modelled thanks to the boundary condition (1) with a right hand-side equal to $C\sin(\omega t)$ on Γ_A^i and zero on Γ_A^a. Initial conditions are uniformly set to zero. Above, C is a constant, and ω is associated to a frequency ν, which can vary.

We compare two numerical solutions. One, which is obtained by taking into account the singular part, i.e. with the *Singular Complement Method* or *SCM*. The other one, by computing only a P^1, Lagrange finite element approximation. The values of ν are set successively to $\nu_1 = 5.10^9$ Hz, and $\nu_2 = 15.10^9$ Hz. In each case, the mesh is such that the number of discretization nodes per wavelength is constant. Numerical dispersion, if it occurs, is therefore comparable.

Results are shown on Figure 1. For the higher frequency ν_2, the height is roughly equal to seven wavelengths and the singular behavior is more localized, near the reentrant corner. Results are rather close. But for ν_1, which corresponds

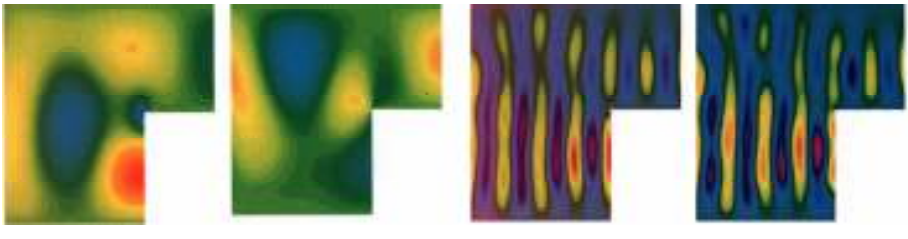


Fig. 1. with (left) and without (right) *SCM* for low frequency ν_1 - with (left) and without (right) *SCM* for high frequency ν_2

to a wavelength comparable to the dimensions of the domain (the height is roughly equal to two wavelengths), the solutions are very different.

5 Conclusion

In this paper, we were interested in the propagation of a wave in a singular waveguide, by studying how the frequency of the ingoing electromagnetic waves influences the singular solution. We developed a numerical method, based on direct, and possibly orthogonal, splittings of the space of electromagnetic solutions. One of the foremost result is that the singular behavior of the solution is more localized, near the reentrant corner for the higher frequencies than for the lower.

References

1. **F. Assous, P. Ciarlet, Jr.**, Solving Vlasov-Maxwell equations in singular geometries, accepted to *Math. and Comput. in Simulation*.
2. **F. Assous, P. Ciarlet, Jr., E. Garcia**, A characterization of the singular electromagnetic fields by an inductive approach, submitted to *Math. Meth. Appl. Sci.*.
3. **F. Assous, P. Ciarlet, Jr., E. Garcia, J. Segré**, Time dependent Maxwell's equations with charges in singular geometries, *Comput. Methods Appl. Mech. Engrg.*, **196**, 665–681, 2006.
4. **F. Assous, P. Ciarlet, Jr., J. Segré**, Numerical solution to the time-dependent Maxwell equations in two-dimensional singular domain: The Singular Complement Method, *J. Comput. Phys.*, **161**, 218-249 (2000).
5. **F. Assous, P. Degond, E. Heintzé, P. A. Raviart, J. Segré**, On a finite element method for solving the three-dimensional Maxwell equations, *J. Comput. Phys.*, **109**, 222-237 (1993).
6. **I. Babuska (1973)**, The finite element method with Lagrange multipliers, *Numer. Math.*, **20**, 179-192 (1973).
7. **F. Brezzi**, On the existence, uniqueness and approximation of saddle point problems arising from Lagrange multipliers, *RAIRO Anal. Numér.*, 129-151, (1974).
8. **P. Ciarlet, Jr., V. Girault**, *Inf-sup* condition for the 3D, P_2-iso-P_1, Taylor-Hood finite element; application to Maxwell equations, *C. R. Acad. Sci. Paris, Ser. I*, **335**, 827-832 (2002).
9. **E. Garcia**, *Résolution des équations de Maxwell avec charges dans des domaines non convexes*, PhD Thesis, University Paris 6, France (2002). (in French)
10. **P. Grisvard**, *Singularities in boundary value problems*, **22**, *RMA* Masson, Paris (1992).

Quantum-Inspired Genetic Algorithm Based Time-Frequency Atom Decomposition*

Gexiang Zhang and Haina Rong

School of Electrical Engineering, Southwest Jiaotong University,
Chengdu 610031 Sichuan, China
gxzhang@ieee.org

Abstract. The main problem of time-frequency atom decomposition (TFAD) lies in an extremely high computational load. This paper presents a fast implementation method based on quantum-inspired genetic algorithm (QGA). Instead of finding the optimal atom in greedy implementation algorithm, this method is to search a satisfactory atom in every iteration of TFAD. Making full use of QGA's advantages such as good global search capability, rapid convergence and short computing time, the method reduces greatly the computational load of TFAD. Experiments conducted on radar emitter signals verify the effectiveness and practicality of the introduced method.

Keywords: Quantum-inspired genetic algorithm, time-frequency atom decomposition, fast implementation algorithm.

1 Introduction

Time-frequency atom decomposition (TFAD), also known as matching pursuit or adaptive Gabor representation [1,2], was introduced independently in [3] and [4]. TFAD is an approach that decomposes any signal into a linear expansion of waveforms selected from a redundant dictionary of time-frequency atoms that well localized both in time and frequency [3]. Unlike Wigner and Cohen class distributions, the energy distribution obtained by TFAD does not include interference terms. Different from Fourier and Wavelet orthogonal transforms, the information in TFAD is not diluted across the whole basis [3]. Hence, TFAD has become an attractive analysis technique in signal processing and harmonic analysis [1, 3-5]. However, the the necessary dictionary of time-frequency atoms being very large, the computational load turns out to be the main problem of TFAD [2]. After a greedy algorithm was presented in [3], several fast implementation algorithms were introduced in [6-9]. Although genetic algorithm was used to reduce the computational load in [1,2,6-8], this problem need be studied further by using some unconventional computation methods with some advantages over conventional genetic algorithms.

* This work was supported by the Scientific and Technological Development Foundation of Southwest Jiaotong University (2006A09) and by the National Natural Science Foundation of China (60572143).

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 243–250, 2007.
© Springer-Verlag Berlin Heidelberg 2007

This paper introduces quantum-inspired genetic algorithm (QGA) into TFAD to lower the computational complexity. Based on the concepts of quantum computing, QGA falls into the latest category of unconventional computation. Due to some outstanding advantages such as good global search capability, rapid convergence and short computing time [10-12], QGA is able to accelerate greatly the process of searching the most satisfactory time-frequency atom in each iteration of TFAD. In Section 2, the detailed algorithm of QGA based TFAD is described. Next, some experiments are conducted on radar emitter signals in Section 3. Finally, some conclusions are drawn.

2 QGA Based TFAD

The structure of QGA Based TFAD is presented in Algorithm 1 and the detailed description is as follows.

Algorithm 1. *Algorithm of QGA Based TFAD*

```
      Begin
(i)     Initialization of TFAD; % Initial iteration T=1
(ii)    While (not termination condition of TFAD) do
          % Searching the most satisfactory atom using QGA
(iii)     Setting initial values of parameters in QGA; % generation g=0
(iv)      Initializing P(g);
(v)       Generate R(g) by observing P(g); %
(vi)      Fitness evaluation; %
(vii)     Store the best solution among P(g) into B(g);
          While (not termination condition of QGA) do
            g=g+1;
(viii)      Generate R(g) by observing P(g-1); %
(ix)        Fitness evaluation; %
(x)         Update P(g) using quantum rotation gate; %
(xi)        Store the best solution among P(g) and B(g-1) into B(g); %
            If (migration condition)
(xii)         Migration operation;
            End if
            If (catastrophe condition)
(xiii)        Catastrophe operation;
            End if
          End while
          T=T+1;
(xiv)     Computing residual signal %
        End while
      End
```

(i) In the first step, the initial iteration T is set to 1. The original signal, represented linearly by time-frequency atoms, is loaded from the database or generated in simulation way. Let f be the original signal, $f \in \mathbf{H}$, where

H is a Hilbert space. By using TFAD, the signal f can be represented with a linear expansion of time-frequency atoms that are dilations, translations, and modulations of a single window function g_γ ($\|g_\gamma\| = 1$) [3]. Let $\mathscr{D} = \{g_\gamma\}_{\gamma \in \Gamma}$, where Γ is the set of the index γ that is composed of y parameters, i.e. $\gamma = (\gamma_1, \gamma_2, \cdots, \gamma_y)$. In this paper, time-frequency atom uses Gabor function:

$$g_\gamma(t) = \frac{1}{\sqrt{s}} g(\frac{t-u}{s}) \cos(\nu t + \omega) \ . \tag{1}$$

where the index γ=(s, u, ν, ω) is a set of parameters and s, u, ν, ω are scale, translation, frequency and phase, respectively. $g(\cdot)$ is a Gauss-modulated window function as

$$g(t) = e^{-\pi t^2} \ . \tag{2}$$

In the index γ=(s, u, ν, ω), there are four parameters to optimize. The method for choosing the four parameters is $\gamma = (c^x, bc^x \Delta u, dc^{-x} \Delta\nu, z\Delta\omega)$, $c = 1.5$, $\Delta u = 1/2$, $\Delta\nu = \pi$, $\Delta\omega = \pi/6$, $0 < x \leq \log_2 L$, $0 < b \leq L/2^{x-1}$, $0 \leq z \leq 12$, $0 \leq d \leq 2^{x+1}$, where L is the length of the signal f [3].

(ii) In TFAD, the maximal number of iteration is usually used as the termination condition.

(iii) The initial values of some parameters in QGA are set. The parameters include population size N, the number y=4 of variables, the number m of binary bits of each variable, the maximal evolutionary generation g of QGA for searching the sub-optimal time-frequency atom in the over-complete dictionary (i.e. in each iteration), immigration generation M_g and catastrophe generation C_g.

(iv) In this step, population $\boldsymbol{P}(g)$=$\{\boldsymbol{p}_1^g, \boldsymbol{p}_2^g, \cdots, \boldsymbol{p}_N^g\}$, where $\boldsymbol{p}_i^g (i = 1, 2, \cdots, N)$ is an arbitrary individual in population and $\boldsymbol{p}_i^g$ is represented as

$$\boldsymbol{p}_i^g = \begin{bmatrix} \alpha_{i1}^g | \alpha_{i2}^g | \cdots | \alpha_{i(my)}^g \\ \beta_{i1}^g | \beta_{i2}^g | \cdots | \beta_{i(my)}^g \end{bmatrix} \ . \tag{3}$$

where $\alpha_{ij}^g = \beta_{ij}^g = 1/\sqrt{2}$ (j=$1, 2, \cdots, my$), which means that all states are superposed with the same probability. Here, α_{ij}^g and β_{ij}^g ($j = 1, 2, \cdots, my$) are the probability amplitudes of one qubit and satisfy the normalization equality: $|\alpha|^2 + |\beta|^2 = 1$, where $|\alpha|^2$ and $|\beta|^2$ are the probabilities that the qubit will be observed in '0' state and in '1' state in the act of observing the quantum state.

(v) According to probability amplitudes of all individuals in $\boldsymbol{P}(g)$, observation states $\boldsymbol{R}(g)$ is constructed by observing $\boldsymbol{P}(g)$. Here $\boldsymbol{R}(g)$=$\{\boldsymbol{a}_1^g, \boldsymbol{a}_2^g, \cdots, \boldsymbol{a}_N^g\}$, where $\boldsymbol{a}_i^g (i = 1, 2, \cdots, N)$ is an observation state of an individual. $\boldsymbol{a}_i^g$ is a binary string with the length of my, that is $\boldsymbol{a}_i^g = b_1 b_2 \cdots b_{my}$, where b_j ($j = 1, 2, \cdots, my$) is one binary number '0' or '1'. Observation states $\boldsymbol{R}(g)$ is generated in probabilistic way. For the probability amplitude $[\alpha \ \beta]^T$ of a qubit, a random number r in the range [0, 1] is generated. If $r < |\alpha|^2$, the corresponding observation value is set to '0', otherwise, the value is set

to '1'. In the process of constructing observation state $\boldsymbol{R}(g)$ using $\boldsymbol{P}(g)$, the decoding operation in conventional genetic algorithm is included. After decoding, the parameter values of all optimization parameters can be obtained.

(vi) The value of fitness function of each individual in population is computed by using the values of y variables and each individual in population is evaluated. Here, the fitness is chosen as $|\langle f, g_\gamma\rangle|$. The reason is as follows. The signal f will be decomposed into $f = \langle f, g_\gamma\rangle g_\gamma + Rf$, where Rf is the residual signal after approximating f in the direction of g_γ. Obviously, g_γ is orthogonal to Rf. So the following equality can be obtained [3].

$$\|f\|^2 = |\langle f, g_\gamma\rangle|^2 + \|Rf\|^2 \,. \tag{4}$$

In (4), only if $|\langle f, g_\gamma\rangle|$ ($g_\gamma \in \mathscr{D}$) is maximum, $\|Rf\|$ will be minimum.

(vii) The best solution in $\boldsymbol{P}(g)$ at generation g is stored into $B(g)$.

(viii) According to probability amplitudes of all individuals in $\boldsymbol{P}(g-1)$, observation states $\boldsymbol{R}(g)$ is constructed by observing $\boldsymbol{P}(g-1)$. This step is similar to step (v). The detailed procedures have been described.

(ix) This step is similar to step (vi).

(x) The probability amplitudes of all qubits in population are updated by using quantum rotation gates given in (5).

$$\boldsymbol{G} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \,. \tag{5}$$

where θ is the rotation angle of quantum rotation gate and θ is defined as

$$\theta = k \cdot f(\alpha, \beta) \,. \tag{6}$$

where k is a coefficient whose value has a direct effect on convergent speed of QGA. In this paper, k is defined as a variable, which is relative to the evolutionary generation g. A kind of definition of k is given in (7).

$$k = \frac{\pi}{2} e^{-\frac{mod(g,100)}{10}} \,. \tag{7}$$

where mod(g,100) is a function for computing modulus after g is divided by 100. k varies from $\pi/2$ to 0, which indicates QGA searches the optimal solution at large grid at the beginning of the algorithm and then the search grid declines gradually to 0 as the evolutionary generation g increases to 100. When the evolutionary generation g amounts to 100, the value of k comes back to $\pi/2$. $f(\alpha, \beta)$ is a function for determining the search direction of QGA to a global optimum. The look-up table of $f(\alpha, \beta)$ is shown in Table 1, in which $\xi_1 = \arctan(\beta_1/\alpha_1)$ and $\xi_2 = \arctan(\beta_2/\alpha_2)$, where α_1, β_1 are the probability amplitudes of the best solution stored in $B(g)$ and α_2, β_2 are the probability amplitudes of the current solution [12]. Thus, the formula for updating an individual $\boldsymbol{p}_i^g$ in $\boldsymbol{P}(g)$ using quantum rotation gate $\boldsymbol{G}$ can be described as

$$\boldsymbol{p}_i^{g+1} = \boldsymbol{G}(g) \cdot \boldsymbol{p}_i^g \,. \tag{8}$$

Table 1. Look-up table of function $f(\alpha, \beta)$ (Sign is a symbolic function)

$\xi_1 > 0$	$\xi_2 > 0$	$f(\alpha, \beta)$	
		$\xi_1 \geq \xi_2$	$\xi_1 < \xi_2$
True	True	+1	-1
True	False	$\text{sign}(\alpha_1 \cdot \alpha_2)$	
False	True	$-\text{sign}(\alpha_1 \cdot \alpha_2)$	
False	False	$\text{sign}(\alpha_1 \cdot \alpha_2)$	$-\text{sign}(\alpha_1 \cdot \alpha_2)$
$\xi_1, \xi_2 = 0$ or $\pi/2$		± 1	

where g is evolutionary generation, $\boldsymbol{G}(g)$ stands for the gth generation quantum rotation gate, $\boldsymbol{p}_i^g$ and $\boldsymbol{p}_i^{g+1}$ are the probability amplitudes at gth generation and at $(g+1)$th generation, respectively. Once the probability amplitudes of all individuals in $\boldsymbol{P}(g)$ are updated using quantum rotation gate, the individuals in the next generation are generated.

(xi) The best solution among $\boldsymbol{P}(g)$ and $B(g-1)$ is stored into $B(g)$.

(xii) When immigration operation is made every M_g generation, the 10% of probability amplitudes in the stored best individual are replaced by the new probability amplitudes generated using the similar method in step (ii). That is, each α is a random value from 0 to 1 and corresponding β equals $\pm\sqrt{1-|\alpha|^2}$. How to choose M_g is discussed in the next section.

(xiii) If the best solution stored in $B(g)$ is not changed in many generations, such as C_g generations, then population catastrophe operation should be performed.

(xiv) The residual signal $R^{(T+1)}f$ is $R^{(T+1)}f = R^T f - \langle R^T f, g_{\gamma_T}\rangle g_{\gamma_T}$. The residual signal is used as the original signal in the next iteration.

After the signal f is decomposed up to the order T, f can be represented with the concatenated sum

$$f = \sum_{i=0}^{T} \langle f, g_{\gamma_i}\rangle g_{\gamma_i} + R^{T+1} f \ . \tag{9}$$

where g_{γ_i} satisfies

$$|\langle R^T f, g_{\gamma_i}\rangle| = \alpha \sup_{\gamma \in \Gamma} |\langle R^T f, g_\gamma\rangle| \ . \tag{10}$$

According to the conclusion [3]: $\lim_{T\to\infty} \|R^T f\| = 0$, the signal f can be represented as

$$f = \sum_{i=0}^{T} \langle f, g_{\gamma_i}\rangle g_{\gamma_i} \ . \tag{11}$$

To compute the correlation between the original signal f and the restored signal f_r with parts of decomposed time-frequency atoms, resemblance coefficient method [13] is used to the correlation ratio C_r of f and f_r.

$$C_r = \frac{\langle f, f_r\rangle}{\|f\| \cdot \|f_r\|} \ . \tag{12}$$

The decaying value d_c is a function of the number of iteration T. d_c is computed using the following formula.

$$d_c = \log_{10} \frac{\|R^T f\|}{\|f\|} \ . \tag{13}$$

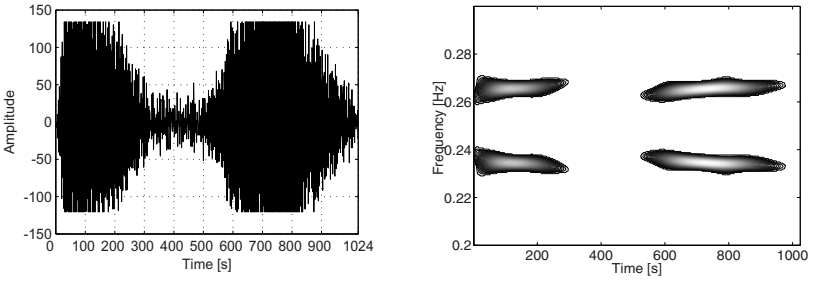
where $R^T f$ is the residual signal after approximating the original signal f using the first T time-frequency atoms [3]. The variable d_c is directly related to convergent speed of TFAD implementation algorithm.

3 Experiments

To test the validity of the introduced method, a real radar emitter signal is used to conduct the experiment, in which greedy implementation algorithm (GIA) [3] is brought into comparison with QGA. Experimental environment is chosen as: the maximal number of iteration is set to 500 as the termination condition of TFAD; From our prior tests, c=1.5 is much better than c=2 in [3]; In QGA, population size N, the number m of binary bits of each variable, the maximal evolutionary generation g, immigration generation M_g and catastrophe generation C_g are set to 20, 10, 200, 15 and 25, respectively. These experiments are carried out on the personal computer with 3.06GHz CPU,1GHz EMS memory and 160GB hard disk. The performances of the two algorithms are evaluated by using computing efficiency and optimization result. Computing efficiency includes computing time and the decay performance given in (13). Optimization result is evaluated by using the correlation ratio of the original signal and approximation signal given in (12). Figure 1 shows the noised radar emitter signal and its time-frequency distribution. Experimental results are given in Fig.2, Table 2 and Fig.3. The decaying curves of GIA and QGA in Fig.2 show the difference between the two algorithms. Figure 3 illustrates time-frequency distributions of restored radar emitter signals from the decomposed time-frequency atoms. As can be seen from Fig.2 and Table 2, the introduction of QGA into TFAD shortens greatly the computing time. In Table 2, the computing time of QGA is about 30 times as small as that of GIA. Figure 3 brings us some important hints that time-frequency distributions constructed by decomposed time-frequency atoms are nearly identical with that of the original radar emitter signals and TFAD is a good decomposition approach of a signal. The experimental results manifest that QGA is a better method than GIA for solving the problem of computing load in time-frequency atom decomposition.

Table 2. Performance comparisons of GIA and QGA

Methods	QGA	GIA
Correlation ratio	0.9903	0.9904
Computing time (Second)	4130.8750	123850.5781



(a) Time-domain signal (b) Time-frequency distribution

Fig. 1. A radar emitter signal with noise

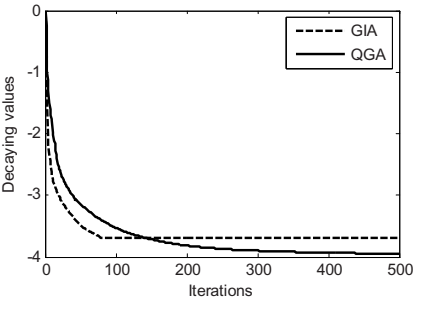
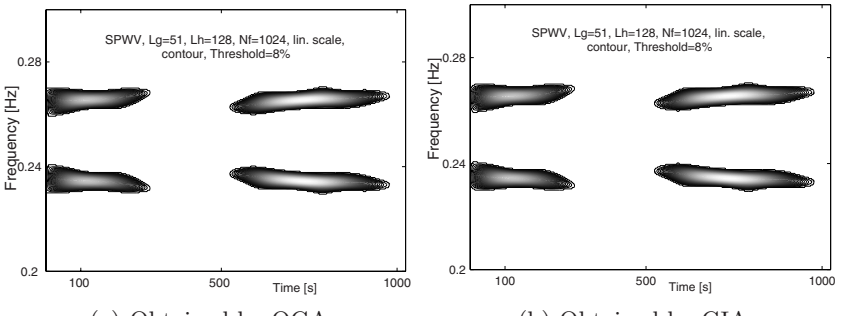


Fig. 2. The decaying curves of GIA and QGA



(a) Obtained by QGA (b) Obtained by GIA

Fig. 3. Time-frequency distributions of restored signals

4 Concluding Remarks

This paper presents a fast and efficient implementation method for TFAD based on QGA. The steps of the method are described in detail. The experiment on a

real radar emitter signal has been carried out to verify the validity and efficiency of the introduced method. Based on the concepts and principles of quantum computing and taking advantage of the strong parallelism of quantum computing, QGA uses its good characteristics of good global search capability, rapid convergence and short computing time to decrease greatly the computational load of TFAD. Our further study will aim at feature extraction and analysis of radar emitter signals by using the introduced method.

References

1. Lopez-Risueno, G., Grajal, J.: Unknown Signal Detection via Atomic Decomposition. In: Proceedings of the 11^{th} IEEE Signal Processing Workshop on Statistical Signal Processing (2001) 174-177
2. Vesin, J.: Efficient Implementation of Matching Pursuit Using a Genetic Algorithm in the Continuous Space. In: Proceedings of 10^{th} European Signal Processing Conference (2000) 2-5
3. Mallat, S.G., Zhang, Z.F.: Matching Pursuits with Time-Frequency Dictionaries. IEEE Transactions on Signal Processing **41** (1993) 3397-3415
4. Qian, S., Chen, D.: Signal Representation Using Adaptive Normalized Gaussian Functions. Signal Processing **36** (1994) 1-11
5. Gribonval, R., Bacry, E.: Harmonic Decomposition of Audio Signals with Matching Pursuit. IEEE Transactions on Signal Processing **51** (2003) 101-111
6. Figueras i Ventura, R.M., Vandergheynst, P.: Matching Pursuit through Genetic Algorithms. LTS-EPFL Tech. Report (2001) 1-14
7. Yin, Z.K., Wang, J.Y., Pierre, V.: Signal Sparse Decomposition Based on GA and Atom Property. Journal of the China Railway Society **27** (2005) 58-61
8. Stefanoiu, D., Llonescu, F.: A Genetic Matching Pursuit Algorithm. In: Proceedings of 7^{th} International Symposium on Signal Processing and Its Applications (2003) 577-580
9. Gribonval, R.: Fast Matching Pursuit with a Multiscale Dictionary of Gaussian Chirps. IEEE Transactions on Signal Processing **49** (2001) 994-1001
10. Zhang, G.X., Jin, W.D., Li, N.: An Improved Quantum Genetic Algorithm and Its Application. In: Wang, G., et al., (eds.): Lecture Notes in Artificial Intelligence, Vol.2639. Springer-Verlag, Berlin Heidelberg New York (2003) 449-452
11. Zhang, G.X., Hu, L.Z., Jin, W.D.: Quantum Computing Based Machine Learning Method and Its Application in Radar Emitter Signal Recognition. In: Torra, V., Narukawa, Y., (eds.): Lecture Notes in Artificial Intelligence, Vol.3131. Springer-Verlag, Berlin Heidelberg New York (2004) 92-103
12. Zhang, G.X., Rong, H.N., Jin, W.D.: An Improved Quantum-Inspired Genetic Algorithm and Its Application to Time-Frequency Atom Decomposition. Dynamics of Continuous, Discrete and Impulsive Systems (2007) (to appear)
13. Zhang, G.X., Rong, H.N., Jin, W.D., Hu, L.Z.: Radar Emitter Signal Recognition Based on Resemblance Coefficient Features. In: Tsumoto,S., et al., (eds.): Lecture Notes in Artificial Intelligence, Vol.3066. Springer-Verlag, Berlin Heidelberg New York (2004) 665-670

Latency Estimation of the Asynchronous Pipeline Using the Max-Plus Algebra

Jian Ruan, Zhiying Wang, Kui Dai, and Yong Li

School of Computer, National University of Defense Technology
Changsha, Hunan, P.R. China

Abstract. This paper presents a methodology to estimate the latency of the asynchronous pipeline without choices. We propose modeling an asynchronous pipeline with the timed event graph and characterizing its specification via the max-plus algebra. An evolution equation of the event firing epoch can be obtained, which is linear under the max-plus algebraic formalism. In terms of the above-mentioned equation, we can formulate the latency of the pipeline successfully. The case study shows that our method is simple, fast and effective.

Keywords: Asynchronous Pipeline, Max-Plus Algebra, Timed Event Graph, Evolution Equation, Latency Estimation.

1 Introduction

Performance analysis of the asynchronous circuit, especially the asynchronous pipeline, is attracting lots of research interest. In [1], Ramamoorthy et al. modeled asynchronous concurrent systems via the timed marked graph with deterministic firing delays, and then used Karp's theorem to determine the average throughput in polynomial time. In [2], Kudva et al. proposed adopting Generalized Timed Petri Nets (GTPNs) to model asynchronous circuits in which component delays are modeled using their mean values. As usual, the model is then converted into a continues time Markov chain (CTMC) and solved for its stable probability distribution. In [3,4], Xie and Beerel proposed a discrete time model that can handle arbitrarily distributed delays. The model is converted into a discrete time Markov chain (DTMC) and solved for its stationary distribution. To mitigate the state explosion problem, they developed a technique called state compression to speed up the symbolic computation of the stationary distribution of their DTMC models. Specifically, they reduced the state space of the DTMC to one of the feedback vertex sets of its state space. Nevertheless, even with these advances, the state explosion problem has limited Markov chain based approaches to small models.

In this paper, we apply the timed event graph to model the asynchronous pipeline. Using the max-plus algebra to describe the specification, we can derive a linear evolution equation of the event firing times under the max-plus algebraic formalism. Based on the periodicity of iterated evolution equation, the latency can be formulated effectively and estimated efficiently.

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 251–258, 2007.
© Springer-Verlag Berlin Heidelberg 2007

The organization of this paper is as follows. Section 2 reviews the basics of the max-plus algebra. Section 3 describes the way to model asynchronous pipelines with the timed event graph. Section 4 illuminates the technique that how to estimate the latency of the asynchronous pipeline by its linear evolution equation in detail. Section 5 validates the conclusion by a 3-stage 4-phase bundled-data linear asynchronous pipeline and section 6 gives some conclusions.

2 Elementary Max-Plus Algebra

In this section we give an introduction to the max-plus algebra which will be used during the performance analysis of asynchronous pipelines. Most of the material presented is selected from [5,6], where a complete overview of the max-plus algebra can be found.

2.1 Basic Concepts

Definition 1. *The max-plus algebra* $\mathbb{R}_{\max}$ *is the set* $\mathbb{R}_\varepsilon \stackrel{def}{=} \mathbb{R} \cup \{-\infty\}$ *equipped with the two internal operations* $\oplus \stackrel{def}{=} \max$ *and* $\otimes \stackrel{def}{=} +$*, the neutral element for operation* $\oplus$ *is* $\varepsilon \stackrel{def}{=} -\infty$ *and the unit element for operation* $\otimes$ *is* $e \stackrel{def}{=} 0$.

$\mathbb{R}_{\max} = (\mathbb{R}_\varepsilon, \oplus, \otimes)$ is an idempotent commutative semi-ring, and the main difference with the conventional algebra $(\mathbb{R}, +, \times)$ is the fact that the operation $\oplus$ is idempotent, $a \oplus a = a$ and does not have an inverse.

Operations $\oplus$ and $\otimes$ can be extended to matrices with the classical construction,

1. $(A \oplus B)_{i,j} = A_{i,j} \oplus B_{i,j}$, where $A, B \in \mathbb{R}_\varepsilon^{m \times n}$;
2. $(A \otimes B)_{i,j} = \bigoplus_{k=1}^{p} (A_{i,k} \otimes B_{k,j})$, where $A \in \mathbb{R}_\varepsilon^{m \times p}$ and $B \in \mathbb{R}_\varepsilon^{p \times n}$;
3. $A^0 = I_n$, $A^k + 1 = A \otimes A^k = A^k \otimes A$, where $A \in \mathbb{R}_\varepsilon^{n \times n}$.[1]

2.2 Equation Solving

Solving linear equations in $(max, +)$ is different from the classical linear case. In general, it may or may not have a solution. Moreover, even if a solution exists, it may not be unique. However, there exists one class of linear equations for which we have a satisfactory theory.

Theorem 1. *Let* $A \in \mathbb{R}_\varepsilon^{n \times n}$, $B \in \mathbb{R}_\varepsilon^{n}$. *If matrix* A *is acyclic,* $A^* \stackrel{def}{=} \bigoplus_{i=0}^{\infty} A^i$ *converges and the vectorial linear equation of the form* $X = A \otimes X \oplus B$ *has a unique solution* $X = A^* \otimes B$.

Matrix $A \in \mathbb{R}_\varepsilon^{n \times n}$ is said to be acyclic iff $\exists (i_1, i_2, \ldots, i_{p-1}, i_p = i_1)$, s.t. $A_{i_1,i_2} \otimes A_{i_2,i_3} \otimes \ldots \otimes A_{i_{p-1},i_p} \neq \varepsilon$.

As for the recursion equation, the most attractive is not its solution but the periodicity.

[1] I_n is the $n \times n$ identity matrix with e on the main diagonal and ε elsewhere.

Theorem 2. *Let A be a $(max,+)$ irreducible matrix, there exists an unique real number λ, integer c and constant n_0, for all $n \geq n_0$, s.t. $A^{n+c} = \lambda^c \otimes A^n$.*

Matrix $A \in \mathbb{R}_{\varepsilon}^{n \times n}$ is said to be irreducible iff $\forall i,\ j,\ \exists (i,\ i_1,\ \ldots,\ i_p = j)$, s.t. $A_{i,i_1} \otimes A_{i_1,i_2} \otimes \ldots \otimes A_{i_{p-1},i_p} \neq \varepsilon$.

Corollary 1. *Let A be a $(max,+)$ irreducible matrix, the evolution equation $X(n+1) = A \otimes X(n)$ will end up in a periodic behavior, that is*

$$\exists n_0, c, \lambda, \quad \forall n \geq n_0, \quad X(n+c) = \lambda^c \otimes X(n).$$

3 Modeling the Asynchronous Pipeline

Event graph is widely used during the design, synthesis and verification of the asynchronous circuit[7] It describes the ordering of the events exactly, whereas the questions pertaining to when the events take place are not addressed. For questions related to the performance evaluation, it is necessary to introduce time.

In this section, we will expand on the way to model an asynchronous pipeline via timed event graph.

3.1 Timed Event Graph

Definition 2. *Petri net[8] is a bi-partite graph given by tuple $\mathcal{G} = (\mathcal{P}, \mathcal{Q}, \mathcal{E}, \mathcal{M}^0)$, where $\mathcal{P} = \{p_0, \cdots, p_m\}$ is the set of places, $\mathcal{Q} = \{q_0, \cdots, q_n\}$ is the set of transitions, $\mathcal{E} \subseteq (\mathcal{P} \times \mathcal{Q}) \cup (\mathcal{Q} \times \mathcal{P})$ is the set of arcs, $\mathcal{M}^0 : \mathcal{P} \to \{0, 1, 2, \cdots\}$ is the initial marking.*

We denote by $p^\bullet$ ($^\bullet p$) the set of downstream (upstream) transitions of place p, $q^\bullet$ ($^\bullet q$) the set of downstream (upstream) places of transition q.

Definition 3. *An event graph is a Petri net where each place has exactly one incoming transition and one outgoing transition.*

$$\forall p \in \mathcal{P}, |{}^\bullet p| = |p^\bullet| = 1$$

To analyze the latency of an asynchronous pipeline, we should consider the delay constraint. This can be done in two basic ways, by associating durations with either transition firings or with the sojourn of tokens in places[2].

1. Associating each place p with holding time σ_p. Holding time is the minimal time that a token must spend in a place before contributing to the enabling of the downstream transition, which can be used to represent the transportation or communication time. $\sigma_p(n)$ means the n-th holding time of place p.

[2] It is possible to convert a event graph with holding times into a event graph with firing times and vice-versa.

2. Associating each transition q with fireing time δ_q. Firing time is the time that elapses between the starting and the completion of the firing of the transition, which can be used to represent the process times in a functional block. $\delta_p\,(n)$ means the n-th firing time of transition p.

Now that the response times of the request and acknowledge between the adjacent pipeline stages are different, the former is much more appropriate.

3.2 Model of Asynchronous Pipeline

It is generally believed that an asynchronous pipeline can be viewed as a traditional "synchronous" data-path, consisting of latches and combinational circuits, which is clocked by some form of handshaking between neighbouring pipeline stages[7]. Taking a linear bundled-data asynchronous pipeline for example, whose structure is shown in Figure 1.

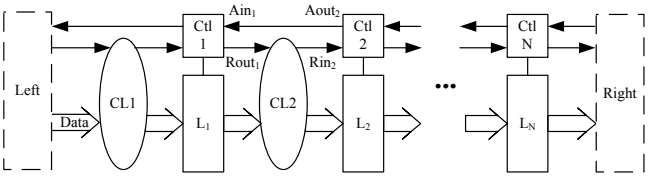


Fig. 1. Linear 4-phase bundled-data asynchronous pipeline

Since the latency of an asynchronous pipeline is only in reference to its structure, we may pass over the influence of the inputing. On the supposition that the environment deals with the request or acknowedge of the pipeline in no time, we may model the linear pipeline, along with its environment, as an autonomous timed event model, which is displayed in Figure 2.

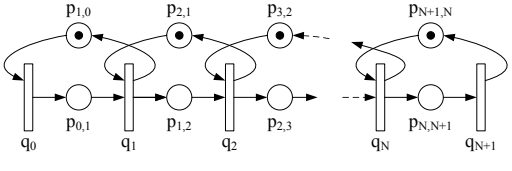


Fig. 2. Timed Event Graph Model

Transition q_0 stands for the left environment, its firing represents the inputing of new data. Transition q_{N+1} stands for the right environment, its firing represents that the computing result has been received. Moreover, transition

q_i $(0 < i \leq N)$ stands for the handshaking latch Hl_i of the pipeline[3], its firing represents that Hl_i is active.

Token in place $p_{i,i+1}$ represents that the input data is handled by stage $i+1$, holding time $\sigma_{i,i+1}$ amounts to the delay from new data on the input of stage $i+1$ to the production of the corresponding output (or to being received by the right environment), provided that the acknowledge signals are in place when data arrives. Token in place $p_{i,i-1}$ represents the reception of the acknowledgment produced by stage i, holding time $\sigma_{i,i-1}$ amounts to the delay from receiving an acknowledge from stage i to the corresponding acknowledge is produced by stage $i-1$ (or to the next request is produced by the left environment), provided that the request is in place when the acknowledge arrives[4].

4 Estimating the Latency of the Asynchronous Pipeline

4.1 Evolution Equation in the Max-Plus Algebra

Let us consider the autonomous timed event graph $\mathcal{G} = \left(\mathcal{P}, \mathcal{Q}, \mathcal{E}, \mathcal{M}^0, \sigma\right)$, we assume that,

1. $\mathcal{G}$ is *live* [5] and the initial marking $\mathcal{M}^0$ has at most one token per place,
2. For any place p, $(p \in \mathcal{P})$, the holding time σ_p is fixed[6].
3. For any token under the initial marking, it is available for the enabling of the downstream transition.

According to the firing rule of the timed event graph, transition i starts its n-th firing at time:

$$\begin{cases} X_i(1) = \max\left\{ \max\limits_{j \in {}^{\bullet\bullet}i} \left\{ X_j(1) + \sigma_{j,i} \middle| \mathcal{M}^0_{j,i} = 0 \right\}, 0 \right\} \\ X_i(n) = \max\limits_{j \in {}^{\bullet\bullet}i} \left\{ X_j\left(n - \mathcal{M}^0_{j,i}\right) + \sigma_{j,i} \right\} \end{cases}$$

Using the $(max, +)$ notation,

$$\begin{cases} X_i(1) = \bigoplus\limits_{j \in {}^{\bullet\bullet}i} \left\{ X_j(1) \otimes \sigma_{j,i} \middle| \mathcal{M}^0_{j,i} = 0 \right\} \oplus 0 \\ X_i(n) = \bigoplus\limits_{j \in {}^{\bullet\bullet}i} \left\{ X_j\left(n - \mathcal{M}^0_{j,i}\right) \otimes \sigma_{j,i} \right\} \end{cases}$$

This equation can be seen as a linear equation between the variables $X_i(n)$, with coefficients $\sigma_{j,i}$.

[3] Handshaking latch Hl_i is composed of conventional latch L_i and control circuit Ctl_i.
[4] The calculation of holding time differes from the circuit implementation styles.
[5] An event graph is live if and only if all circuits are marked under the initial marking.
[6] σ_p can be expressed as $\sigma_{i,j}$, if ${}^{\bullet}p = q_i$ and $p^{\bullet} = q_j$.

When written in vectorial form, it becomes

$$\begin{cases} X(1) = & A_0 \otimes X(1) \oplus \Gamma \\ X(n) = & A_0 \otimes X(n) \oplus A_1 \otimes X(n-1) \end{cases} \tag{1}$$

where $X(n)$ is the vector $\left[X_1(n), \cdots, X_N(n)\right]'$, Γ is a zero vector of size N and for $k \in \{0,1\}$, A_k is a $N \times N$ matrix defined by

$$\left(A_k\right)_{i,j} \stackrel{def}{=} \begin{cases} \sigma_{j,i} & \text{if } \mathcal{M}^0_{j,i} = k \\ \varepsilon & \text{otherwise.} \end{cases}$$

Since $\mathcal{G}$ is live, there is no empty circuit under its initial marking. As a consequence, matrix A_0 is acyclic. Applying Theorem 1, the unique solution of Equation (1) is:

$$\begin{cases} X(1) & = A_0^* \otimes \Gamma \\ X(n) & = A_0^* \otimes A_1 \otimes X(n-1) \stackrel{def}{=} \widetilde{A} \otimes X(n-1) \end{cases} \tag{2}$$

4.2 Evolution Equation of Asynchronous Pipeline

In terms of the timed event graph modeling of the linear asynchronous pipeline, we may get

$$A_0 = \begin{pmatrix} \varepsilon & \varepsilon & \dots & \varepsilon & \varepsilon \\ \sigma_{1,2} & \varepsilon & \dots & \varepsilon & \varepsilon \\ \varepsilon & \sigma_{2,3} & \dots & \varepsilon & \varepsilon \\ \vdots & & & & \\ \varepsilon & \varepsilon & \dots & \varepsilon & \varepsilon \\ \varepsilon & \varepsilon & \dots & \sigma_{N-1,N} & \varepsilon \end{pmatrix}, \quad A_1 = \begin{pmatrix} \varepsilon & \sigma_{2,1} & \varepsilon & \dots & \varepsilon \\ \varepsilon & \varepsilon & \sigma_{3,2} & \dots & \varepsilon \\ \varepsilon & \varepsilon & \varepsilon & \dots & \varepsilon \\ \vdots & & & & \\ \varepsilon & \varepsilon & \varepsilon & \dots & \sigma_{N,N-1} \\ \varepsilon & \varepsilon & \varepsilon & \dots & \varepsilon \end{pmatrix}$$

Using the rule of power operation,

$$\left(A_0^m\right)_{ij} = \begin{cases} \bigotimes_{k=j}^{i-1} \sigma_{k,k+1} & \text{if } i = j+m \\ \varepsilon & \text{otherwise.} \end{cases}$$

According the definition of A^* and $\widetilde{A}$, we can derive

$$\left(A_0^*\right)_{ij} = \begin{cases} \varepsilon & \text{if } i < j \\ e & \text{if } i = j \\ \sum_{k=j}^{i-1} \sigma_{k,k+1} & \text{if } i > j \end{cases}, \quad \widetilde{A}_{ij} = \begin{cases} \varepsilon & \text{if } i < j-1 \\ \sigma_{j,i} & \text{if } i = j-1 \\ \sum_{k=j-1}^{i-1} \sigma_{k,k+1} + \sigma_{j,j-1} & \text{if } i \geq j \end{cases}$$

Proposition 1. *With regard to the linear asynchronous pipeline, matrix* $\widetilde{A}$ *is irreducible.*

Proof. According to the expression of $\widetilde{A}$ shown above, we can discover that for all $i,\ j$,

1. if $i \geq j-1$, $\left(\widetilde{A}\right)_{ij} \neq \varepsilon$,
2. if $i < j-1$, there exists sequence $(i_1 = i,\ i_2 = i_1 + 1,\ \ldots,\ i_p = j)$, s.t. $\left(\widetilde{A}\right)_{i_1,i_2} \otimes \left(\widetilde{A}\right)_{i_2,i_3} \otimes \cdots \otimes \left(\widetilde{A}\right)_{i_{p-1},i_p} \neq \varepsilon$.

□

4.3 Latency of Asynchronous Pipeline

In this subsection, we will formulate the latency of a linear asynchronous pipeline by its evolution equation under the max-plus algebraic formalism.

Definition 4. *Latency is the delay between the input of a data item until the corresponding output data item is produced, i.e., the time difference between the firing of transition* q_0 *and that of transition* q_{N+1}.

$$T_{latency} = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \left[X_0(i) - X_{N+1}(i) \right] \tag{3}$$

According to Proposition 1 and Corollary 1, we can formulate the latency

$$T_{latency} = \frac{1}{c} \sum_{i=n_0}^{n_0+c} \left[X_0(i) - X_{N+1}(i) \right] \tag{4}$$

5 Case Study

Now we apply our method to a 3-stage linear asynchronous pipeline composed of 4-phase bundled-data pipeline stages, illustrating that the analysis is quite feasible.

The linear pipeline's control circuit is implemented by the semi-decoupled latch controller[9]. SPICE analysis has been performed on the latch controller's layout for worst case conditions (Vdd=4.6V, slow-slow process corner, at 100^oC). The data processing function between latch stages is modeled with an delay elment, where $Cl_1 = 21.3nS$, $Cl_2 = 12.4nS$ and $Cl_3 = 27.2nS$. According to the SPICE analysis results and the matching delays of the processing logic, we may calculate the holding times and estimate the latency, shown in Table 1. The FIFO represents a pipeline with no data processing logics between latch stages.

From Table 1, we may find that the estimated results of $(max, +)$ method are very close to those of the SPICE simulation. On the one hand, it lies in the precise holding time from the SPICE analysis of the 4-phase bundled-data latch control circuit. On the other hand, it confirms the truth of our method, which is uncomplicated and in effect.

Table 1. Latency Estimation

	Holding Time (nS)								Result (nS)	
	$\sigma_{0,1}$	$\sigma_{1,2}$	$\sigma_{2,3}$	$\sigma_{3,4}$	$\sigma_{1,0}$	$\sigma_{2,1}$	$\sigma_{3,2}$	$\sigma_{4,3}$	$(max,+)$	SPICE
FIFO	4.1	4.1	4.1	0	0	4.7	4.7	4.7	16.4	14.1
Pipeline	24.4	16.5	31.3	0	0	4.7	4.7	4.7	103.5	94.7

6 Conclusion

The contribution of this paper is introducing a method for estimating the latency of an asynchronous pipeline, which is effective and efficient.

Nevertheless, the technique has significantly more room for improvement. More sophisticated stochastic holding time sequences can be used to replace the fixed time so as to characterize the performance precisely.

Acknowledgment

This work is supported by the Chinese National Natural Science Foundation under grant No.90407022.

References

1. Ramamoorthy, C.V., Ho, G.S.: Performance evaluation of asynchronous concurrent systems using Petri nets. IEEE Transitions on Software Engineering. **6(5)** (1980) 440–449
2. Kudva, P., Gopalakrishnan, G., Brunvand, E., Akella, V.: Performance analysis and optimization of asynchronous circuits. Proc. International Conf. Computer Design (ICCD). (1994) 221–225
3. Xie, A., Beerel, P.A.: Accelerating Markovian analysis of asynchronous systems using state compression. IEEE Transactions on Computer-Aided Design. **18(7)** (1999) 869–888
4. Xie, A., Beerel, P.A.: Symbolic techniques for performance analysis of asynchronous systems based on average time separation of events. Proc. International Symposium on Advanced Research in Asynchronous Circuits and Systems (ASYNC). IEEE Computer Society Press. (1997) 64–75
5. Baccelli, B., Cohen, G., Olsder, G.J., Quadrat, J.-P.: Synchronization and Linearity. Wiley (1992)
6. Altman, E., Gaujal, B., Hordijk,A.: Discrete-Event Control of Stochastic Networks: Multimodularity and Regularity. Springer (2003)
7. Sparsø, J., Furber, S.: Principles of Asynchronous Circuit Design: A Systems Perspective. Kluwer Academic Publishers (2001)
8. Murata, T.: Petri nets: properties, analysis and applications. Proceedings of the IEEE, **77(4)** (1989) 541-580
9. Furber, S.B., Day, P.: Four-phase micropipeline latch control circuits. IEEE Trans. Very Large Scale Integr. Syst. **4(2)** (1996) 247–253

A Simulation-Based Hybrid Optimization Technique for Low Noise Amplifier Design Automation

Yiming Li[1], Shao-Ming Yu[2], and Yih-Lang Li[2]

[1] Department of Communication Engineering, National Chiao Tung University, 1001 Ta-Hsueh Rd., Hsinchu City 300, Taiwan
ymli@faculty.nctu.edu.tw
[2] Department of Computer Science, National Chiao Tung University, 1001 Ta-Hsueh Rd., Hsinchu City 300, Taiwan
smyu@mail.ymlab.org, ylli@cs.nctu.edu.tw

Abstract. In this paper, a simulation-based optimization technique for integrated circuit (IC) design automation is presented. Based on a genetic algorithm (GA), Levenberg-Marquardt (LM) method, and circuit simulator, a window-interfaced prototype of computer-aided design (CAD) is developed for IC design. Considering low noise amplifier (LNA) IC, we simultaneously evaluate specifications including S parameters, K factor, noise figure, and input third-order intercept point in the optimization process. If the simulated results meet the aforementioned constraints, the prototype outputs the optimized parameters. Otherwise, CAD activates GA for global optimization; simultaneously, LM method searches solutions with the results of GA. The prototype then calls a circuit simulator to compute and evaluate newer results until all specifications are matched. More than fifteen parameters including device sizes, passive components, and biasing conditions are optimized for the aforementioned constraints. For LNA IC with $0.18\mu m$ metal-oxide-silicon filed effect transistors, benchmark results confirm the functionality of the implemented prototype.

Keywords: Optimization, computational performance, circuit design, parameter tuning, DaCO, LNA.

1 Introduction

Computer-aided circuit simulation and analysis plays a crucial role in radio-frequency (RF) integrated-circuit (IC) design and implementation. Designers are requested to tune parameters of circuits (e.g., low noise amplifier (LNA)) to achieve a specified communication system. The parameters includes active and passive device parameters, device size, circuit layout, width of wires, and biasing condition. It may require a very experienced electronic engineer to accomplish complicated works; in particular, for RF complementary metal-oxide-semiconductor (CMOS) ICs [1, 2, 3, 4, 5]. Circuit simulation tool has been used in IC design in the past decades. Optimization techniques continuously benefit the communities of electronic design automation [6, 7, 8].

In this paper, based on hybrid evolutionary and gradient-based methods [6, 7, 8, 9, 10, 11, 12] and circuit simulation tool [13], we propose a simulation-based hybrid optimization technique for optimal circuit design with application to LNA ICs. This

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 259–266, 2007.
© Springer-Verlag Berlin Heidelberg 2007

optimization methodology was proposed for model parameter extraction of sub-100 nm CMOS devices and optimal characterization of heterojunction bipolar transistor in our recent works [6, 8]. This approach is generalized to IC design optimization; in particular, for analog and RF circuits. For a given LNA circuit, the hybrid optimization methodology simultaneously considers the electrical specifications [1,2,3] such as S_{11}, S_{12}, S_{21}, S_{22}, K factor, the noise figure, and the input third-order intercept point in the optimization process. First of all, preliminary parameters as well as the netlist for circuit simulation [6, 13] are loaded. A circuit simulation tool will be performed for the circuit simulation and then the results are used in the evaluation of specification. Once the specification meets the aforementioned seven constraints, the optimized parameters will be outputted. Otherwise, we activate the genetic algorithm (GA) for the global optimization; in the meanwhile, the Levenberg-Marquardt (LM) method searches the local optima according to the GA's results. The numerical optimization method does significantly accelerate the evolution process. We repeatedly call circuit simulator to compute and evaluate newer results until the specification is matched. A window-interfaced prototype of the computer-aided design (CAD), along with the proposed methodology, is successfully implemented and used for optimal design of LNA ICs. One of well-known circuit simulation tools, HSPICE [13], is successfully integrated in our CAD prototype. Other simulators, such as ELDO, SPECTRE, SmartSPICE can similarly be integrated by utilizing the command mode of these tools. To verify the validity of the developed prototype, several testing experiments are organized, where we simultaneously consider the characteristic optimization of passive and active devices, and power dissipation. More than fifteen parameters including device sizes, capacitance, inductance, resistance, and biasing conditions are optimized with respect to the aforementioned seven constraints. For the LNA with the $0.18\ \mu m$ CMOS technology, benchmark results including the convergence property and the sensitivity of optimized parameters computationally confirm the robustness and efficiency of the proposed simulation-based hybrid optimization technique.

This paper is organized as follows. In Sec. 2, we introduce the framework of the proposed optimization technique. In Sec. 3, the achieved results are discussed. Finally, we draw conclusions.

2 The Simulation-Based Hybrid Optimization Technique

In modern IC design, it is a real challenge to solve multidimensional optimization problems by using either conventional numerical method [6, 9, 10, 11, 12] or soft computing techniques [6, 7, 8]. Our idea of the simulation-based hybrid optimization technique takes a GA to perform global search, and while the evolution seems to be saturated, the LM method is then enhancing the searching behavior to perform the local search.

Two stages, shown in Fig. 1(a), are performed in the optimal design of ICs. The model parameters of CMOS transistors are first optimized [6]. For a set of given measured current-voltage (I-V) data of CMOS devices and a selected equivalent circuit model, the model parameters are optimized by using the developed hybrid intelligent optimization methodology [6]. With the optimized device parameters, we then move to the part of circuit design optimization. For a circuit to be optimized, we automatically

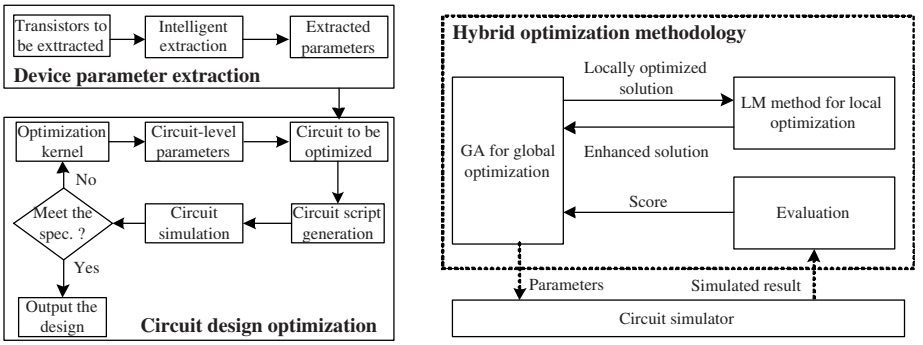


Fig. 1. (a) Functional blocks and (b) optimization flow for the proposed simulation-based hybrid optimization technique

parse and generate the corresponding netlist of the circuit for the circuit simulation and result evaluation. If the result meets the target, we then output the final optimized data. If the error between the target and result does not meet the stopping criterion, the established optimization kernel will enable the circuit parameter extraction in a global sense, where the number of parameters to be extracted depends upon the specification that we want to be achieved. According to the optimized results, the netlist is updated and the next optimization is repeated.

The GA, shown in Fig. 1(b), firstly searches the entire problem space. During this period, the GA-searched candidates are passed to an adopted circuit simulator [13], to retrieve the results of circuit simulation, where a set of ordinary differential equations is numerically solved. According to the specified targets of S_{11}, S_{12}, S_{21}, S_{22}, K factor, the noise figure, and the input third-order intercept point, the results are then evaluated

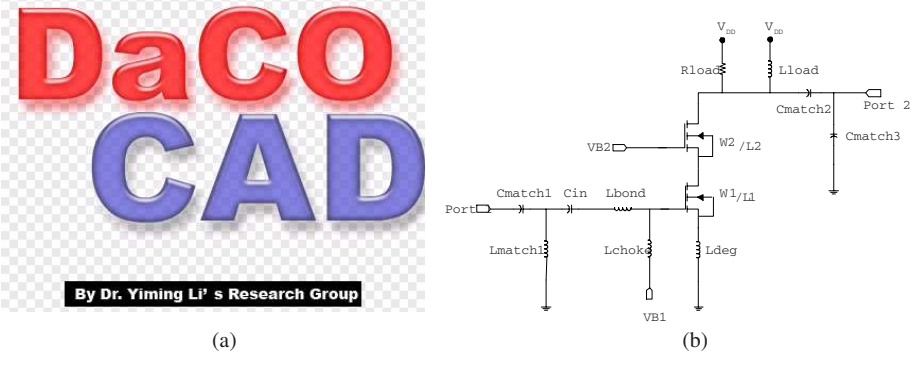


Fig. 2. (a) A logo of the developed CAD prototype and (b) The explored LNA circuit two cascade 0.18 μm MOSFETs

to measure the fitness score, which guides the evolutionary process of GA. Once a solution is obtained, the LM method will perform local searches. The local optima are right the GA's initial values for further optimization. The simulation-based hybrid optimization technique has successfully been implemented, where a window-based CAD prototype–DaCO (i.e., device and circuit optimization) is advanced, shown in Fig. 2(a).

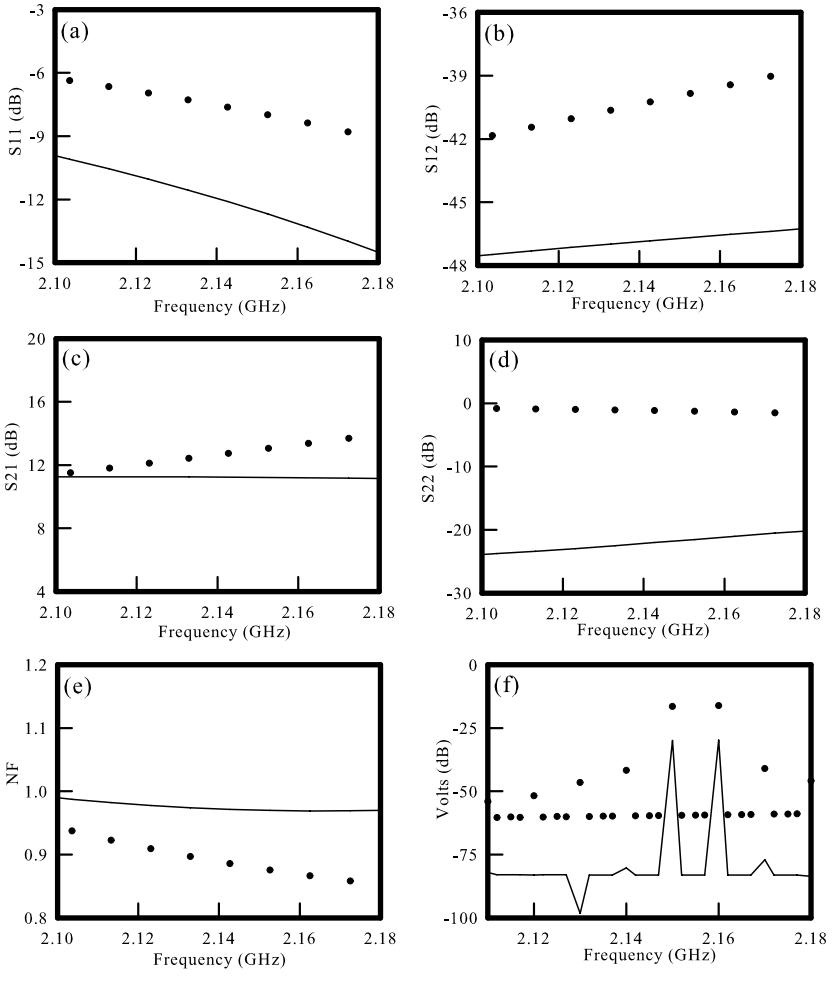


Fig. 3. The initial state (symbol) and a final optimized result (line) for the parameter of (a) S_{11}, (b) S_{12}, (c) S_{21} (d) S_{22}, (e) NF, and (f) IIP3, respectively

3 Results and Discussion

The explored LNA IC, shown in Fig. 2(b), focuses on the working frequencies ranging from 2.11GHz to 2.17GHz, where the Lload and Rload are the compact models of

on-chip spiral inductors. The choke inductor Lchoke working at high frequency is assumed to be fixed at 1uH and Cin is an external signal couple capacitor which is fixed at 20pF. The compact model BSIM 3v3 is adopted for the 0.18 μm metal-oxide-semiconductor field effect transistors (MOSFETs). There are more than fifteen parameters to be extracted in the designed 0.18 μm MOSFET LNA IC. Figure 3(a) shows the initial state (symbol) and an optimized result (line) of S_{11} parameter. The acceptable result is when $S_{11} < -10$ dB within the working frequency range. It is clearly that Fig. 3(a) states that the result has achieved to this goal. We note that the amplitude of the input sinusoidal signal within the working frequency range is with 1.0 V, VB1 = 0.75 V, and VB2 = 2.7 V shown in Fig. 2(b). Figure 3(b) is result of S_{12} parameter. The acceptable result is when $S_{12} < -25$ dB within the working frequency range. It confirms the achieved results with much more improvements than the original one. Figure 3(c) is the result of S_{21}, where a larger S_{21} is expected in the optimization process. Typically, we do not define an engineering specification for S_{21} in this testing case. However, a large S_{21} is good for optimal design of LNA ICs. Compared with the initial data, the obtained result is slightly shifted due to a compromise among all physical constraints at the same time. It could be further improved by performing more evolutions. Figure 3(d) shows the result of S_{22}. The goal is the same with the parameter S_{11}, (i.e., the result is acceptable if $S_{22} < -10$ dB within the working frequency range. Very good result, -20 dB, is achieved when we refer to the setting of standard goal. Once the S parameters are optimized, not shown here, we can also calculate the K factor accordingly [1,2,3]. Figure 3(e) shows the result of the noise figure (NF). The desired specification is that NF < 2 within the working frequency range. The obtained result is a little bit shifted away due to a global compromise among all electrical characteristics. Figure 3(f) indicates the initial state and an optimized result of the input third-order intercept point (IIP3). For the optimization criterion of IIP3, we ask that the amplitude of output $>$ -20dB and is as large as possible. As shown in Fig. 3(f), the optimized IIP3 = -26.

Considering low power consumption for portable communication systems, the VB1 and VB2 are two parameters to be optimized now. Table 1 shows the optimized parameters of the investigated experiment and the table 2 shows the corresponding optimized characteristics for the experiment. Results shown in both tables confirm the validity of the method. The sensitivity examination is now designed as follow. The prototype optimizes a category of parameter meanwhile locks other parameters. The LNA IC parameters to be optimized are classified into three categories illustrated in the table 3. It reveals that the geometry parameters, shown in Fig. 4(a), contribute the most improvement, while the input and output categories have little improvement after 120 generations. This phenomenon indicates that the geometry of the active devices is sensitive and more difficult to be optimized than passive device parameters. By considering the testing case, Fig. 4(b) shows a comparison of the score convergence behavior among population sizes, where the mutation rate is fixed at 0.5. The fitness score versus the number of generation suggests that the convergency does not have a satisfied result if the population size is too small. According to our experience, the population size = 50 is good for the optimal design of LNA IC. In addition, Fig. 5(a) shows the fitness score convergence behavior for the circuit optimization with different mutation rate,

Table 1. A list of the optimized parameters for the testing experiment

Element	Unit	Range	Test
Cmatch1	F	$300 \sim 800$	657.738 f
Cmatch2	F	$1 \sim 10$	4.505 p
Cmatch3	F	$1 \sim 10$	4.951 p
Lbond	H	$1 \sim 10$	1.058 n
Ldeg	H	$0.1 \sim 5$	1.155 n
Lmatch1	H	$1 \sim 10$	5.257 n
Rload	Ω	$1P5 \sim 5P5$	3P5.l
Lload	H	$1P5 \sim 5P5$	3P5.l
VB1	V	$0.5 \sim 1.5$	0.69 V
VB2	V	$0.5 \sim 5$	1.96 V
L	H	$0.13 \sim 0.3$	0.25 u

Table 2. A list of the extracted results

Specification	Target	Test
S_{11}	$<$ -10dB	-35.1dB
S_{22}	$<$ -10dB	-19.1dB
S_{12}	$<$ -25dB	-38.3dB
S_{21}	as large as possible	11.3dB
K	> 1	11.1
NF	< 2	1.17
IIP3	$>$ -10	0.3

Table 3. Three categories of the circuit parameter of the LNA IC

Category	Parameters
Geometry	L_1, W_1, L_2, and W_2
Input	Cmatch1, Lmatch1, Lbond, Lchoke, Cin, Ldeg, VB1, and VB2
Output	Lload, Rload, Cmatch2, and Cmatch3

where the population size = 50. The results suggest that the mutation = 0.5 keeps the population diversity and finally has better evolutionary results.

Finally, the computational efficiency of the proposed hybrid optimization technique is investigated. Figure 5(b) shows the score convergence behavior comparison of the standard GA and the hybrid optimization technique. The setting is with the population size = 50 and mutation = 0.5. As shown in this figure, the proposed methodology is superior to the pure GA after 60 generations. The proposed method shows no significant advantage at the beginning because the LM method has not been triggered yet. Once the LM method is activated, based on the result of GA to perform local optimization, the GA follows the local optima obtained by the LM method to keep evolving. Under this mechanism, our hybrid optimization methodology shows better trend of convergence and the robustness of our proposed methodology hence is held.

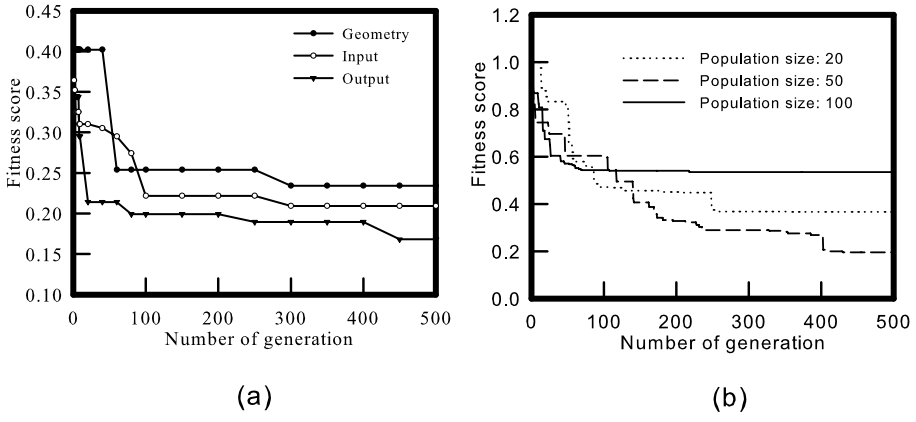


Fig. 4. (a) A verification of the sensitivity analysis for the three cataloged parameters. (b) The fitness score versus the number of generations with respect to different population size.

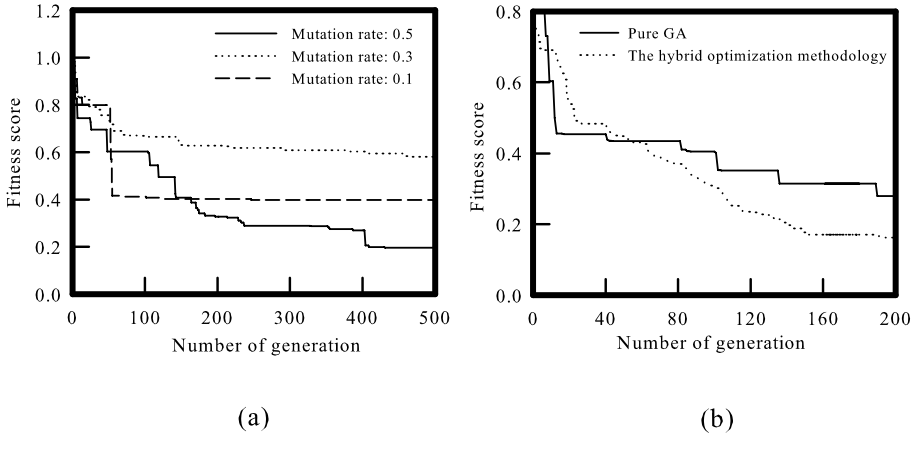


Fig. 5. (a) The fitness score versus the number of generations with respect to different (a) mutation rates and (b) for the pure GA and the proposed hybrid optimization methodology

4 Conclusions

In this paper, a simulation-based hybrid optimization technique for optimal design of LNA IC has been advanced. Based upon the GA, the LM, and one of the well-known circuit simulators, HSPICE, a CAD prototype named DaCO has successfully been implemented. Different testing cases for the designed LNA circuit with the 0.18 μm MOSFETs have been examined to show the validity, efficiency, and robustness of the method. This window-based CAD prototype can interface to any existing CAD

software and benefits advanced IC design and chip fabrication. The developed prototype is now available in the public domain [14].

Acknowledgments

This work was supported in part by the National Science Council of TAIWAN under Contract NSC-95-2221-E-009-336, Contract NSC-95-2752-E-009-003-PAE, and by the MoE ATU Program, Taiwan, under a 2006-2007 grant.

References

1. Pozar, D. M.: Microwave Engineering. John Wiley & Sons (2005)
2. Misra, D. K.: Radio-Frequency and Microwave Communication Circuits Analysis and Design. John Wiley & Sons (2004)
3. Leung, B.: VLSI for Wireless Communication. Prentice Hall (2002)
4. Gupta, R., Allstot, D. J.: Parasitic-aware design and optimization of CMOS RF integrated circuits. Dig. IEEE MTT-S Int. Microwave Symp., (1998) 1867–1870
5. Vancorenland, P., Van der Plas, G., Steyaert, M., Gielen, G., Sansen, W.: A layout-aware synthesis methodology for RF circuits. Proc. IEEE/ACM Int. Conf. Computer Aided Design (2001) 358–362
6. Li, Y., Cho, Y.-Y.: Intelligent BSIM4 Model Parameter Extraction for Sub-100 nm MOSFET Era. Jpn. J. Appl. Phys. **43** (2004) 1717–1722
7. Man, K. F., Tang, K. S., Kwong, S., Halang, W. A.: Genetic Algorithm for Control and Signal Processing. Springer (1997)
8. Li, Y., Cho, Y.-Y., Wang, C.-S., Huang, K.-Y.: A Genetic Algorithm Approach to InGaP/GaAs HBT Parameter Extraction and RF Characterization. Jpn. J. Appl. Phys.. **42** (2003) 2371-2374
9. Li, Y.: A Parallel Monotone Iterative Method for the Numerical Solution of Multidimensional Semiconductor Poisson Equation. Comput. Phys. Commun. **153** (2003) 359–372
10. Li, Y., Sze, S. M., Chao, T.-S.: A Practical Implementation of Parallel Dynamic Load Balancing for Adaptive Computing in VLSI Device Simulation. Eng. with Comput. **18** (2002) 124–137
11. Zhang, J. Z., Chen, L. H.: Nonmonotone LevenbergVMarquardt Algorithms and Their Convergence Analysis. J. Optim. Theory Appl. **92** (1997) 393–418
12. Han, Q. M.: A Levenberg-Marquardt method for semidefinite programming. J. Numer. Methods Comput. Appl. **19** (1998) 99–106
13. Shur, M.S., Fjeldly, T.A.: Silicon and Beyond, Advanced Circuit Simulators and Device Models. World Scientific Publishing Co. (2000)
14. DaCo – a prototype for device and circuit optimization. Available: http://140.113.87.143/ymlab/

Spectral Collocation Technique for Absorbing Boundary Conditions with Increasingly High Order Approximation

Zhenli Xu and Houde Han

Department of Mathematics, University of Science and Technology of China, Hefei, Anhui 230026, China
xuzl@ustc.edu

Abstract. An efficient treatment is developed for the Schrödinger equation with a class of local absorbing boundary conditions, which are obtained by high order Padé expansions. These boundary conditions are significant in the simulation of open quantum devices. Based on the finite difference approximation in the interior domain, we construct a spectral collocation layer on the cell near the artificial boundary, in which the wave function is approximated by the Chebyshev polynomials. The numerical examples are given by using this strategy with increasingly high order of accuracy up to the ninth order.

Keywords: Absorbing boundary conditions, Schrödinger equation, semiconductor devices, Chebyshev spectral collocation, difference schemes.

1 Introduction

For numerical simulations of quantum mechanical models for semiconductor devices, it is important and challenging to design an effective open boundary condition which absorbs outgoing waves [1]. In this paper, we consider high-order local absorbing boundary conditions (ABCs) for the Schrödinger equation

$$i\hbar\psi_t(x,t) = -\frac{1}{2}\hbar^2\Delta\psi(x,t) + V(x,t)\psi(x,t), \quad x \in \mathbb{R}^d, \tag{1}$$

which models quantum waveguides and resonant tunneling structures. Here we assume that the initial data $\psi(x,0) = f(x)$ is compactly support in a closed region Ω, and the given potential V is assumed to be a constant V_0 outside Ω so that there is no incoming wave on the truncated boundaries $\partial\Omega$.

One way is to construct the transparent boundary condition (TBC) by approximating the exact solution of the exterior problem restricted to the finite computational domain Ω. There are many papers [1,2,3,4,5,6,7,8] developing TBCs and studying their difference approximations and stability. However, the obtained TBC through this way is nonlocal in t, thus requiring all the history data in memory. Moreover, the computational effort of an ad-hoc discretization is unacceptable high.

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 267–274, 2007.
© Springer-Verlag Berlin Heidelberg 2007

To attain a method with low computational cost, another way is to construct ABCs [9,10,11,12,13] by approximating the nonlocal operator in TBC with polynomials. This class of boundary conditions is local in time, and easy to implement. In addition, they can be applied to nonlinear problems such as optical fiber communications [14] through a local time-splitting approach [15]. An efficient implementation of ABCs with high-order rational polynomial approximations is urgently required for the accurate solution of the Schrödinger equation (for a recent review paper, see Hagstrom [16]). However, it is also very difficult due to the high-order derivatives of the ABCs and their weak illposedness [12]. For this, we construct a spectral collocation layer on the cell near the boundary. The Chebyshev spectral method, which is very efficient for high-order equations [17,18], can be practically implemented in the layer, together with finite difference discretizations in interior domain.

The organization of this paper is as follows. In §2, high-order local ABCs are formulated. In §3, numerical issues are discussed. In §4, several numerical examples are given to illustrate the performance of the method. §5 gives some concluding remarks.

2 Construction of Absorbing Boundary Conditions

Consider the one-dimensional Schrödinger equation

$$i\psi_t(x,t) = -\psi_{xx}(x,t) + V(x,t)\psi(x,t), \tag{2}$$

which models the transient behavior of electrons in a quantum waveguide. Here the coefficients are absorbed into the derivatives for simplicity. We truncate the unbounded domain to a finite one $[-L, L]$ so that the initial value $f(x) = 0$ and the potential $V(x,t) = V_0$ for $|x| \geq L$.

We concentrate on the right boundary. Similar discussion can be performed on the left boundary. General wave packets propagating to the right at $x = L$ are represented by

$$\psi(x,t) = \int_{V_0}^{+\infty} e^{i(\sqrt{\omega - V_0}x - \omega t)}\hat{\psi}(x,\omega)d\omega, \tag{3}$$

where $\hat{\psi}$, satisfying $\hat{\psi}(L,\omega) = 0$ for $\omega < V_0$, denotes the Fourier transform in t under the dual relation $\omega \leftrightarrow i\partial_t$ between the frequency domain and the time domain. From Eq. (3), the TBC [1] in transformed space

$$i\hat{\psi}_x + \sqrt{\omega - V_0}\hat{\psi} = 0, \tag{4}$$

which annihilates all the outgoing waves, is given by

$$\psi_x(L,t) = -\sqrt{\frac{1}{\pi}}e^{-i(\frac{\pi}{4}+V_0 t)}\frac{d}{dt}\int_0^t \psi(L,\tau)e^{iV_0\tau}\frac{d\tau}{\sqrt{t-\tau}}. \tag{5}$$

It is a nonlocal boundary condition. In order to get local boundary conditions which allow a minor reflection, we approximate the square root $\sqrt{\omega - V_0}$ by using

polynomials or rational polynomials. Note that $k = \sqrt{\omega - V_0}$ is positive. The zero- and first-order Taylor approximations centered ω_0 with $k_0 = \sqrt{\omega_0 - V_0} > 0$ to the square root give

$$\sqrt{\omega - V_0} = k_0 + O(\omega - \omega_0),$$
$$\sqrt{\omega - V_0} = k_0 + \frac{\omega - V_0}{2k_0}$$

where the wavenumber parameter k_0 can be adaptively picked through a windowed Fourier transform [19]. These approximations at once lead to the first- and second order ABCs

$$B_1\psi = i\psi_x + k_0\psi = 0,$$
$$B_2\psi = i\psi_t + i2k_0\psi_x + (2k_0^2 - V_0)\psi = 0.$$

Higher-order Taylor approximations usually lead to ill-posed problems as shown for the hyperbolic wave equation in Engquist and Majda [20] and for the convection-diffusion equation in Halpern [21], in which the authors showed that a hierarchy of Padé approximations shall lead to well-posed problems. For the Schrödinger equation (2), the (n, n)-Padé approximations to $\sqrt{\omega - V_0}$ can be deduced through a recursive relation, as

$$\sqrt{\omega - V_0} = k_0 P_n/Q_n + O(\omega - \omega_0)^{2n+1}, \tag{6}$$

with $P_0 = Q_0 = 1$ and

$$P_n = (\omega - V_0 + k_0^2)P_{n-1} + 2(\omega - V_0)Q_{n-1},$$
$$Q_n = (\omega - V_0 + k_0^2)Q_{n-1} + 2k_0^2 P_{n-1}.$$

And for $(n+1, n)-$Padé approximations, the recursive relation is

$$\sqrt{\omega - V_0} = k_0 P_n/Q_n + O(\omega - \omega_0)^{2n+2}, \tag{7}$$

with $P_0 = \omega - V_0 + k_0^2$, $Q_0 = 2k_0^2$ and

$$P_n = (\omega - V_0 + k_0^2)P_{n-1} + 2(\omega - V_0)Q_{n-1},$$
$$Q_n = (\omega - V_0 + k_0^2)Q_{n-1} + 2k_0^2 P_{n-1}.$$

We let $\{\tilde{P}_n, \tilde{Q}_n\}$ represent the dual operators in physical space of $\{P_n, Q_n\}$ in frequency domain. Then after transforming Eqs. (6) and (7) back to the physical space by using the dual relation and using Eq. (4), we obtain high order approximations to the TBC: for $(2n+1)$-st order ABC by the (n, n)-Padé approximation,

$$B_{2n+1}\psi = (i\tilde{Q}_n\partial_x + k_0\tilde{P}_n)\psi = 0, \tag{8}$$

with $\tilde{P}_0 = \tilde{Q}_0 = 1$, and

$$\tilde{P}_n = (i\partial_t - V_0 + k_0^2)\tilde{P}_{n-1} + 2(i\partial_t - V_0)\tilde{Q}_{n-1}, \tag{9}$$
$$\tilde{Q}_n = (i\partial_t - V_0 + k_0^2)\tilde{Q}_{n-1} + 2k_0^2\tilde{P}_{n-1}; \tag{10}$$

and, for $(2n+2)$-nd order ABC by the $(n+1,n)$-Padé approximation,

$$B_{2n+2}\psi = (i\tilde{Q}_n\partial_x + k_0\tilde{P}_n)\psi = 0, \tag{11}$$

with $\tilde{P}_0 = i\partial_t - V_0 + k_0^2$, $\tilde{Q}_0 = 2k_0^2$, and the same recursive formulae as (9) and (10). For the $(n,n+1)$-Padé approximation, we remark that the obtained ABCs are ill-posed, as is discussed in Alonso-Mallo and Reguera [12].

3 Numerical Approximation

3.1 Discretization of Interior Domain

Denote the approximation of ψ on the grid point (x_j,t^n) by ψ_j^n for $0 \le j \le J$, with $x_j = x_0 + j\Delta x$ and $t^n = n\Delta t$. In the interior domain, the Schrödinger equation is approximated by the Crank-Nicholson scheme

$$i\frac{\psi_j^{n+1}-\psi_j^n}{\Delta t} = (-D^+D^- + V_j^{n+\frac{1}{2}})\frac{\psi_j^{n+1}+\psi_j^n}{2}, \quad 1 \le j \le J-1, \tag{12}$$

where D^+ and D^- represent the forward and backward differences, respectively.

3.2 Order Reduction for Time Derivatives of Boundary Conditions

In order to numerically discrete the boundary conditions with high order derivatives, it is necessary to reduce the time derivatives to first order. Thus, a two-layer scheme can be easily performed. For this, we use the original equation $i\psi_t = -\psi_{xx} + V_0\psi$ to substitute ∂_t with terms ψ_{xx} and ψ. For example, for the (n,n)-Padé approximations, the boundary condition $B_3\psi = 0$ is of the first order with respect to t when $n = 1$. And when $n > 1$, we can write the operators $\tilde{P}_n$ and $\tilde{Q}_n$ as

$$\tilde{P}_n = (-\partial_{xx} + k_0^2)\tilde{P}_{n-1} - 2\partial_{xx}\tilde{Q}_{n-1}, \quad \tilde{Q}_n = (-\partial_{xx} + k_0^2)\tilde{Q}_{n-1} + 2k_0^2\tilde{P}_{n-1}. \tag{13}$$

The boundary condition $B_{2n+1}\psi = 0$ then keeps first order with respect to t. Therefore, the midpoint rule can be easily achieved as a time discretization method.

3.3 Approximation of Spectral Collocation Layers

We now concentrate on the discretizations of boundary conditions, in which we shall use a technique of the spectral collocation method (for the theory and application of spectral methods, see monographs [22,23] for details). The collocation points are distributed on the cell near the artificial boundary; i.e., on $[x_0, x_1]$ at the left boundary and $[x_{J-1}, x_J]$ at the right. We consider the right boundary again. The approximation of difference scheme (12) at x_{J-1} can be regarded as a boundary condition of the right collocation domain, and thus the collocation domain is well-posed together with the ABC (8) or (11) at x_J.

To introduce the spectral collocation method, we transform the domain $[x_{J-1}, x_J]$ into $[-1, 1]$ by using scaling

$$\xi = s[x - (x_J + x_{J-1})/2], \text{ with } s = 2/(x_J - x_{J-1}). \tag{14}$$

Eq. (2) near the boundary is then written as

$$i\tilde{\psi}_t(\xi, t) = -s^2\tilde{\psi}_{\xi\xi}(\xi, t) + V_0\tilde{\psi}(\xi, t), \tag{15}$$

where $\tilde{\psi}(\xi, t) = \psi(x, t)$. Similarly, the derivatives in the recursive formulae of ABCs can be scaled. Let $T_m(\xi) = \cos(m \arccos(\xi))$ for $0 \le m \le M$ be the Chebyshev polynomials, and $\xi_m = \cos(m\pi/M)$ be the Chebyshev-Gauss-Labatto collocation points. The Chebyshev interpolating polynomial in space with M collocation points reads

$$(\mathcal{P}_M\tilde{\psi})(\xi, t) = \sum_{m=0}^{M} \tilde{\psi}_m(t)\phi_m(\xi),$$

where $\tilde{\psi}_m(t)$ represents the point value of $\tilde{\psi}(\xi, t)$ at ξ_m, and the basis function is

$$\phi_m(\xi) = \frac{2}{\alpha_m M} \sum_{l=0}^{M} \frac{1}{\alpha_l} T_l(\xi_m)T_l(\xi),$$

with $\alpha_m = 1$, except for $\alpha_0 = \alpha_M = 2$. If we approximate the p-th spatial derivatives of $\tilde{\psi}(\xi, t)$ at the collocation points, ξ_m, we have

$$\tilde{\boldsymbol{\psi}}^{(p)} = D^{(p)}\tilde{\boldsymbol{\psi}} = D^p\tilde{\boldsymbol{\psi}},$$

where $\tilde{\boldsymbol{\psi}}$ denotes the vector $\tilde{\boldsymbol{\psi}} = [\tilde{\psi}(\xi_0), \tilde{\psi}(\xi_1), \cdots, \tilde{\psi}(\xi_M)]^T$. The corresponding differentiation matrix D has the entries

$$\begin{aligned} D_{ml} &= \frac{\alpha_m}{\alpha_l} \frac{(-1)^{m+l}}{\xi_m - \xi_l}, \qquad m \ne l, \\ D_{mm} &= -\frac{\xi_m}{2(1 - \xi_m^2)}, \qquad 0 < m < M \\ D_{00} &= -D_{MM} = \frac{2M^2 + 1}{6}. \end{aligned}$$

Using the above approximations in space and the midpoint rule in time, the equation in the collocation layer can be expressed as a system of algebraic equations which is connected with scheme (12) in the interior domain at $x = x_{J-1}$. The formed algebraic system is a modified triangular linear matrix which contains two blocks in the upper left and lower right corners.

4 Numerical Examples

In this section, we shall give some numerical examples for the Schrödinger equation (2) in homogeneous media; i.e., we set $V \equiv 0$.

Example 1. In the first example, we are going to consider the evolvement of the right travelling Gaussian beam $\psi(x,0) = e^{-x^2}e^{iqx}$ with $q = 5$ in a truncated region $[-10,\ 10]$. Similar tests can be found in[12] and in[24], which simulated optical beam propagation in the Fresnel approximation. By taking the parameter $k_0 = q$ and $2q$ respectively in the calculations, we compute the solution up to $T = 10$, and calculate the reflection ratio $r = \sum_{j=0}^{J} |\psi_j^n|^2 / \sum_{j=0}^{J} |\psi_j^0|^2$ at time $t^n = T$. In Tables 1 and 2, we show the results of the reflection ratios for different boundary conditions with increasing high order approximation for (n,n)-Padé and $(n+1,n)$-Padé polynomials, respectively. We see that a higher order boundary condition has usually a better effect of absorbing outgoing waves. The instability of the numerical solution also increases with the order of the ABCs. For the eighth and ninth order ABCs, blowups appear; and the behavior is worse, the smaller grid size is. However, these results are reasonable due to the weak ill-posedness of the high order ABCs which induce ill-conditioned matrices. Therefore, a better solver is desired to the linear systems.

Table 1. Reflection ratios with $\Delta x = \Delta t$ for the difference scheme, and M collocation points for the spectral collocation layer

		$k_0 = 5$				$k_0 = 10$			
Δx	M	Padé-(1,1)	(2,2)	(3,3)	(4,4)	(1,1)	(2,2)	(3,3)	(4,4)
0.1	13	7.58e-5	5.69e-5	5.49e-5	5.38e-5	1.87e-4	7.63e-5	6.33e-5	5.78e-5
	17	7.58e-5	5.69e-5	5.49e-5	Blowup	1.87e-4	7.63e-5	6.33e-5	7.17e-5
0.05	13	1.59e-5	5.60e-6	3.89e-6	Blowup	9.19e-5	2.50e-5	1.17e-5	Blowup
	17	1.59e-5	5.60e-6	3.89e-6	Blowup	9.19e-5	2.50e-5	1.16e-5	Blowup

Table 2. $(n+1,n)$-Padé approximations with the same settings as Table 1

		$k_0 = 5$				$k_0 = 10$			
Δx	M	Padé-(1,0)	(2,1)	(3,2)	(4,3)	(1,0)	(2,1)	(3,2)	(4,3)
0.1	13	1.01e-4	5.90e-5	5.54e-5	5.39e-5	1.48e-3	9.35e-5	6.82e-5	6.01e-5
	17	1.01e-4	5.90e-5	5.54e-5	Blowup	1.48e-3	9.35e-5	6.82e-5	5.84e-5
0.05	13	3.98e-5	8.47e-6	4.41e-6	Blowup	3.24e-4	4.25e-5	1.65e-5	1.11e-5
	17	3.98e-5	8.47e-6	4.42e-6	Blowup	3.24e-4	4.25e-5	1.65e-5	Blowup

Example 2. The second example is for the initial data

$$\psi(x,0) = e^{-x^2} \int_0^5 (e^{iqx} + e^{-iqx})/2dq = e^{-x^2} \sin(5x)/x$$

in a truncated region $[-10,\ 10]$, which is composed of waves with different group velocities. We compute the solution up to $T = 2$. The parameter k_0 in the boundary conditions is now a function of time; we take a linear function $k_0(t) = (11-t)/2$, because at time t the components with $\pm k_0(t)$ wavenumbers arrive at the boundaries. In the calculations, we take $M = 13$, $\Delta x = 0.05$ and $\Delta t = 0.005$. The

numerical solution with the same mesh sizes in a large domain $[-40, 40]$ is taken to be a reference "exact" solution, since the analytic solution is unknown. The numerical solutions for ABCs up to seventh order are illustrated in Fig. 1. It is a good example to show why high-order ABCs are required. We see from the curves that lower-order ABCs reflect many outgoing waves which pollute the solution in the computational domain, and that higher-order ABCs demonstrate better results.

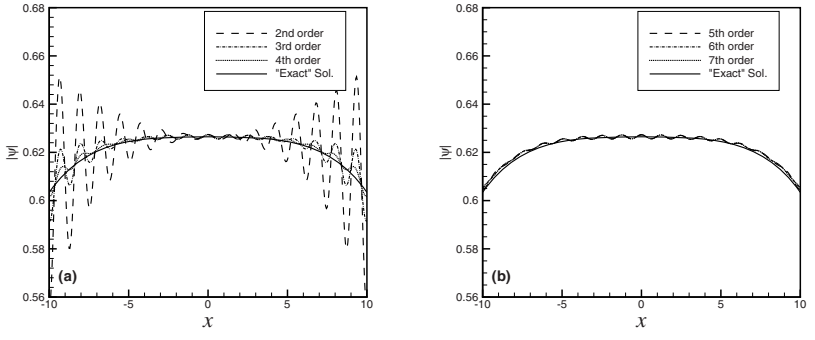


Fig. 1. Numerical results of wave packets $|\psi|$ at $t = 2$ for different ABCs

5 Concluding Remarks

High-order absorbing boundary conditions (ABCs) for the linear Schrödinger equation are implemented through using a spectral collocation technique. Numerical results illustrate the attractive performance of higher-order cases. However, the instability appears with the increase of the order, which motivates further investigation on the improvement of the approach. The method is outlined for one-dimensional problem but the extensions to multidimensional cases of the technique are straightforward in Cartesian grids. Even so, a detailed report for multidimensional problems is also necessary, which is under consideration as a possible part of the future research.

Acknowledgments. This work is supported by the National Natural Science Foundation of China (Grant Nos. 10471073 and 40674037).

References

1. Arnold, A.: Numerically absorbing boundary conditions for quantum evolution equations. VSLI Design **6** (1998) 313–319
2. Schmidt, F., Yevick, D.: Discrete transparent boundary conditions for Schrödinger-type equations. J. Comput. Phys. **134** (1997) 96–107
3. Ehrhardt, M.: Discrete transparent boundary conditions for general Schrödinger-type equations. VLSI Design **9**(4) (1999) 325–338

4. Arnold, A., Ehrhardt, M., Sofronov, I.: Discrete transparent boundary conditions for the Schrödinger equation: Fast calculation, approximation, and stability. Commun. Math. Sci. **1** (2003) 501–556
5. Antoine, X., Besse, C.: Unconditionally stable discretization schemes of non-reflecting boundary conditions for the one-dimensional Schrödinger equation. J. Comput. Phys. **188** (2003) 157–175
6. Han, H., Huang, Z.: Exact artificial boundary conditions for Schrödinger equation in R^2. Commun. Math. Sci. **2** (2004) 79–94
7. Han, H., Jin, J., Wu, X.: A finite-difference method for the one-dimensional time-dependent Schrödinger equation on unbounded domain. Comput. Math. Appl. **50** (2005) 1345–1362
8. Sun, Z.Z., Wu, X.: The stability and convergence of a difference scheme for the Schrödinger equation on an infinite domain by using artificial boundary conditions. J. Comput. Phys. **214** (2006) 209–223
9. Shibata, T.: Absorbing boundary conditions for the finite-difference time-domain calculation of the one dimensional Schrödinger equation. Phys. Rev. B **43** (1991) 6760
10. Kuska, J.P.: Absorbing boundary conditions for the Schrödinger equation on finite intervals. Phys. Rev. B **46** (1992) 5000
11. Fevens, T., Jiang, H.: Absorbing boundary conditions for the Schrödinger equation. SIAM J. Sci. Comput. **21** (1999) 255–282
12. Alonso-Mallo, I., Reguera, N.: Weak ill-posedness of spatial discretizations of absorbing boundary conditions for Schrödinger-type equations. SIAM J. Numer. Anal. **40** (2002) 134–158
13. Szeftel, J.: Design of absorbing boundary conditions for Schrödinger equations in R^d. SIAM J. Numer. Anal. **42** (2004) 1527–1551
14. Agrawal, G.: Nonlinear fiber optics, 3rd Ed. Academic Press, San Diego (2001)
15. Xu, Z., Han, H.: Absorbing boundary conditions for nonlinear Schrödinger equations. Phys. Rev. E **74** (2006) 037704
16. Hagstrom, T.: New results on absorbing layers and radiation boundary conditions. Lect. Notes Comput. Sci. Eng. **31** (2003) 1–42
17. Elbarbary, E.M.E., El-Sayed, S.M.: Higher order pseudospectral differentiation matrices. Appl. Numer. Math. **55** (2005) 425–238
18. Mai-Duy, N.: An effective spectral collocation method for the direct solution of high-order ODEs. Commun. Numer. Meth. Eng. **22** (2006) 627–642
19. Xu, Z., Han, H., Wu, X.: Adaptive absorbing boundary conditions for Schrödinger-type equations: application to nonlinear and multi-dimensional problems. J. Comput. Phys., to appear (2007) arXiv: math.NA/0610642.
20. Engquist, B., Majda, A.: Radiation boundary conditions for acoustic and elastic wave calculations. Commun. Pure Appl. Math. **32** (1979) 313–357
21. Halpern, L.: Artificial boundary conditions for the linear advection diffusion equation. Math. Comput. **46**(174) (1986) 425–438
22. Trefethen, L.N.: Spectral Methods in MATLAB. SIAM: Philadelphia, PA (2000)
23. Shen, J., Tang, T.: Spectral and High-Order Methods with Applications. Science Press, Beijing (2006)
24. Yevick, D., Yu, J.: Optimal absorbing boundary conditions. J. Opt. Soc. Am. A **12** (1995) 107–110

Shockwave Detection for Electronic Vehicle Detectors

Hsung-Jung Cho* and Ming-Te Tseng

Department of Transportation Technology and Management,
National Chiao Tung University,
Taiwan
hjcho@cc.nctu.edu.tw

Abstract. Although shockwaves have been extensively adopted in traditional traffic flow theory, how to detect shockwaves using an electronic vehicle detector has not been explored. Therefore, this study illustrates, for the first time, not only how to detect shockwaves, but also how to obtain shockwaves from three new traffic parameters: Stopped, Moving, and Empty. The Stopped parameter attempts to identify a newly arrived shockwave equation when a traffic queue approaches the electronic vehicle detector. The Moving and Empty parameters derive another new arrival shockwave equation when the electronic vehicle detector fails to identify any queue. An algorithm is also created to demonstrate how to use these parameters and equations to detect shockwaves. Additionally, numerous simulations are conducted to identify the behaviors of new traffic parameters and the effectiveness of the proposed algorithm. Results of this study demonstrate that the computing algorithm for electronic vehicle detectors can accurately detect shockwaves.

Keywords: Shockwave, electronic vehicle detector, stopped, moving, empty.

1 Introduction

Shockwave analysis has long been adopted in traffic flow analysis.[1] Shockwaves are defined as boundary conditions in the time-space domain that indicate a discontinuity in flow-density conditions.[2] Shockwaves result from sudden temporal or spatial changes in roadway density due to capacity or volume changes. Shockwave analysis is an effective means of analyzing flow and queuing problems.[3,4] Numerous techniques have been proposed via which shockwaves can be plotted and employed to forecast traffic system performance.[1] Applications to freeway bottlenecks and to traffic signals have also been developed.[5,6,7]

Most previous analyses and applications obtain shockwave values either via direct provision or by computation based on changes in flow and density (shockwave, $w = \Delta q / \Delta k$). No previous investigation has studied methods for direct shockwave detection for electronic vehicle detectors. This investigation is the first to illustrate a method for shockwave detection for electronic vehicle detectors. Moreover, this investigation obtains, for the first time, shockwaves from three new traffic parameters

* Corresponding author.

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 275–282, 2007.
© Springer-Verlag Berlin Heidelberg 2007

besides flow and density. The three new traffic parameters are Stopped, Moving, and Empty. Two shockwave equations are derived from these three parameters. When a traffic queue reaches an electronic vehicle detector, the shockwave equation can be derived from the Stopped parameter. When no traffic queue is present at an electronic vehicle detector, the Moving and Empty parameters are used to obtain another shockwave equation. An algorithm is also devised to demonstrate how to use these parameters and equations. Finally, CORSIM simulations were run for different combinations of scenarios to identify the behavior of new traffic parameters and the effectiveness of the computing algorithm.

2 Formulation

The formulation comprises four components. The first is the definitions of three new traffic parameters: Stopped, Moving and Empty. The second is the use of Stopped to detect shockwaves. The third is the use of Moving and Empty to detect shockwaves. The fourth is the algorithm, which combines the three new traffic parameters and two shockwave equations for shockwave detection.

2.1 Three Traffic Parameters: Stopped, Moving and Empty

Figure 1 provides the definitions of Stopped, Moving and Empty. Stopped refers to the detection of a vehicle that continues beyond a specified time interval. Moving means that the presence of a vehicle continues for less than a specified time interval. Empty means that no presence of a vehicle is detected during a specified time interval. If the specified time interval is short and the traffic flow is high, the short Moving and short Empty states alternate. Figure 1 shows the high sensitivity electronic detector. If the specified time interval is long and the traffic flow is high, the situation alternates between the long Stop, long Moving, and short Empty states. Figure 1 also illustrates the low sensitivity electronic detector. An electronic detector can have a low sensitivity, high sensitivity or both.

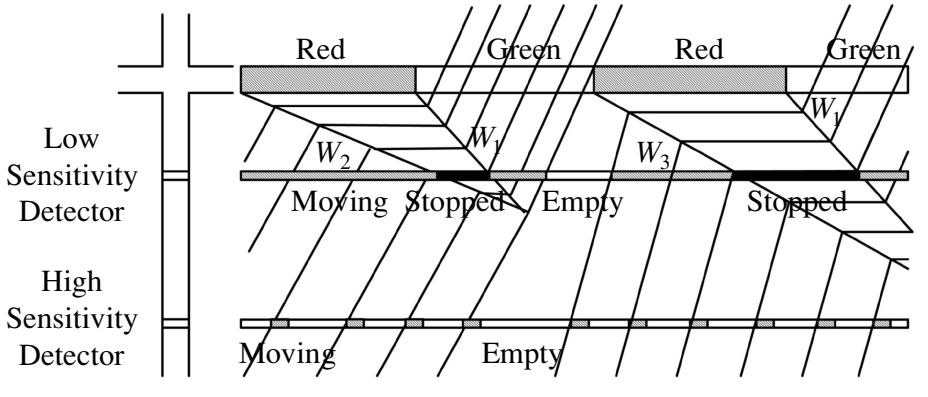


Fig. 1. Definitions of Stopped, Moving and Empty at a signalized intersection

2.2 Using Moving and Empty Time for Shockwave Detection

When a high sensitivity electronic detector receives only Moving and Empty time, the traffic queue does not reach the electronic detector. Figure 2 illustrates the Moving time, vehicle and detection zones for such a traffic condition. Two vehicles with speed V and length L_V pass through a high sensitivity electronic detector with zone length L_Z . Each vehicle triggers the moving time as $(L_v + L_z)/v$. Intuitively, the summary of the Moving time is related to traffic flow. Hence, the arrival shockwave W_3 during time interval ΔT can be expressed as

$$W_3 = \frac{L_{Queue}}{\Delta T} = (\sum_n \frac{L_v + L_z}{v}) \frac{C}{Moving + Empty} = C \frac{Moving}{Moving + Empty}. \quad (1)$$

where C is a constant given by

$$C = \frac{v(AvgL_V + AvgL_{gap})}{AvgL_V + L_Z}. \quad (2)$$

where L_{Queue} denotes the queue length for arriving vehicles, $AvgL_V$ represents the average vehicle length and $AvgL_{gap}$ is the average gap length between two stopped vehicles.

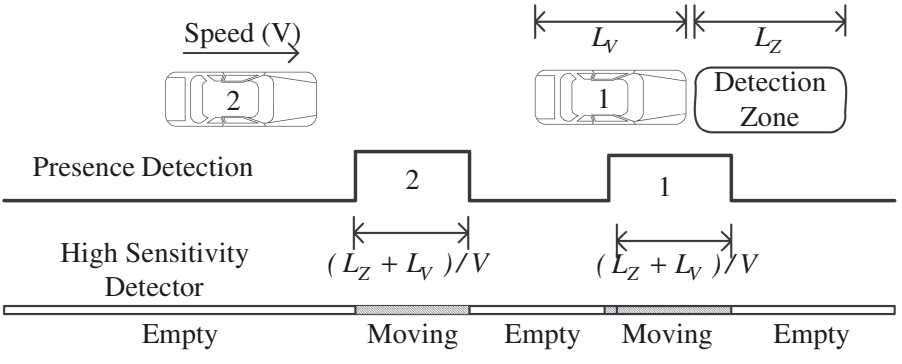


Fig. 2. Moving time, vehicles and detection zone

2.3 Using Stopped Time for Shockwave Detection

When traffic flow increases, traffic queue length gradually increases. If the traffic queue extends back to the position of the electronic vehicle detector during a red phase of a signalized intersection, the Stopped time is triggered. Generally, the Stopped time of the electronic detector changes slowly for continued signal cycles. Figure 3 displays the Stopped time for two continued signal cycles. The Stopped time of the first cycle is $\overline{OA}$, while that of the second cycle is $\overline{OC}$. The Stopped difference

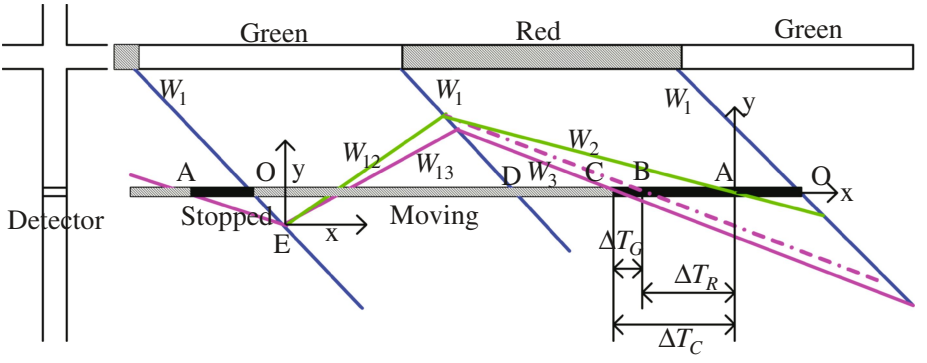


Fig. 3. Arrival shockwave and Stopped time

for these two continued cycles is $\overline{AC}$, or ΔT_C. Meanwhile, $\overline{AC}$ is the sum of $\overline{AB}$ and $\overline{BC}$. Furthermore, $\overline{BC}$, or ΔT_G, results from the difference in flow between shockwave W_{12} and shockwave W_{13} during the green phase G. Moreover, $\overline{AB}$, or ΔT_R, results from the flow difference between shockwave W_2 and shockwave W_3 during the red phase R. Furthermore, W_3 denotes the shockwave of an arrival flow, while W_2 represents the shockwave of a special arrival flow in which total queued vehicles during the red phase can be discharged based only on the green phase time (in Fig.1). W_1 is the shockwave of the saturated discharge flow.

Therefore

$$\Delta T_C = \Delta T_G + \Delta T_R . \tag{3}$$

Let point A be (0, 0) and set the x-y coordinate axis as in Fig.3. By using linear algebra,

$$\Delta T_R = \frac{W_1 R(W_3 - W_2)}{W_3(W_1 - W_2)} . \tag{4}$$

Similarly, let point E be (0, 0) and set the x-y coordinate axis. Thus,

$$\Delta T_G = \frac{W_1 G(W_1 - W_3)(W_{12} - W_{13})}{W_3(W_1 + W_{12})(W_1 + W_{13})} . \tag{5}$$

Suppose the traffic flow follows the model of Greenshield [8]:

$$U = U_f(1 - \frac{K}{K_j}) . \tag{6}$$

where U denotes speed, U_f represents free flow speed, K is density, and K_j denotes traffic jam density. The following equations are then obtained:

$$W_{12} = \frac{RW_1W_2}{GW_1 - CW_2}; W_{12} = U_f - W_1 - W_2; W_{13} = U_f - W_1 - W_3. \quad (7)$$

Substituting W_{12} and W_{13} (Eq.7) into Eq.(5) yields the new ΔT_G:

$$\Delta T_G = \frac{(GW_1 - CW_2)^2(W_3 - W_1)(W_3 - W_2)}{(W_1 - W_2)W_3((GW_1 - CW_2)W_3 + CW_2^2 - GW_1^2)}. \quad (8)$$

To summarize, the following equation can be obtained.

$$\Delta T_C = \frac{(GW_1 - CW_2)^2(W_3 - W_1)(W_3 - W_2)}{(W_1 - W_2)W_3((GW_1 - CW_2)W_3 + CW_2^2 - GW_1^2)} + \frac{W_1R(W_3 - W_2)}{W_3(W_1 - W_2)} \quad (9)$$

where parameters ΔT_C, G, R, W_1, and W_2 are all constants. Thus it is easy to obtain $W_3 = f(\Delta T_C, G, R, W_1, W_2)$.

2.4 Algorithm

The previous subsections propose two arrival shockwave equations. Equation (1) demonstrates how to use Moving and Empty time to detect arrival shockwave.

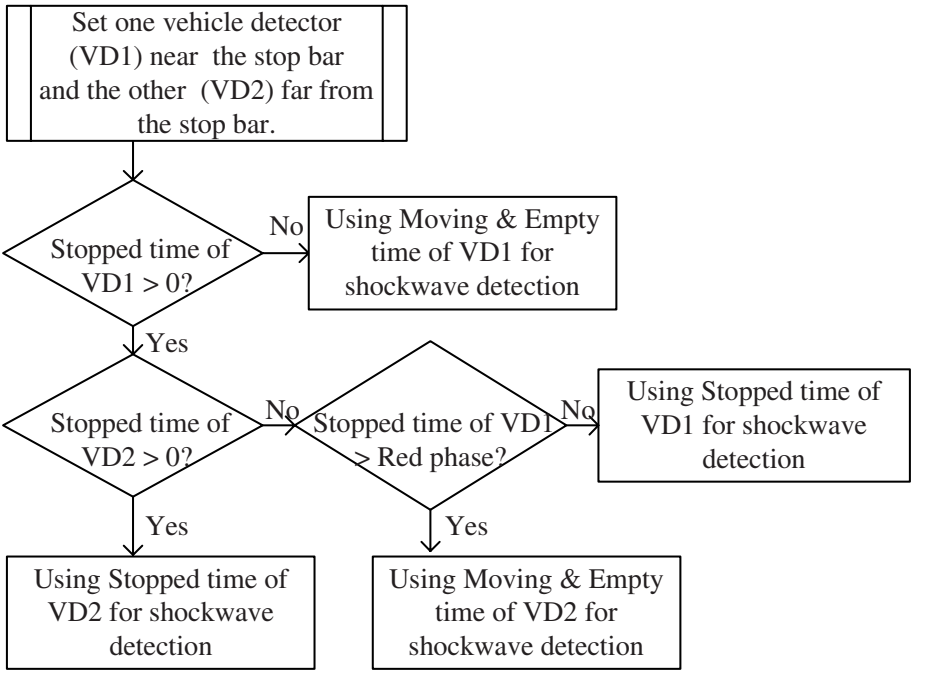


Fig. 4. Shockwave detection algorithm

Meanwhile, Equation (9) demonstrates how to use Stopped time to detect arrival shockwave. In Fig.4, the algorithm illustrated how to combine these two equations for arrival shockwave detection.

3 Simulation Results

Two CORSIM frameworks have been created for assessing the three new traffic parameters and the proposed shockwave detection algorithm. One framework is designed to identify the relations between environment and the three new traffic parameters. The other is for testing the shockwave detection algorithm.

3.1 Three Traffic Parameters and Environment

The test example shown in Table 1 involves two adjacent intersections with a 700 feet link at a fixed-timed traffic signal operating with a 60-s cycle length and a 23-s effective green interval. To determine the features of the three parameters and the effects of environment change, eight scenarios were prepared and listed in Table 1: traffic flow changing from 48 vph to 2880 vph, different detector zone size (20, 50 feet), different vehicle detector distances to the stop bar (0, 30, 60,90,120,240 feet), and next intersection spill back condition. Table 1 lists the test results. The arrows in Table 1 indicate the trend of traffic parameters.

Table 1. Relation between three traffic parameters and the environment

			Empty	Stopped	Moving
Flow	48 → 2880	20 feet	↘ 6	↗ 38	↗ 16
		50 feet	↘ 0	↗ 38	↗ 22
		VD to stop bar	far → near	near → far	near → far
		Spill over	no → yes	no → yes	no → yes
	2880 → 48	20 feet	6 ↗	38 ↘	16 ↘
		50 feet	0 ↗	38 ↘	22 ↘
		VD to stop bar	far → near	near → far	near → far
		Spill over	no → yes	no → yes	no → yes

3.2 Shockwave Detection

The other test example for arrival shockwave features two adjacent intersections with a 997 foot link at a fixed-timed traffic signal operating with a 60-s cycle length and a 30-s effective green interval. Two electronic vehicle detectors are located 300 and 570 feet from the stop bar. The change sequence for the traffic flow is 650, 550, 600, 500, 700, 550, 400, 600 and 1000vph for every 900 seconds. Since the capacity is 600 vph, the v/c values vibrate around 1 and some traffic queues are formed. Figure 5 shows three traffic parameters in two electronic vehicle detectors. Notably, the Stopped time of VD2 is always zero, indicating that no traffic queue exists over VD2. Furthermore, the Stopped time of VD1 is occasionally near the effective red time (30). Restated, some traffic queues are shown on VD1. Figure 6 summarizes the results from two different shockwave equations proposed in subsections 2.2 and 2.3. According to this figure, each equation can only detect arrival shockwaves accurately during a specific time interval. Figure 7 illustrates that the algorithm can obtain good results. The detected arrival shockwave is almost identical to Corsim's shockwave.

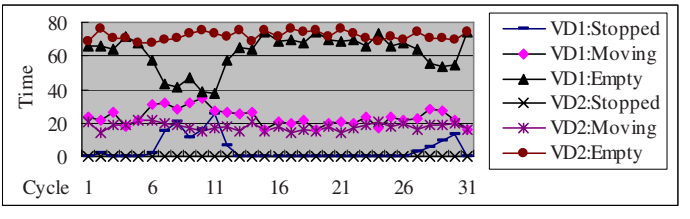


Fig. 5. Three traffic parameters in VD1 and VD2

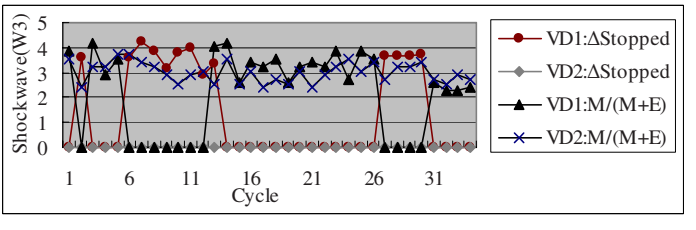


Fig. 6. Using Stopped, Moving and Empty of VD to detect arrival shockwave

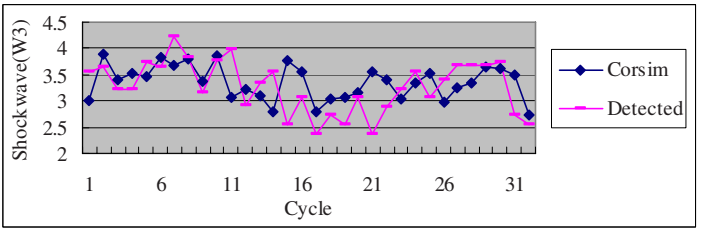


Fig. 7. The final result of the shockwave detection algorithm

4 Conclusions

This work proposes, for the first time, three new traffic parameters and a new approach to shockwave detection for electronic vehicle detectors. The proposed shockwave detection approach comprises three major components: a shockwave equation derived from the new Stopped parameter, the other shockwave equation derived from the new Moving and Empty parameters, and an algorithm showing how to use the new parameters and equations to more accurately identify shockwaves. Different scenarios involving simulated environments were used to identify the behaviors of the three traffic parameters. The proposed shockwave detection methodology was tested for a case involving two adjacent intersections in a simulated environment. Simulated data from two electronic vehicle detectors were used. Each electronic vehicle detector obtained two Shockwave equations. The proposed algorithm selected the best of two equations as the final shockwave. This study has demonstrated the feasibility of using the two shockwave equations and algorithm for shockwave detection in urban intersections. The proposed algorithm is easily applied and cost-effective since low cost electronic detectors can be used to generate these three new parameters and detect shockwaves. Further research is required to extend the concept to freeways and highways.

Acknowledgements

The authors would like to thank the Ministry of Education of the Republic of China, Taiwan (MOE ATU Program) and the National Science Council of the Republic of China, Taiwan (Contract No. NSC-95-2221-E-009-346, NSC-95-2221-E-009-347 and NSC-95-2752-E-009-010-PAE) for financially supporting this research.

References

1. Gazis, D.C.:The Origins of Traffic Theory, Operation Research, Vol. 50, No. 1, 2002, pp. 69-77.
2. May, A. D.:Traffic Flow Fundamentals, Prentice Hall, Englewood Cliffs, New Jersey 07632, 1990.
3. Zhang, H.M.: A theory of nonequilibrium traffic flow, Transportation Research B 32 (7), 1998, pp. 485-498.
4. Cho, H.-J., Lo, S. C.: Modeling of Self-consistent Multi-class Dynamic Traffic Flow Model, Physica A., Vol. 312, 2002, pp. 342-362.
5. Cho, H.-J., Tseng, M.-T.: A Novel Computational Algorithm for Traffic Control SoC, WSEAS Transactions on Mathematics, Issue 1, Volume 5, 2006, pp. 123-128.
6. Dion, F., Rakha, H., Kang, Y.-S.: Comparison of delay estimates at under-saturated and over-saturated pre-timed signalized intersections, Transportation Research-B, Volume 38, 2004, pp. 99-122.
7. Abu-Lebdeh, G., Benekohal, R. F.: Development of Traffic Control and Queue Management Procedures for Oversaturated Arterials, Transportation Research Record, No. 1603, 1997, pp. 119-127.
8. Greenshields, B.D.: A study of traffic capacity. Proceedings of the Highway Research Board 14, 1934, pp. 448-477.

Contour Extraction Algorithm Using a Robust Neural Network

Zhou Zhiheng, Li Zhengfang, and Zeng Delu

College of Electronic & Information Engineering, South China University of Technology, 510641 Guangzhou, China
crenna@163.com

Abstract. For the contour extraction from the images, the traditional active contour models may trap in local minimum and strongly depend on initial contour. A contour extraction algorithm based on a robust neural network is proposed in this paper. A searching series of circles are used in obtained feature pixels with adaptive threshold for the final curve function approaching by neural network. Robust back propagation algorithm has been used to control the final curve shape. The simulations also show that the proposed algorithm has a great performance for different kinds of images.

Keywords: contour extraction, neural network.

1 Introduction

Object extraction has a wide variety of applications in computer vision and image processing. Snake [1,2] and geodesic active contours models [3] have been developing rapidly during past years. These models usually take an initial contour defined for an object contour and make it evolve until the contour satisfactorily approximates the actual object contour. The initial contour is a rough approximation and the final contour is an accurate representation of the object boundary. The models can be achieved by introducing an energy function and then minimizing it. In order to deal with topological changing of the evolving contour, Chop[4], Caselles et al.[5] and Malladi et al.[6] introduced level set theory.

But the minimization of all these methods may trap in local minimum, especially in the case of the contour with long concavities. On the other hand, the initial contour must fully be inside or outside the object contour, and the direction of contour evolution must be given to the algorithm in advance.

Addressing all mentioned problems, a contour extraction algorithm based on neural network is proposed in this paper. Assume the object contour can be viewed as a close curve in the two-dimensional surface. We intend to approach the curve function from some training feature pixels obtained by a series of searching circles with adaptive threshold.

2 Feature Pixels Searching

The algorithm chooses a circle in the image, and then searches the feature pixels on the circle according to some rules. After creating a new circle through changing of

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 283–290, 2007.
© Springer-Verlag Berlin Heidelberg 2007

radiuses, it searches the feature pixels on the new circle as before. The above procedure repeats till the radiuses tend to zero. These series of circles are called as "Searching Circles".

Firstly, we choose a pixel in the middle (as possible) of the image as the initial center $O_0 = (x_0, y_0)$, and initial radiuses r_0 large enough to cause the initial circle C_0 to encircle the objects in the image.

2.1 Radius Changing

Generally, in order to stop the searching process the radiuses of the circles are supposed to become less and less. So, the key problem is how to reduce the radius. If the reduction of radiuses is too much, some parts of the contour would possibly be missed. If it is too less, the computational costs will be enhanced. Evidently, we can choose a big reduction when the curve is far away from the real contour, and vice-versa. If we denote $I(x, y)$ as the value at pixel (x, y), r_n as the radius of the current searching circle C_n, then reduction of the radius

$$\Delta r_n = r_n - r_{n+1} \triangleq f_{\Delta r_n}\left(\max\{|\nabla I(x, y)| \,|\, (x, y) \in C_n\}\right) \tag{1}$$

where $\nabla I(x, y)$ is the gradient of the image at (x, y) and

$$f_{\Delta r}(u) = \frac{\ln(255/2)}{\ln(2+u)} \ . \tag{2}$$

So, $f_{\Delta r}(u)$ is a decreasing function with respect to x, which means Δr becomes smaller as the searching circle is closed to the objects boundaries meanwhile the absolute gradients become bigger. In equation (2), avoiding the denominator being 0, we use "2+*u*". On the other hand, if the gradient is bigger than 255/2, there is a strong enough edge passing this pixel. So, *u* is comparable to 255/2.

2.2 Adaptive Threshold

The simplest method for judging contour feature pixels is to use gradient threshold, nevertheless a single threshold is not fit for the whole image. In order to save more computations, the threshold *T* is adaptively determined by the method, which is given as following:

(a) Select an initial estimate for *T*.

(b) Divide all pixels of the former *n* searching circles $C_0, \cdots, C_{n-1}$ into two groups, such as

$$G_1 = \left\{(x, y) \,|\, |\nabla I(x, y) > T|\right\}, G_2 = \left\{(x, y) \,|\, |\nabla I(x, y) \le T|\right\}. \tag{3}$$

(c) Compute the average gradient values μ_1 and μ_2 for the pixels in G_1 and G_2, respectively.

(d) For the *n*+1th searching circle C_n, use new threshold value

$$T = \frac{1}{2}(\mu_1 + \mu_2) \tag{4}$$

(e) Repeat steps (b) through (d) until searching stops or the difference in T in successive iteration is smaller enough.

It can be f ound that when the pixels on the contour and pixels in other smooth region occupy comparable numbers in the image, a good initial value for T is the average gradient value of the image. But, when the former is smaller than the later, the average gradient value is no longer a good initial choice and median value between the maximum and minimum gradient values will be better.

3 Contour Approaching

The object contour can be viewed as a close curve in the two dimensional surface. The feature pixels, obtained in the second section, can be viewed as some training samples of the curve. So, the contour can be obtained by approximated the curve function by these samples. Hornik [7] used the Stone-Weirstrass theorem to prove that a two-layer feedforward network can approximate a continuous real function on a compact subset of $\Re^n$ arbitrarily closely in terms of the L^∞ norm. We try to use neural network to approach the curve function, and then obtain the whole object contour.

3.1 Function Approximation by Feed-Forward Network

Denote the contour feature pixels set by $\{(x_p, y_p); p = 1, \cdots, n\}$, where $y_p = f(x_p)$. The task of the neural network is to learn an estimation $\hat{y} = \hat{f}(x)$ of function $f(x)$, so that the error will approach the minimum in a certain kind of measurement. Multi-layer networks are highly nonlinear models for function approximations.

In general, an L-layer feedforward network with fixed nonlinear activation functions can be parameterized by a set of L-1 weights w^l, $l = 0,1,\cdots,L-2$. The weight w^l relates the lth layer output x^l to the l+1th layer activation A^{l+1} by means of an affine transformation, i.e., $A^{l+1} = w^l x^l$, $0 \le l \le L-1$. Notice that the 0th layer output is the network input x. For an input x, the network estimated output $\hat{y}$ is determined by the following forward propagating recursive equations which characterize the network dynamics.

$$x^0 = x, \quad x^l = G^l(A^l), \tag{5}$$

where $A^l = w^{l-1}x^{l-1}$, $l = 1,\cdots,L-1$, $\hat{y} = \hat{f}(x;w) = x^{L-1}$.

Here x^0 and x^{L-1} are the input and network output, respectively, and G^l is a function, which applies a nonlinear activation function $g^l(\ )$ to each component of its argument.

In particular, the parametric model represented by a three-layer network with one input node, a single output node, and *N* nodes in the hidden layer has the form

$$\hat{f}(x;w)=\sum_{i=1}^{N}\beta_i g_i(\alpha_i x) \tag{6}$$

where β_i, $1\le i\le N$ are weights connecting *N* hidden nodes to the output node, α_i, $1\le i\le N$ are weights connecting input layer node to the *i*th hidden layer node and the $g_i(\ )$s are the hidden layer activation functions.

3.2 Robust Back Propagation Algorithm

Back propagation (BP) algorithm allows multi-layer feed-forward neural networks to learn input-output mappings from training samples. The BP algorithm iteratively adjusts the network weights to minimize the least squares energy function

$$E(w)=\sum_{p=1}^{n}\left(y_p-\hat{y}_p\right)^2 \tag{7}$$

where y_p is the desired output for the *p*th training sample, and $\hat{y}_p$ is the corresponding estimated output. It can be found that $E(w)$ reaches zero when the estimator $\hat{f}$ given by the network interpolates all training pixels, i.e., $y_p=\hat{y}_p$.

The final contour is determined by feature pixels obtained in the second section. But directly approximation from feature pixels will make the object's shape complicated, in another words, the function of the curve be undesired fluctuating, because of noise pixels or misjudged pixels in the second section. All these feature pixels can be viewed as outliers. We intend to make the final curve detours these outliers, so we must reduce the big error deviation causing by them. Robust Statistics is an important research field for solving this kind of problem.

Based on the theory of robust statistics, the energy function can be rewritten as

$$E(w,a)=\sum_{p=1}^{n}\rho_p\left(y_p-\hat{y}_p,a_p\right) \tag{8}$$

where $\rho_p(u)$ is a Huber function [8] as shown in Fig.1,

$$\rho_p(u,a_p)=\begin{cases}u^2 & |u|\le a_p\\ a_p^2+2a_p(|u|-a_p) & |u|>a_p\end{cases},\ a>0. \tag{9}$$

a_p is a parameter for the outliers adjusting influence on the energy function. It can be found that if a_p goes bigger, the less influence of outliers will be added on the energy function, and vice versa. On the other hand, if a feature pixel with smaller absolute gradient value, it tends to be outliers and the final curve will detours it. And

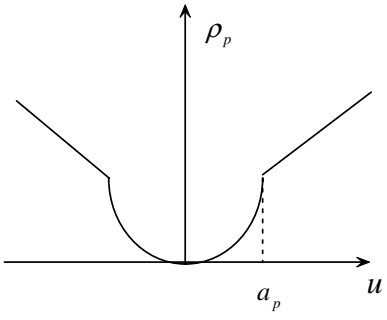


Fig. 1. Huber function

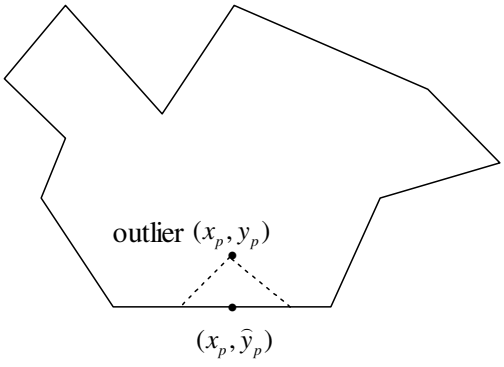


Fig. 2. The curve detours the outlier

on the contrary, the feature pixel with bigger absolute gradient value is on the final curve. So, we construct the relative expression between a_p and absolute gradient value

$$a_p = \lg\left(\left|\nabla I\left(x_p, y_p\right)\right| + 1\right). \tag{10}$$

Fig.2 shows how the curve detours the outlier. In this case, $\left|\nabla I\left(x_p, y_p\right)\right|$ is relatively small. The dashed lines are the curve whose function is directly approximated from training data. And the real lines are the curve whose function is approximated by robust energy function. The robust processing will make the final curve closer to all feature pixels with bigger absolute gradient values.

4 Simulation Results

This section we are going to present the performance of our robust approaching algorithm after simulation. The design of initial searching circle C_0: Assume the resolution of the observed image is $M \times N$, we choose initial center $O_0 = (\lfloor M/2 \rfloor, \lfloor N/2 \rfloor)$ and initial radius $r_0 = \lfloor \min(M, N)/2 \rfloor$ to cover the whole part of the image.

In order to verify the performance of the proposed algorithm, simulations will be implemented in different kinds of images, such as complicated background, and noisy boundary.

4.1 Complicated Background

If the image contains complicated background, the contour of an object may be composed of different parts of boundaries, or objects' contours are very different from each other. Traditional contour extraction algorithm mainly depends on the variation of gradients to do the judgments. Simple judging mode cannot discriminate different parts of contour.



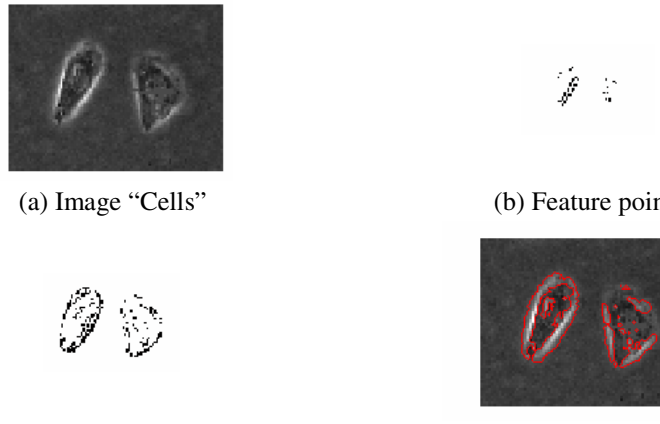
(a) Image “Hawk” (b) Feature points (c) Extracted contour

Fig. 3. Robust NN for image “Hawk”

The proposed contour extraction algorithm is based on adaptive threshold. We use image “Hawk” with complicated background and 128 128 in size to do the simulation. In Fig.3, the final contour can approximate the actual object contour of image “Hawk”.

4.2 Noisy Boundary

In traditional active contour model, if the evolving curve reaches the object contour, the gradient-directed function will tend to zero. In fact, this function has some limitations. On one hand, as pointed in the famous C-V model [9], image gradients are bounded. On the other hand, if the object boundary is blurring, the gradients does not change obviously. In these cases, the function does not tend to zero even if the evolving curve reaches the object contour. It will cause the evolving curve pass through the actual contour. For image noise, active contour model cannot give solutions. But the proposed robust method can solve the problem efficiently.



(a) Image “Cells” (b) Feature points

(c) Robust NN extracted contour (d) C-V model extracted contour

Fig. 4. Robust NN and C-V model for image “Cells” with time cost 3.1s and 42.8s, respectively

We use image "Cells" with size 83 65 to do the simulation. In Fig.4, it can be found that the final contour obtained by the proposed algorithm is quite accurate. It can be found that C-V model does not superior to robust NN, but it costs much more time than robust NN. We also do other comparisons for Gaussian white noise contaminated image between robust NN and C-V model as shown in Table 1.

Table 1. Computational cost time (s) comparisons with Gaussian white noise contaminated image with size 256×256

Variance	C-V model	Robust NN
0	100.8	29.9
0.01	140.6	41.7
0.02	187.6	43.8
0.03	221.3	59.8

5 Conclusions

Object detection in the image has a wide variety of applications in computer vision. An approaching algorithm for object contours based on a robust neural network is proposed in this paper. A series of searching circles are used in obtained feature pixels with adaptive threshold for the final curve function approaching by neural network. Robust back propagation algorithm has been used to control the final curve shape. The simulations also show that the proposed algorithm has a great performance in contour extraction. Whatsoever, the contributions in this paper only serve as a prelude to our future work, in hopes of better results, to integrate this method with snakes, geodesic active contour, etc.

Acknowledgments. Supported by National Natural Science Foundation of China (Grant No. 60325310, No. U0635001), the Science Foundation of Guangdong Province for Program Research Team (Grant No. 04205783), the Specialized Prophasic Basic Research Projects of Ministry of Science and Technology, China (Grant No. 2005CCA04100), China Postdoctoral Science Foundation (Grant No.20060390728).

References

1. Kass, M., Witkin, A., Terzopoulos, D.: Snakes: Active Contour Models. Int. J. Computer Vision, Vol. 1, (1988) 321-332
2. Jacob, M., Blu, T., Unser, M.: Efficient Energies and Algorithms for Parametric Snakes. IEEE Trans. on Image Processing, Vol.13, No.9, (2004) 1231-1243
3. Paragios, N., Deriche, R.: Geodesic Active Contours and Level Sets for the Detection and Tracking of Moving Objects. IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol.22, No.3, (2000) 266-280
4. Chop, D.: Computing Minimal Surfaces via Level Set Curvature-Flow. J. Computational Physics, Vol. 106, (1993) 77-91
5. Caselles, V., Kimmel, R., Sapiro, G.: Geodesic Active Contours. Proc. IEEE Int. Conf. Computer Vision, (1995)

6. Malladi R, Sethian JA, Vemuri BC: Shape Modeling with Front Propagation: A Level Set Approach. IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol.17, No.2, (1995) 158-175
7. Hornik, K., Stinchcombe, M., White, H.: Multilayer Feed-forward Networks Are Universal Approximators. Neural Network, Vol. 2, (1989) 359-366
8. Huber, P. J.: Robust Statistics. New York: Wiley, (1981) 73-106
9. Chan, T.F., Vese L.A.: Active Contours without Edges. IEEE Trans. on Image Processing, Vol.10, No.2, (2001) 266-277

A Discrete Parameter-Driven Time Series Model for Traffic Flow in ITS

Yow-Jen Jou[1] and Yan-Chu Huang[2]

[1] Department of Finance and Information Management,
National Chiao Tung University, Taiwan
[2] Institute of Statistics, National Chiao Tung University, Taiwan

Abstract. The traffic condition information is critical to signal control and traffic queue management for intelligent transportation systems (ITSs). This work formulates the dynamic traffic flows by a parameter-driven discrete time series model. A modified Expectation-Maximization (EM) algorithm is proposed to estimate the parameters of the model. The conditional expectation given y in the E step of the EM algorithm is replaced by the marginal expectation. Applications of the model and the algorithm are illustrated by analyzing the traffic flow data. Through the data collected by the detectors, the traffic condition about the network can be estimated, smoothed and predicted. This model can be embedded in a chip for signal control and queue management procedures of ITSs.

1 Introduction

Intelligent transportation systems deal with subjects involving the application of advanced technologies such as information and communication technologies to improve the traffic management. Traffic congestion occurs daily in most urban areas, posing a serious threat to the environment. Appropriate traffic management procedures are increasingly required with the deployment of more efficient traffic control systems within the ITSs complex that handles information, including traffic flow prediction, on a real-time basis. Although signalized intersections have been investigated in recent decades using various methods, the merits of signal control have not been fully realized [2]. Managing traffic more efficiently requires sustaining the road system in a free-flow condition as long as possible. Therefore, this work presents, for the first time, a parameter-driven model for a time series of counts to describe the evolution of the traffic flow. The parameters are then estimated, followed by forecasting of the future traffic flow.

Discrete time series model has attracted considerable attention recently [1], [3], [4]. Cox [1] characterized time series models into two categories: parameter-driven and observation-driven models. These models result in very complex likelihoods which require computation intensive solution like expectation-maximization (EM) algorithm. In this paper, the conditional expectation given y in the E step of the EM algorithm is replaced by the marginal expectation. This is a more efficient algorithm for the estimation of the parameter-driven model. Therefore, the proposed model can be embedded in the signal control chip as a system on chip (SoC) [2]. The chip can be used for advanced traffic management systems, advanced traveler information systems and advanced public transportation systems.

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 291–294, 2007.
© Springer-Verlag Berlin Heidelberg 2007

2 Model Specifications and Methodology

This section describes the parameter-driven model and discusses the modified EM algorithm. Two methods are proposed to smoothen and estimate the response variable.

2.1 The Model

Let $\{Y_t\}_{t=1}^N$ denote a time series of counts with nonnegative integer values which comprise the response process. This study use a parameter-driven model to account for the $\{W_t\}$, which is assumed to be a stationary Gaussian AR(1) latent process. Conditional on $\{W_t\}$ the observations are independent Poisson random variables with means λ_t, written as

$$Y_t|\lambda_t \sim Po(\lambda_t), \tag{1}$$

where $\{\lambda_t\}$ is defined as $\lambda_t = \exp(W_t + \alpha' U_t)$, and

$$W_t = \rho W_{t-1} + \varepsilon_t, \tag{2}$$

where $\{\varepsilon_t\}$ is iid $N(0, \sigma_\varepsilon^2)$. The parameter vector $\theta = (\rho, \sigma_\varepsilon^2, \alpha')$.

Although an elegant and straightforward means of describing the relationship between the response variable and the covariates, the parameter-driven model makes it intractable to analyze the likelihood function analytically. This problem can be solved using currently available computational power.

2.2 The Modified EM Algorithm

Let $X_t = (y_t, W_t)$ denote the complete-data random vector of interest at time t, let $X = (X_1, \ldots, X_N)$ represent the realization of the complete-data, let $Y = (y_1, \ldots, y_N)$ be the observed data and let $W = (W_1, \ldots, W_N)$ be the latent process which will be treated as missing values.

In the presence of latent process, which is treated as missing data, only the log likelihood function $l_X(\theta|y) = \log \int f_X(y, W|\theta) dW$ is observed. Each iteration of the EM algorithm involves two steps:

- E step: From $Q(\theta'|\theta) = E_\theta(l_X(\theta')|y)$;
- M step: Maximize $Q(\bullet|\theta)$.

$E_\theta(\bullet|y)$ denotes the conditional expectation given $Y = y$, where θ represents the true parameter and $l_X(\theta')$ is the log-likelihood of X. Because the conditional expectation given y of $l_X(\theta|y)$ is intractable, the EM algorithm is adjusted by replacing the marginal expectation with the conditional expectation given y. The procedure of the modified EM algorithm is completed when the changes are within the tolerance. To determine the precision of the parameter estimation, the estimated information matrix is calculated by $I(\hat{\theta}) = E(-\partial^2 l_X(\theta)/\partial\theta\partial\theta'|_{\theta=\hat{\theta}})$, and the estimated variance-covariance matrix is $\hat{V}ar(\hat{\theta}) = I^{-1}(\hat{\theta})$.

2.3 Fitting and Prediction

The models in Eqns. (1) and (2) are considered, and the EM algorithm is found to be readily applicable. Furthermore, since

$$E_\theta(\lambda_t|y) = \exp(\alpha' U_t) E_\theta(\exp(W_t)|y) \tag{3}$$

for all t, fitting and prediction are easily conducted.

The l-step prediction can be obtained from

$$E_\theta(\lambda_{N+l}|y) = \exp(\alpha' U_{N+l} + (1-\rho^{2l})\sigma_\varepsilon^2/2(1-\rho^2)) E_\theta(\exp(\rho^l W_N)|y) \tag{4}$$

Two methods are proposed for smoothing and predicting the response variable. In method 1, unconditional expectations are used for all the terms involving the latent process. Meanwhile, in method 2, the estimated conditional expectation of the latent process is substituted by the simulated results.

3 Numerical Example: ITS Traffic Flow

To management traffic more efficiently, maximizing the periods for which road systems are in a free-flow condition is preferable. The data analyzed in this study were obtained by RTMS (Remote Traffic Microwave Sensor) from the intersections before the rush hour. The data are taken at ten-second intervals. The notations used in the model are listed as follows:

Y : Traffic flow (volume); this value indicates the detected flow,
u : Occupancy; the percentage of time during which vehicles traverse a given point, and
N : The number of data.

This study applied the parameter-driven model to analyze traffic flow as an illustration of the application of the proposed approach. The covariate vector in the model is given by $U_t = (1, u_t)'$. The previous section adopted the latent AR(1) model, and the modified EM algorithm was used for the estimation. The starting values for ρ and σ_ϵ^2 are 0.5 and 0.5. Furthermore, the starting values of α are obtained by fitting a log-linear model to the data, assuming no temporal dependence, namely $(\alpha_0, \alpha_1)'$ are $(0.682636, 2.550254)'$. The value of $Q(\bullet|\theta)$ is maximized and the approximate ML estimates are obtained by $\hat{\rho} = 0.707559(0.022361), \hat{\sigma}_\epsilon^2 = 0.332942(0.014899)$, $\hat{\alpha}' = (0.350308(0.025593), 2.580916(0.129314))$. The figures in parentheses are the estimated standard errors, obtained from the inverse information matrix. Fitted results are shown in Fig. 1 in which dots represent the observations. The thinner solid line connects the fitted data simulated using method 1, and the average error between the observations and the fitted data is $\sum_{t=1}^{N} |y_t - \hat{y}_t|/n = 1.181$. The thicker solid line connects the fitted data simulated using method 2, and the average error between the observations and the fitted data is $\sum_{t=1}^{N} |y_t - \hat{y}_t|/n = 2.277$. Clearly the average error between observations and the fitted data from method 1 is smaller than that obtained by method 2, an instability that may partly result from the variability associated with the latent process W_t.

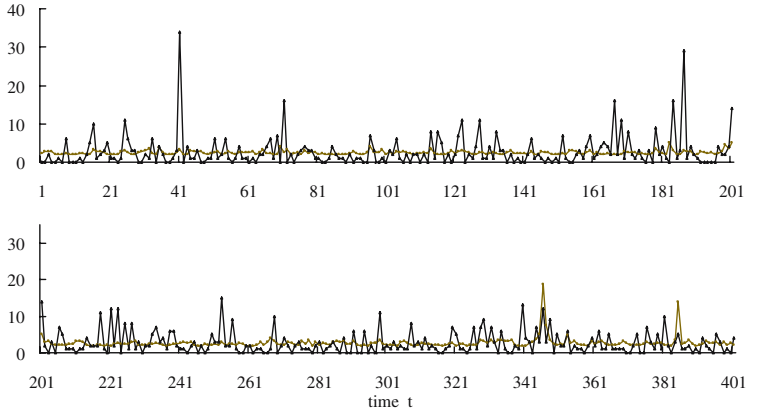


Fig. 1. Time sequence of the traffic flow (the first 400 data of complete data set). The dots represent observations. The thinner solid line connects the fitted data simulated using method 1. Meanwhile, the thicker solid line connects the fitted data simulated using method 2.

4 Conclusions

In this work, a modified MCEM algorithm is proposed to the analysis of a parameter-driven model for a time series of count data. The applicability of the proposed algorithm is demonstrated by analyzing the relationship between traffic flow and occupancy, which is modelled by assuming the existence of an AR(1) Gaussian latent process. The proposed algorithm also possesses the advantages of being economical in terms of both computational time and memory space. This model can be used for signal control and queue management procedures of ITSs.

Acknowledgement

This research was partially supported by the Ministry of Education of the Republic of China, Taiwan, MOE ATU Program and partially support by National Science Council, Taiwan, under grant number NSC-95-2221-E-009-347 and NSC-95-2622 -E-009-007-CC3.

References

1. Cox, D.R.: Statistical analysis of time series: some recent developments (with discussion), Scandinavian Journal of Statistics, Vol. 8, 1981, pp. 93-115.
2. Cho, H.J., Tseng, M.T.: A novel computational algorithm for traffic signal control SoC, WSEAS Transactions on Mathematics, Issue 1, Vol. 5, 2006. pp. 123-128.
3. Chan, K.S., Ledolter, J.: Monte Carlo EM estimation for time Series models involving counts, Journal of the American Statistical Association, Vol. 90, No. 429, 1995, pp. 242-252.
4. Freeland, R.K., McCabe, B.P.M.: Analysis of low count time series data by Poisson autoregression, Journal of Time Series Analysis, Vol. 25, No.5, 2004, pp. 701-722.

Peer-Based Efficient Content Distribution in Ad Hoc Networks*

Seung-Seok Kang

Department of Computer Science
Seoul Women's University
Seoul 139-774, Korea (ROK)
msukang@swu.ac.kr

Abstract. Mobile devices pay the telecommunication cost of downloading Internet data proportional to the amount of data transferred. This paper introduces a special *ad hoc* network in which several mobile devices, called peers, to cooperate each other to reduce the overall cost to download Internet content. Each peer downloads a specific portion of the content over 3G connection and exchanges the portion with other peers over the *ad hoc* connection in order that all participating peers are able to reconstruct the whole content. This paper proposes a peer-based content distribution method, and compares its performance with a similar one named *per-packet* based distribution method. The simulation results indicate that *per-peer* based method outperforms the *per-packet* based method. In addition, approximately 90% of the telecommunication cost is saved with as few as 10 peers.

Keywords: ad hoc network, 3G, peers, content distribution.

1 Introduction

Current wireless telecommunication services provide high speed Internet access as well as voice service. Fig. 1 displays a situation in which many nearby mobile devices connect to the Internet via their wireless telecommunication links to their ISPs and then access their favorite content. Suppose the devices in the figure try to download the same file stored at a content provider (CP), such as mobile game programs, MP3 files, or movie clips. This situation may happen in many places. For example, a teacher may want to share some educational content with his/her students indoor and outdoor. Some friends may download an interesting mobile game program that they want to store on their mobile devices and play together interactively. In a sports stadium, spectators may want to retrieve the records of their home team and favorite players in a game. Since it is expected that the 3G connection cost to download data from the Internet is a function of the amount of data downloaded, the cost of the telecommunication connections to access the Internet may be reduced when each mobile device is assigned to download a given portion of the target file and shares its portion with other devices. This

* This work was supported by a research grant from Seoul Women's University (2006).

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 295–302, 2007.
© Springer-Verlag Berlin Heidelberg 2007

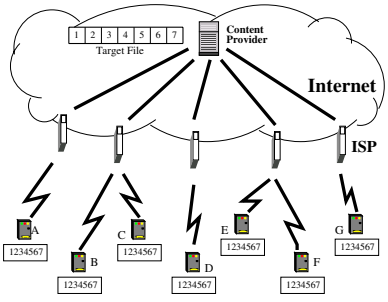


Fig. 1. Mobile devices download same target file in the Internet

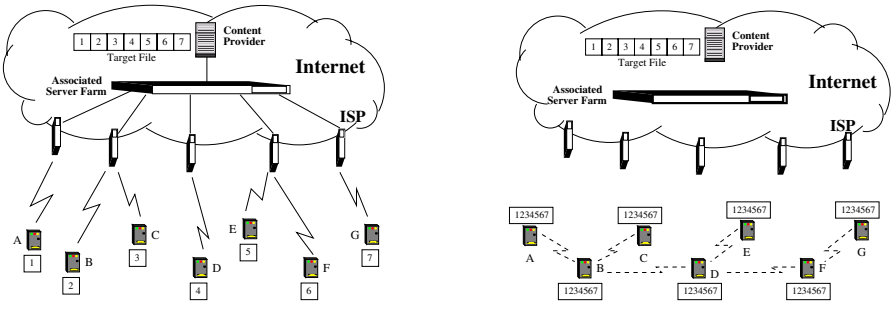


Fig. 2. Downloaded portions are exchanged over the *ad hoc* connection

paper proposes a low-cost sharing schemes in which mobile devices download in parallel their assigned portions of the target file via their 3G connections, and then build an *ad hoc* network (that has no fee for data transferred) to exchange the remaining portions of the file. This may be possible if each mobile device has both a 3G interface for wireless wide area network (WWAN) access such as UMTS or CDMA2000 and a wireless local area network (WLAN) interface, such as 802.11 or Bluetooth, to form an *ad hoc* network [1].

The left side of Fig. 2 illustrates the mechanism for partitioning the target file and downloading the assigned portions by the mobile devices. Each mobile device is assigned a portion of the target file to download. Each device connects to its favorite ISP with its 3G link and contacts the CP with the aid of its associated server in order to request its assigned portion of the file. The main difference between Fig. 1 and the left side of Fig. 2 is the thickness of the 3G connection, which represents the amount of data downloaded for each mobile device. In Fig. 1, each mobile device downloads the *entire* content from the CP. However, the thin line in the left side of Fig. 2 indicates that each device downloads only a *portion* of the target file. Upon completion of downloading the specified portion of the file, the mobile devices use their *ad hoc* connections to exchange

their content with other member devices. The right side of Fig. 2 shows that all mobile devices participate by exchanging their partial content with others in order to reconstruct the target file.

The idea of the cooperating *ad hoc* network concept is often described as *peer-to-peer* computing [2]. In [3], the peers download an assigned portion of a target file and distribute their portions to all other peers based on the *per-packet* based method. Fairness issues of the multiple sources to single receiver (multipoint-to-point) is studied in [4]. SplitStream [5] allows a source peer to split a file into k stripes and to multicast each stripe using k multicast trees. In BitTorrent [6], when many downloaders try to download a file(s) from a URL-specified location, the downloaders upload to each other concurrently to help reducing the load of the source. The authors in [7] proposed an approach for computing a schedule for coordinated data collection with avoiding congested links and maximizing the network resource utilization.

This paper focuses mainly on the data distribution method among participating peers. This method is based on a *per-peer* policy in which all peers are ordered to transmit their packets one peer at a time. They all know which peer is its predecessor. When a peer transmits some number of data packets, the content is delivered only to the sender's one-hop neighbors. At the end of the current peer's transmission, the peer sends a special packet that triggers the transmission of the next peer. The transmission process finishes when all peers receive the complete target file.

The rest of the paper is organized as follows. Section 2 deals with two distribution methods. Simulation results are explained in section 3. Section 4 draws conclusions.

2 Data Distribution Methods

2.1 Network Formation

The work in [3] describes the special *ad hoc* network formation. One of the device users, called a peer, initiates a connection to its ISP and contacts a special server on the Internet. The server becomes the associated server of the peer. Each peer needs to have its associated server. The associated server may be located at the ISPs, the CPs, or some other place within the Internet. A server may be associated with several peers. For some cases, all peers may connect to a single server that controls the operation of the *ad hoc* network and deals with any license issues and fees with CPs at a bulk rate on behalf of the peers. The associated server of the initiating peer plays an additional role of the master server. The master server deals with several management tasks for the special *ad hoc* network including the computation of the download scheduling that decides which peers to download which portions of the target file.

Each peer downloads only a portion of the target file, but needs to acquire complete content. In order to reconstruct the file, every peer becomes a sender as well as a receiver. However, if all peers, as senders, transmit their content in an uncontrolled manner, the broadcast storm problem [8] may arise. Each

peer needs a controlled way of broadcasting its received data to other member peers. A *per-packet* based distribution method [3] selects the rebroadcasting set of peers in order to propagate content to other peers because some peers are out of transmission range of others. When a peer transmits one unit-sized packet of downloaded data that is new to other peers, the data packet travels over the *ad hoc* network by rebroadcasting from some selected rebroadcasting peers, if needed. If all peers receive the packet, the scheduled next peer transmits another unit-sized packet of downloaded data. *Per-packet* transmission repeats until all peers receive the complete content.

2.2 Per-Peer Based Distribution

In the *per-packet* distribution method, one packet from one peer is forwarded to all other peers at a time. Then, the next scheduled peer takes a turn to transmit its next unit-sized packet. This delivery method, however, shows some inefficiency while transmitting packets. Peers located at the central area may experience heavy packet collisions. In addition, for some moment, non-data packets such as DONE packets may dominate the *ad hoc* network, which increases the completion time and degrades the transmission performance.

A *per-peer* based distribution method decreases the chance of packet collision and increases the time for more data packets transmitted, which results in shorter completion time, less 3G communication costs, and less power consumption than that of the *per-packet* based method. In the *per-peer* based distribution, only one peer has a chance to transmit several unit-sized data packets to its neighbors at a time. In addition, the data packets are not immediately forwarded, which reduces the possibility of packet collision. Each peer independently decides which unit-sized data to broadcast depending on the reception status of its neighbors. The neighbors do not rebroadcast any packet immediately, but wait for their transmission turn. When the current transmitting peer has finished broadcasting a given number of packets, it broadcasts a DONE packet. If a scheduled next peer is directly connected with the current peer, the next peer resumes its transmission after selecting which data to send. If the next peer is out of range from the current peer, the DONE packet contains a list of rebroadcasting peers that forward the DONE packet toward the next-scheduled transmitting peer.

The master server creates and maintains a transmission sequence as a circular list of peers using global topology information. Fig. 3 illustrates one example of a 3-hop *ad hoc* network. The master server constructs the list with the following rules. The server prefers the peer that has the largest number of neighbors. Then it selects the peer that is directly connected with the previously selected peer. If there is no unselected peers that have direct connection with the recently selected peer, the server chooses the peer with the minimum number of hops from the previously selected peer. Fig. 3 (b) is one possible example of the transmission sequence as a circular list. Peer B is selected first and subsequently peer E, A, and D are choosen. Because peer D and F have no direct connection, the link

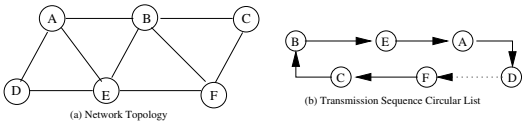


Fig. 3. A 3-hop network topology and its transmission sequence circular list

is shown as a dotted line. Peer D keeps a rebroadcasting set that includes peer E for rebroadcasting a DONE packet. In addition, each peer maintains two-hop neighborhood information. Much research have proposed distributed algorithms that utilize two-hop neighborhood information [9,10,11,12]. When a peer receives a packet, it immediately knows which peer also receives the same packet. This knowledge may obviate some unnecessary transmissions of data packets.

2.3 Benefit Value

One important value that each peer maintains for each unit-sized data segment is the number of neighbors that do not have the segment. Because the target file size is known and the data packet size can be defined, e.g., 500 bytes, each peer is able to compute how many unit-sized data are needed to construct the target file. Suppose the target file size is 1 Mbyte and the unit packet size is 500 bytes, 2048 data packets are necessary. Each peer maintains both a bitmap size of 2048 bits and the same-sized integer array called the *benefit* value. The bit in the bitmap indicates the existance of the corresponding data segment stored in the peer. The integer benefit value implies the number of neighbors that do not have the corresponding data segment. In case that a peer downloads some portion of the target file from the Internet, the benefit value(s) of the data segment(s) is(are) set to the number of its neighbors, because no neighbor has the data segment. The corresponding bit is also set, because the segment is in the peer. When each peer resumes transmission, it selects the unit-sized data that has the largest benefit value, because the corresponding number of neighbor peers do not have the data and they will get a benefit to fill the gap in the target file. This also leads to increases a chance to discourage unnecessary data transmission by other peers with the aid of the two-hop neighborhood information. In Fig. 3, assume peer B sends a data packet whose benefit value is four, and peer A, C, E, and F receive the packet. Peer B's benefit value for the data becomes zero after the transmission, because all its neighbor peers hold the data this time. Both peer A and E computes a value of 1 for the data's benefit value, because only peer D does not have the data. Further assume that peer E transmits the data in its turn. Now peer A's benefit value of the data becomes zero and peer A is unnecessary to transmit the data in case of its turn to transmit. If a peer has all zero benefit values, it immediately sends DONE packet when it is the peer's turn. This allows other peers with positive benefit values to transmit data.

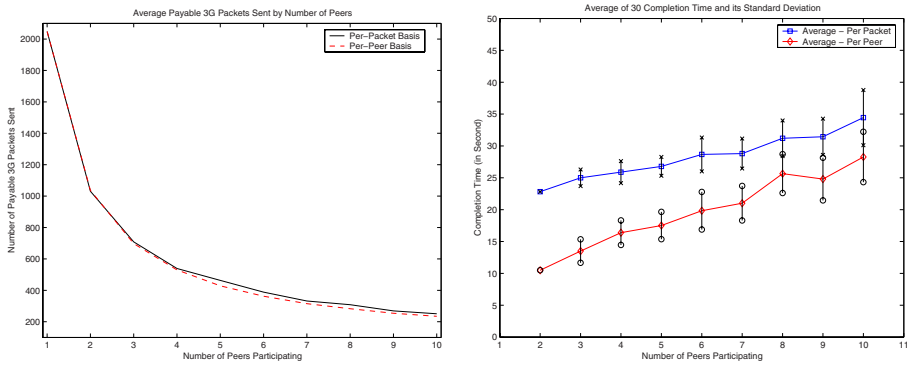


Fig. 4. Number of packets on the 3G link and completion time by varying number of peers

3 Simulation Results

The *ns2* network simulator [13] is used in this simulation. The simulation model assumes each peer downloads a portion of the target file using its 3G connection that the master server schedules to download, while it exchanges the portions with other peers using its *ad hoc* network. The peers do not experience buffer underrun when they transmit their portions to other peers. The unit packet size of the 3G connection is set to 500 bytes. The *ad hoc* network uses the same size of data packets. The peers download the target file size of 1 Mbytes that consists of 2048 unit packets. Each peer has the 802.11 MAC with the transmission range of 250 meters. The number of peers varies from 2 to 10. Because of the short amount of completion time described in [3], the mobility of peers is not considered. All peers are located in a 400 meter by 400 meter grid unless specified otherwise. In case of *per-peer* based distribution, each peer transmits the maximum of 90 packets at a time because the packet buffer size is set to 100 in the simulation. Each peer uploads its new neighborhood information in a second. In all cases, an *ad hoc* network is connected. That is, each peer is connected to at least one other peer.

The left side of Fig. 4 displays the average number of 3G packets used by the peers that may have a fee charged by a telecommunication provider. Each value in the figure is the average of 30 runs. As the number of peers increases, the number of fee-based packets decreases substantially. However, an additional peer only reduces a marginal cost when there are already enough peers, whereas, the new peer receives the same cost reduction. Approximately 90% of the telecommunication cost is saved with as few as 10 peers. The *per-packet* based method uses slightly more packets on the 3G telecommunications link than that of *per-peer* based method. This is due mainly to the recovery process in which each peer uploads its bitmap reporting any data gaps to its associated server. *Per-peer* based distribution resolves the gap by exchanging bitmaps with its neighbors over the cost-free *ad hoc* connection.

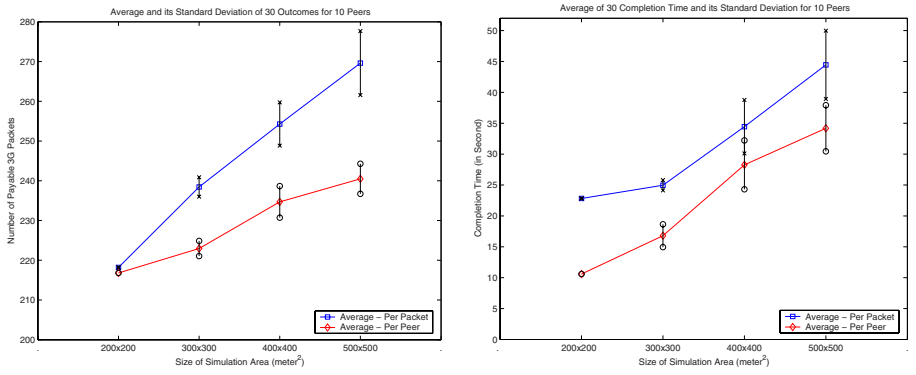


Fig. 5. Number of packets on the 3G link and completion time by varying size of simulation area

The completion time of each simulation is shown in the right side of Fig. 4. Both horizontal lines are the average of 30 runs and the short vertical lines represent the standard deviation of the 30 results. The average completion time increases slowly as more peers participate. This is due to the increasing number of hops in the network. As the number of peers increases, they may disperse over the grid, which may increase the number of hops. More hops cause more time to complete the distribution. The *per-peer* based distribution completes the content distribution faster than that of the *per-packet* based one. One important point is that as the number of peers increases, the difference of the two completion times decreases. This result indicates that the number of hops in a network is a more dominant factor than the density because *per-peer* based distribution hardly experiences any collision.

When all peers are located within the transmission range of each other, no collisions are expected because only one peer transmits at a time. Fig. 5 illustrates how the performance changes due to the change of the grid size. The simulation runs with 10 peers and repeats 30 times with 30 different network topologies. The left side of Fig. 5 shows the number of packets on the 3G telecommunications link as the size of simulation area increases. A larger area may result in a larger number of hops in a network, and more time to take to complete the content distribution. Because peers upload their neighborhood information periodically, longer completion time leads to increase the number of packets on the 3G link as well. In addition, peers upload their bitmaps and download the recovery instructions for *per-packet* based distribution, which consumes packets on the 3G link. Due to more collisions in a larger area, the *per-packet* based method uses more 3G packets than that of the *per-peer* based method. The right side of Fig. 5 shows the completion time by varying the simulation area size. The largest number of hops was 3 among 120 different topologies used in the figure. As the simulation size increases, so does the number of hops, which causes longer completion time. Overall, the *per-peer* based distribution method

outperforms the *per-packet* based method in terms of the completion time, and uses less packets on the 3G link.

4 Conclusion

Telecommunication cost may be one of the crucial factors for mobile device users to access Internet content. This paper describes a special *ad hoc* network in which mobile peers save telecommunication cost by sharing their partially downloaded data with other peers. Each peer agrees to download a specified portion of the target file located in the Internet using its fee-based WWAN connection. Each participating peer distributes its downloaded portion to all other member peers over the cost-free WLAN *ad hoc* connection so that all participating peers can generate the complete target file. The *per-peer* based distribution method utilizes the reception status of its neighbors using 2-hop neighbor information and the benefit value, and outperforms the *per-packet* based method. In addition, both distribution methods save approximately 90% of the telecommunication cost with as few as 10 peers.

References

1. Xiao, Y., Leung, K., Pan, Y., Du, X.: Architecture, mobility management, and quality of service for integrated 3G and WLAN networks. Wireless Communications and Mobile Compting **5** (2005) 805–823
2. Oram, A.: Peer-to-Peer : Harnessing the Power of Disruptive Technologies. 1st edition edn. O'Reilly & Associates (2001)
3. Kang, S., Mutka, M.: Efficient Mobile Access to Internet Data via a Wireless Peer-to-Peer Network. IEEE Int'l Conference on Pervasive Computing and Communications (2004) 197–205
4. Karbhari, P., Zegura, E., Ammar, M.: Multipoint-to-Point Session Fairness in the Internet. In: Proceedings of INFOCOM 2003. (2003) 207–217
5. Castro, M., Druschel, P., Kermarrec, A., Nandi, A., Rowstron, A., Singh, A.: SplitStream: High-Bandwidth Multicast in Cooperative Environments. (2003) 298–313
6. BitTorrent: The BitTorrent file distribution system, `http://www.bittorrent.org`. (2006)
7. Cheng, W., Chou, C., Golubchik, L., Khuller, S., Wan, Y.: Large-scale Data Collection: a Coordinated Approach. In: Proceedings of INFOCOM 2003. (2003) 218–228
8. Tseng, Y., Ni, S., Chen, Y., Sheu, J.: The Broadcast Storm Problem in a Mobile Ad Hoc Network. Wireless Networks **8** (2002) 153–167
9. Lim, H., Kim, C.: Multicast Tree Construction and Flooding in Wireless Ad Hoc Networks. (2000) 61–68
10. Qayyum, A., Viennot, L., Laouiti, A.: Multipoint Relaying for Flooding Broadcast Messages in Mobile Wireless Networks. (2002) 3866–3875
11. Peng, W., Lu, X.: AHBP: An Efficient Broadcast Protocol for Mobile ad hoc Networks. Journal of Science and Technology (JCST) - Beijing, China **16** (2001) 114–125
12. Calinescu, G., Mandoiu, I., Wan, P., Zelikovsky, A.: Selecting Forwarding Neighbors in Wireless Ad Hoc Networks. (2001) 34–43
13. ns2: The network simulator, `http://www.isi.edu/nsnam/ns`. (2006)

Session Key Reuse Scheme to Improve Routing Efficiency in AnonDSR

Chunum Kong, Min Young Chung, and Hyunseung Choo*

School of Information and Communication Engineering
Sungkyunkwan University
440-746, Suwon, Korea +82-31-290-7145
{cukong,mychung,choo}@ece.skku.ac.kr

Abstract. The importance of security in ad hoc network is increasing gradually to deliver information safely among nodes in hostile environment. The data is encrypted using various encryption techniques for the security reinforcement or for hiding the communication path. AnonDSR which offers anonymity to encrypt communication path, guarantees anonymity efficiently. The anonymity of source and destination nodes is guaranteed through 3 protocols. However, secret keys and route pseudonyms must be newly created for security, whenever an anonymous communication session occurs. It generates large overhead.Therefore, the proposed scheme reduces overhead of AnonDSR to reuse symmetric keys and route pseudonyms during a certain period which is defined by user. It is possible so that the data is encrypted by symmetric key which is shared between source and destination nodes, because intermediate nodes cannot decrypt the data. This scheme maintains security features of AnonDSR to perform anonymous communication, and only performs anonymous data transfer protocol when duplicate session is occurred. Then, the route setup time is improved a minimum of 47.1% due to the decrease of route setup procedure.

Keywords: Anonymous Ad Hoc Routing, Key Reuse, and Onion.

1 Introduction

When nodes achieve wireless communication in hostile environment, the nodes vulnerable to attack from malicious nodes can become the target of forgery or modification. Anonymous routing schemes which provide security for communication path are researched recently, and the necessity in several fields is increased for more secure communication.

There are representative schemes such as SDAR [2], ANODR [1], and Anon DSR [3]. These schemes maintain anonymity using the trapdoor technique [4]. This technique is method to distinguish destination node. The destination node is decided if a node can decrypt the encrypted message at source node, because this message is encrypted by secret key which is shared between source and

* Corresponding author.

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 303–311, 2007.
© Springer-Verlag Berlin Heidelberg 2007

destination nodes. SDAR uses a public key in the trapdoor. The data are attached continuously after encryption by a temporary public key. ANODR uses a symmetric key in the trapdoor, and does not use secret key when this scheme encrypts data. And AnonDSR is used to mix a symmetric key and a public key. This encrypts the data using the onion technique [5].

AnonDSR solves the limitation of SDAR and ANODR. It consists of 3 protocols. First, the security parameter establishment protocol sets the symmetric key, shared only with source and destination nodes, for secure communication. Second, the anonymous route discovery protocol is that source and destination nodes have route pseudonyms and symmetric keys instead of intermediate nodes' identity for anonymous communication. Finally, the anonymous data transfer protocol enforces security of data using the onion technique which uses symmetric keys of each node. AnonDSR creates a new session to guarantee anonymity. This session must always update the symmetric key and route pseudonym. Otherwise, it causes the problem that the source node's identity is exposed from the attack of malicious nodes. Both create considerable overhead by regenerating the symmetric key and route pseudonym for anonymous communication.

Therefore, the proposed scheme reuses symmetric keys and route pseudonyms during a certain period which is defined by user, in spite of changing a session. Because it improves encryption of AnonDSR by adding a further encryption, which shares the symmetric key with source and destination nodes, and intermediate nodes cannot decrypt this symmetric key. As a result, the proposed scheme only executes anonymous data transfer protocol instead of 3 protocols when duplicate session occurs at anonymous communication. The route setup time is decreased as the number of occurrence at duplicate session and security is enhanced increasing the number of encryption at data transfer protocol. If duplicate session happens, performance of route setup time is improved by a minimum of 47.1%. At this time, as the used keys have a persistence of a certain period, it leads to confusion if session key terminology is used. Consequently, this is called a long term session key.

The rest of the paper is organized as follows. Section 2 describes anonymous ad hoc routing schemes. The proposed scheme is illustrated in Section 3. Section 4 analyzes security and anonymity of the proposed scheme compared to that of existing schemes. Section 5 concludes this paper.

2 Related Works

We review the comparison of ANODR and AnonDSR for anonymous ad hoc routing schemes and discuss their characteristics and deficiencies. Each scheme exchanges messages in RREQ and RREP, and is decided features according to encryption techniques. We use the notations and terminology shown in Table 1.

Table 2 shows RREQ and RREP format of anonymous routing schemes that include common message type, route pseudonym, trapdoor, and route encryption format. RREQ and RREP mean a message type. The route pseudonym identifies a node by random number, instead of ID, which guarantees anonymity.

Table 1. Notations and terminology

ID_A	Identity for node A	K_X	A random symmetric key
N_X	A random nonce	K_A	Symmetric key for node A
N_A	A random nonce for node A	H()	A one way hash function
PK_{temp}	Temporary public key	PK_A	Public key for node A
SK_A	Private key for node A	P	Padding
PL	Padding length	$Sign_A$	Signature for node A
$E_K(M)$	A message M encrypted with a symmetric key K	$E_{PK}(M)$	A message M encrypted with a public key K

Table 2. RREQ and RREP format of anonymous routing schemes

	Phase	Message type	Security level	Temporal public key	Route pseudonym	Unique sequence num	Real ID	Routing pass	Trapdoor	Route encryption format
ANODR	–	RREQ	–	–	–	Seqnum	–	–	tr_{dest}	onion
	–	RREP	–	–	N_{dest}	–	–	–	pr_{dest}	onion
AnonDSR	1	RREQ	SecType	–	–	Seqnum	ID_{src}, ID_{dest}	RRec	SecPara	–
		RREP	SecType	–	–	Seqnum	ID_{src}, ID_{dest}	RRec	SecPara	–
	2	ANON-RREQ	–	PK_{temp}	–	–	–	–	tr_{dest}	onion
		ANON-RREP	–	–	N_{next}	–	–	–	–	onion
	3	ANON-DATA	–	–	N_{src}	–	–	–	–	onion

The trapdoor represents a technique where only the destination node can check the message received from the source node. Generally, this is encrypted data using the shared secret key between source and destination nodes. The route encryption format represents a method of encrypting data. The onion technique repeatedly encrypts data its own secret key at a node.

3 Proposed Scheme

This scheme consists of 3 protocols: Security Parameter Establishment Protocol (SPEP), Anonymous Route Discovery Protocol (ARDP), and Anonymous Data Transfer Protocol (ADTP). SPEP and ARDP consist of RREQ and RREP, ADTP exchanges data.

We assume that firstly, like AnonDSR, the public key of each node is distributed to all the nodes in a network by Certificate Authority (CA), and secondly, the proposed scheme does not use the general concept of session key, and it defines a long term session key in order to use secret key and route pseudonym continuously even if a session is changed. Also, the route pseudonym created between nodes uses the original node continuously and is available to control the terms of validity by the user.

3.1 Security Parameter Establishment Protocol

To ensure secure communication, the SPEP exchanges a secret key between source and destination nodes. The secret key is shared and managed by only

the source and destination nodes. Route pseudonym (N_T) and symmetric key (K_T) are created at the source node. N_T and K_T mean secret index and shared secret key, respectively. The source node should maintain the N_T and K_T as a table. It is possible to check the previous communication session by this table. If there is the same secret key in the table, only anonymous data transmission is performed. 3 protocols are performed sequentially when the same secret key does not exist in the table. Route pseudonym (N_T) and symmetric key (K_T) are used as the concept of long term session key.

In the RREQ, the source node broadcasts RREQ messages and uses the trapdoor technique, which encrypts the public key of the destination node. The RREQ is composed as follows:

<RREQ, SecType, seqnum, ID_{src}, ID_{dest}, RRec, SecPara>

Where RREQ is the type of message; SecType chooses the degree of security in RREQ; seqnum is a unique sequence number; ID_{src} and ID_{dest} are the identity of source node and destination node, respectively; RRec is the source route record [3]; and SecPara is the security factor provided by the source node. When security and anonymity are required, SecPara is used by the trapdoor technique keeping $E_{PK_dest}(N_T, K_T, \text{Para}), Sign_{src}$ where PK_{dest} is a public key of the destination node. $Sign_{src}$ is a signature that encrypts basic elements of the source node using a hash function. Only the destination node has a private key (SK_{dest}) and can confirm a route pseudonym (N_T) and symmetric key (K_T) decrypting PK_{dest}.

The RREP broadcasts the RREP message from the destination node to neighborhood nodes. RREP is composed as follows:

<RREP, SecType, seqnum, ID_{src}, ID_{dest}, RRec, SecPara>

RREP is the type of message. It is identical with RREQ process except that the public key in SecPara is only replaced with PK_{src}, the public key of the source node.

3.2 Anonymous Route Discovery Protocol

The source and destination nodes can have the entire route using the trapdoor and onion mechanism. The non secure and secure communication methods of AnonDSR are not changed and the process of anonymous communication method is modified.

The RREQ improves the encryption process of the onion, which is the Path Discovery Onion (PDO). It is encrypted data at the source node using the symmetric key (K_T), shared only with source and destination nodes of the SPEP, by use of a long term session key. The RREQ is composed as follows:

<ANON-RREQ, PK_{temp}, tr_{dest}, onion>

Where ANON-RREQ is the type of RREQ message that requires anonymous communication; PK_{temp} is a temporary public key created at the source node,

and is used to encrypt the data of intermediate nodes; tr_{dest} is a trapdoor technique and only the destination node can decrypt it by encrypting a symmetric key. For example, $tr_{dest} = N_T$, $E_{K\text{-}T}(ID_{dest}, SK_{temp})$; the onion, route encryption format, is used to encrypt a symmetric key (K_X) and route pseudonym (N_X) which each session is created newly at intermediate nodes. The onion, route encryption format, process is called the Path Reverse Onion (PRO) and it is created via PDO in reverse order. Fig. 1 shows the PDO and PRO processes.

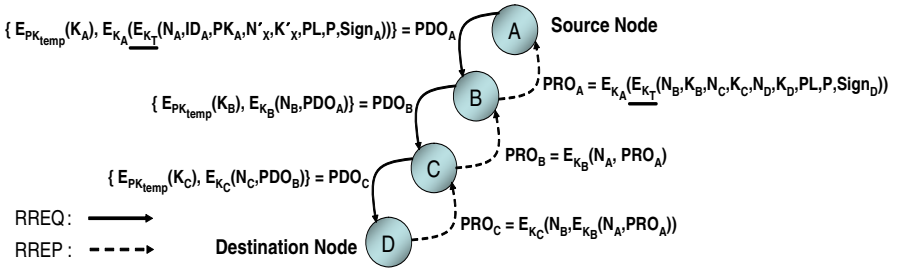


Fig. 1. PDO and PRO processes

Prior to encryption, it is appended the process of data encryption by symmetric key (K_T) generated in the SPEP. PDO at the source node is composed as follows:

$$PDO_A = E_{PK\text{-}temp}(K_A),\ E_{K\text{-}A}(E_{K\text{-}T}(N_A,ID_A,PK_A,N'_X,K'_X,\text{PL,P},Sign_A))$$

Where a route pseudonym (N_X) and symmetric key (K_X) are the long term session key to create each node independently; $Sign_A$ is the signature of ID, N_X, and K_X which are encrypted the private key (SK_A) of the source node (A); N'_X, K'_X means the used secret key and route pseudonym when the next Session has another route. The source node encrypts its own information using its own symmetric key (K_A) and the sharing symmetric key (K_T). This symmetric key (K_A) is encrypted by the temporary public key (PK_{temp}) created at the source node, and this process repeats whenever a node moves. If the destination node is reached, tr_{dest} as the trapdoor technique is decrypted by symmetric key (K_T). Then, the destination node can decrypt the symmetric keys of each node after it obtains the temporary private key (SK_{temp}), and decrypts the data of PDO encrypted by the temporary public key (PK_{temp}).

The RREP improves the encryption process of onion. The PRO adds the encrypting process by symmetric key (K_T), the long term session key, prior to encrypting the symmetric key of each node. RREQ is composed as follows:

$$<\text{ANON-RREP},\ N_{next},\ \text{PRO}>$$

Where ANON-RREP is the type of RREP message that requires anonymous communication; N_{next} is updated whenever a node moves, because this means the next route pseudonym. PRO, route encryption format, represents the onion technique and is created in PDO reverses order; PRO at the source node is composed as follows:

$$PRO_A = E_{K_A}(E_{K_T}(N_B,K_B,N_C,K_C,N_D,K_D,\mathrm{PL},\mathrm{P},Sign_D))$$

This encrypts the symmetric key of each node by the onion method in reverse route to inform total route pseudonyms and symmetric keys on route to the source node (A). The PRO is encrypted by the shared symmetric key (K_T) at first.

3.3 Anonymous Data Transfer Protocol

The source and destination nodes already have all symmetric key and route pseudonym on the routing route. These are used to encrypt data. Only each node on the routing route is encrypted a part of data by the onion mechanism. Prior to encryption, security is augmented by appending the data encryption process with the symmetric key (K_T) generated in the SPEP. If the intermediate node that has all symmetric keys (K_X) on the route does not contain the sharing symmetric key (K_T), security is enhanced as the encrypted data cannot be decrypt. The anonymous data transfer message is composed as follows:

$$<\text{ANON-DATA}, N_{src}, \text{onion}>$$

Where ANON-DATA is a message that informs a data transmission; N_{src} represents a route pseudonym of a starting node initially and is shifted by the route pseudonym of the next node whenever arriving at a node on the path; Onion, route encryption format, encrypts the data with symmetric keys which are collected at the previous two protocols. When transmitting, it is called Anonymous communication Data Onion (ADO) and when receiving, it is called Reverse anonymous communication Data Onion (RDO). Fig. 2 shows ADO and RDO processes.

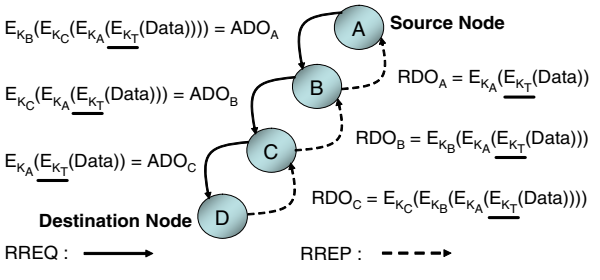


Fig. 2. ADO and RDO processes

4 Performance Evaluation

4.1 Analysis of Security, Anonymity, and Scalability

The simulation is performed on an Intel Pentium 4 Processor with 2.60GHz clock generator and, 768MB RAM. The network consists of 500 nodes and each node has 4 neighbors. It is implemented in C. The processing overhead of each encryption technique [1,3] is given Table 3.

The route setup time of each scheme can be calculated using the actual encrypting and decrypting time of encryption schemes in Table 3, and the encrypting and decrypting number in Table 4. The route setup time of the proposed scheme and AnonDSR are almost identical. However, overhead is decreased and security is enhanced to increase the encrypting number of data if duplicate session is occurred. Therefore, the proposed scheme is more efficient than AnonDSR.

Table 3. Processing overhead of encryption schemes

Mechanism		Encrypting Time	Decrypting Time
AES(128bit)		128Mbps	128Mbps
RSA	(1024bit)	1ms	97ms
	(2048bit)	4ms	712ms
SHA-1		161Mbps	161Mbps

Table 4. The number of encryptions/decryptions in anonymous routing schemes

Contents		Protocols	SDAR	ANODR	AnonDSR	Proposed Scheme
RREQ	Intermediate Nodes	Symmetric Key (Encrypting/Decrypting)	N	2n	2n	2n
		Public Key (Encrypting)	n	0	n	n
		Private Key (Decrypting)	n	0	0	0
	Source and Destination Nodes	Symmetric Key (Encrypting/Decrypting)	1	3	3	4
		Public Key (Encrypting)	L	0	2	2
		Private Key (Decrypting)	L	0	L+1	L+1
RREP	Intermediate Nodes	Symmetric Key (Encrypting/Decrypting)	n	n	n	n
		Public Key (Encrypting)	0	0	0	0
		Private Key (Decrypting)	0	0	0	0
	Source and Destination Nodes	Symmetric Key (Encrypting/Decrypting)	L+1	1	L+1	L+3
		Public Key (Encrypting)	0	0	1	1
		Private Key (Decrypting)	0	0	1	1
Summation		Symmetric Key (Encrypting/Decrypting)	2n+L+2	3n+4	3n+L+4	3n+L+7
		Public Key (Encrypting)	n+L	0	n+3	n+3
		Private Key (Decrypting)	n+L	0	L+2	L+2

The total number of encryption and decryption concerned with symmetric keys and public keys in anonymous routing schemes are compared Table 4 to analyze the scalability of computing. AnonDSR and the proposed scheme consider the SPEP and the ARDP since SDAR and ANODR do not have the ADTP.

In Table 4, n means the number of different RREQ and RREP messages on an ad hoc network and L means the number of hop of a RREQ and RREP message from source node to destination node. Total route setup time and scalability are identical almost, because the total number of decryption concerned with public key in the proposed scheme are identical with AnonDSR in each anonymous scheme.

4.2 Performance Comparison of Proposed Scheme and AnonDSR

The proposed scheme is more secure than AnonDSR as the proposed scheme has a higher encrypting number of ADTP. In case of the same communications session, the encrypting and decrypting number of the entire process must be known including the ADTP. Therefore, it can calculate the route setup time using the encrypting and decrypting time of encryption schemes in Table 3.

The proposed scheme and AnonDSR are calculated using the encrypting and decrypting number and the processing time of the encryption scheme. If the same session occurs, the overhead of the proposed scheme is reduced, as shown in Fig. 3, when the duplicate session occurs initially. The efficiency improves 63.3% $\sim$ 64.5% when the same session occurs twice.

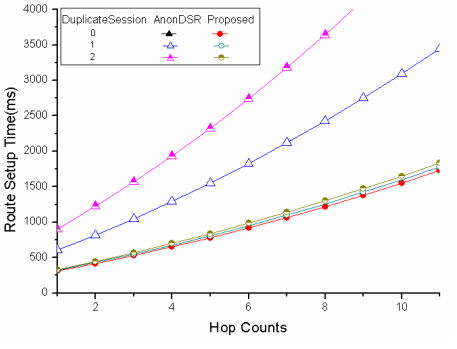


Fig. 3. Route setup time of duplicate session.

5 Conclusion

In this paper, we research a better anonymous routing scheme to solve the problems of AnonDSR, which uses the trapdoor and onion method effectively. The proposed scheme improves the encryption techniques of AnonDSR using a long term session key which is kept for a certain period of time. The proposed scheme can use the symmetric key and route pseudonym continuously, despite a changing session. Because it improves encryption of AnonDSR by adding an additional encryption procedure, which shares the symmetric key only with source and

destination nodes. As a result, We know that the route setup time is improved a minimum of 47.1%, and the efficiency of route setup time is improved increasing the occurrence of duplicate session.

Acknowledgments. This research was supported by MIC, Korea under ITRC IITA-2006-(C1090-0603-0046).

References

1. Kong, J., Hong, X.: ANODR: Anonymous on Demand Routing with Untraceable Routes for Mobile Ad-Hoc Networks. ACM Symposium (2003) 291-302
2. Boukerche, A., El-Khatib, K., Korba, L., Xu, L.: A Secure Distributed Anonymous Routing Protocol for Ad Hoc Wireless Networks. Journal of Computer Communications (2004)
3. Song, R., Korba, L., Yee, G.: AnonDSR: Efficient Anonymous Dynamic Source Routing for Mobile Ad-Hoc Networks. SASN (2005) 32-42
4. Yao, A.: Theory and Applications of Trapdoor Functions (Extended Abstract). Symposium on Foundations of Computer Science (1982) 80-91
5. Goldschlag, D., Reed, M., Syverson, P.: Onion Routing for Anonymous and Private Internet Connections. Communication of the ACM (1999) 39-41

Clustering in Ad Hoc Personal Network Formation

Yanying Gu, Weidong Lu, R.V. Prasad, and Ignas Niemegeers

Center for Wireless and Personal Communications (CWPC),
Delft University of Technology, Delft, The Netherlands
{y.gu,w.lu,vprasad,i.niemegeers}@ewi.tudelft.nl

Abstract. Personal Network (PN) is a user-centric concept to realize the interconnection of the users' various devices and networks, such as home networks, corporate networks and vehicle area networks, at any time and any place. These networks may be geographically separated from each other and are usually organized in an ad hoc fashion, where the devices in one network can share content, data, applications and resources with each other, and communicate with the outside world. Clustering for personal networks intends to organize these ad-hoc networked heterogeneous devices into hierarchical clusters. A new clustering scheme, Personal Network Clustering Protocol (PNCP), is proposed for PNs and compared with clustering algorithms in the literature in terms of various performance metrics. PNCP performs the cluster formation based on heterogeneity of PNs in a distributed way, forms clusters with a moderate number of nodes, limits the cluster radius to *k*-hops, decreases the change in the cluster composition, extends the average membership time in the clusters providing stability and generates less overhead traffic. Though it was aimed for PNs, this protocol can also be used for any other heterogeneous ad hoc network for clustering.

Keywords: Personal network, Clustering, Ad hoc network, Personal Network Clustering Protocol.

1 Introduction

Although several existing technologies can offer a part of the solutions to communication needs of a person such as WPAN, WLAN, the Internet, and UMTS networks etc., little has been done to integrate these technologies and emerging ones to meet the future demands of a user. Thus the necessity is to combine all the devices of a user and networks into one single network which is transparent to the user. A Personal Network (PN) [1] supports a user's need to have access to his personal and public services with his devices. A PN covers all types of devices that can be used by a person in different places by using the service of interconnecting.

From the point of view of a user, the devices belonging to the owner of that PN are called personal devices (or personal nodes) [1], as shown in Fig. 1. The nodes having full networking functionalities such as packet forwarding, mobility management and computational capacities are full function nodes (FFN) [2]; the other nodes are called

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 312–319, 2007.
© Springer-Verlag Berlin Heidelberg 2007

reduced function node (RFN) [2]. For example, a RFN does not forward packets or only forward (signaling) packets with high QoS requirements or packet catering to special purposes. Moreover, to form a secure private network and to achieve efficient communication between many devices of a PN, personal devices are naturally organized in groups, in which personal nodes can share their content, resources, etc. and cooperate with each other to support all the activities of the owner.

When personal devices are spread geographically, clustering is a way to organize the devices of a PN in the close vicinity of each other into hierarchical groups, which can facilitate efficient routing, addressing and self-organization of a PN. The clustered architecture guarantees a better performance in an ad hoc mobile network with a large number of mobile terminals [3]. Based on the properties of PNs, there are some specific requirements for the clustering scheme in PNs: (a) it must be distributed and self-organized [1]; (b) considering the privacy and the security in PNs, personal nodes and foreign nodes need to be included in personal clusters and foreign clusters [2]; (c) the proposed clustering scheme should give solutions for both FFN and RFN [2]; (d) since personal nodes are usually moving together in groups, the proposed clustering method should reduce the influence of the group mobility on the existing cluster relationship [5]; (e) the heterogeneity of personal nodes should be considered in the selection of a cluster head in each cluster [2]; (f) the life-time of the personal clusters should be long so that the influence of re-clustering on the applications and in turn the traffic load could be reduced [3]; (g) the cluster radius and the number of nodes in each personal cluster should be controlled in order to offer efficiency in PNs [3].

To efficiently manage a PN, we introduce a clustering scheme called Personal Network Clustering Protocol (PNCP) to generate a cluster structure to bridge different technologies and offer a homogeneous and a transparent view of a PN to the user. PNCP considers the heterogeneity of PNs in cluster formation in a distributed and self-organized way, forms clusters with a moderate number of nodes, limits the maximum number of hops of cluster radius [3], results in less variations cluster dynamics change, longer average membership time and less overhead compared to other clustering schemes.

The remainder of this paper is organized as follows. In Section 2, some clustering algorithms reported in the literature are presented. In Section 3, the descriptions of the proposed clustering protocol (PNCP) are highlighted. Section 4 offers explanations of various simulation results, which include the performance analysis and comparison between PNCP and other clustering algorithms. Finally, Section 5 summarizes our paper with conclusions, which include the contributions and performance improvements of PNCP.

2 Related Work

Many papers focused on presenting effective and efficient clustering schemes [3] for ad hoc networks. In the Lowest-ID clustering algorithm [6], the node with lowest ID in its neighborhood acts as a cluster head (CH). However, the node with a lower ID would stay as a CH for a long period of time, which may cause the node draining the battery quickly. Least Cluster head Change (LCC) algorithm [3] still selects the nodes with lower ID as CHs and reduces the frequent re-clustering in Lowest_ID. Re-clustering is

event-driven when two CHs move into the radio range of each other, one of them gives up the cluster head role. LCC significantly improves cluster stability by relinquishing the requirement that a CH should always bear some specific attributes in its local area.

Connectivity-based *k*-hop clustering (Highest_C) algorithm [8] is proposed to use connectivity as a primary and lower ID as a secondary criterion in selecting cluster heads. It generalizes the connectivity to count all *k*-hop neighbors of the given node. However, Highest_C does not offer a stable cluster structure, because topology changes may dramatically influence nodes' degree or connectivity.

Ohta et al, [9] have proposed a clustering scheme, which controls the number of nodes in each cluster between an upper and a lower bound. However, it results in frequently cluster division and merging. Mobility based clustering (MOBIC) introduces a novel relative mobility metric to offer a more stable cluster structure for mobile ad hoc networks [5]. MOBIC is feasible and effective for ad hoc networks with group mobility behavior.

In the On Demand Weighted Clustering Algorithm [11], a combined-metric factor is calculated to select a CH. The combined metric considers node degree, node distance, absolute speed and battery power. However, the relative mobility should be used instead of absolute mobility as a metric in the CH selection [5].

Different clustering approaches typically focus on one specific performance metric. However, the requirements of a PN clustering scheme explained in Section 1 have not been taken into account at the same time in the above algorithms. PNCP incorporates all the aforementioned requirements and fulfill these to the extent possible.

3 Personal Network Clustering Protocol (PNCP)

Personal Network Clustering Protocol (PNCP) enables distributed and self-organized cluster formation and maintenance without any information, such as location, offered by other systems, such as GPS. Fig.1 gives an overview of the cluster structure in PNs. PNCP organizes personal nodes [1] to form personal clusters (*Requirement b* in Section 1). Master Node (MN) [2] is a new term introduced in PN, which provides many PN specific functions as well as all the functionalities of a cluster head (CH) of a personal cluster. In this paper, we only discuss MN functionalities as a CH. In order to comply with PN terminology, we refer to CH as MN, and use these terms interchangeably in the rest of this paper. The selection of MN is based on the node's capability to act as a MN, which is described by Node Capacity (*NC*) and Mobility (*M*)

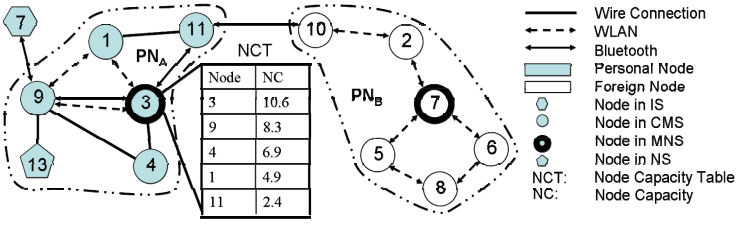


Fig. 1. Personal network clustering protocol

(*Requirement e*). We describe the protocol using a model in which we associate each node with a state, which is explained in Fig. 1. Master Node State (MNS) and Cluster Member State (CMS) are defined for the MNs and CMs [2]. Initial State (IS) is taken by a personal node not belonging to any cluster, such as node 7 in Fig. 1. Further, we define a Non-clustered State (NS) for two kinds of personal nodes: (i) a RFN node chooses a CM as its gateway to access a personal cluster (*Requirement c*); and (ii) a CM may temporarily move to *k+1* hops away from the master node and the node can set another CMS node in the same cluster as a gateway to extend its membership of the cluster (*Requirement f*) [10].

3.1 Node Capacity and Mobility

Many earlier clustering schemes perform cluster formation based on a single metric, for example, node identity, connectivity, mobility, etc. [3]. The On Demand Weighted Clustering Algorithm [11] tried to use a combined metric to achieve overall improvement. *NC* explained in Eq. (1) is also a combined metric and proposed to consider the heterogeneous properties of personal nodes in PNs for the role of a MN. *NC* considers Connectivity (*C*) and available Resources (*R*) of a personal node, where *C* indicates the influence of neighbors on the *NC* and *R* represents the internal capacity of the node.

$$NC = \alpha_1 * C + \alpha_2 * R \tag{1}$$

where α_1 and α_2 are the weight factors of *C* and *R*. Geng Chen [8] has used connectivity as the criteria to select a CH, where each neighbor in *k*-hop neighborhood make an equal contribution to *C*. Due to the location and capability difference of these neighbors, they may have different impacts on *C*. So we define connectivity *C* in Eq. (2)

$$C = \sum_{i=1}^{K} \frac{1}{i} \sum_{j}^{\mathrm{Ni(v)}} (T_j - M_j) \tag{2}$$

where *i* is the i^{th} neighborhood of node *v*, *k* is the cluster radius in each cluster, *Ni(v)* denotes the nodes in the i^{th} neighborhood, T_j and M_j are the transmission range and the mobility of the j^{th} node in the i^{th} neighborhood. The calculation method of M_j is the same as M_Y in [5]. Because the contribution of a neighbor decreases as the hop distance increase, we divide the contribution of each neighbor by its hop distance. Further for each neighbor, when it has a higher value of mobility, it may leave the neighborhood easily; with a larger transmission range it is not easy for the node to leave the neighborhood. Thus we use the result of *T* minus *M* to indicate the heterogeneity of each neighbor.

For *R*, we take into account available memory (*A*), battery power (*E*) and CPU resources (*P*). We calculate the available resources, R_i, for node *i*, in Eq. (3):

$$R_i = (\chi_1 * A_i + \chi_2 * P_i) * E_i \tag{3}$$

The parameters, χ_1 and χ_2 are the weight factors. Although a node may have more available memory and CPU resources, it still can not use these resources if it

has a low level of battery power. Thus we multiply the sum of the available memory and the CPU resources by E_i to indicate the decisive role of E_i in the calculating of R.

For Mobility, we use relative mobility ($M_i^{rel}(j)$) instead of absolute mobility, which has been proposed in [5]. And the Mobility (M_i^j) of a personal node j at a node i is the absolute value of $M_i^{rel}(j)$, so the mobility metric is always positive. Thus a IS node chooses a MN with a lower relative mobility (M) and joins the cluster to act as a CM; and a IS node with highest ability (NC) can change to MN, which improves the stability of the formed cluster structure by the effort made by both MNs and CMs. NC and M are used extensively in the cluster formation and maintenance algorithm, which will be explained in the following sections.

3.2 Cluster Formation

The actions taken in the cluster formation algorithm are explained as follows:

1: A IS node discovers its personal neighbors, calculates its NC and relative M, and saves its NC and all its personal neighbors' NC in node capacity table (NCT) as shown in Fig. 1.
2: The IS node, as shown by node 7 in Fig. 1, looks for a MN with lower M from the existing neighbor personal clusters.
3: The IS node unicasts a hello message as a request to the MN as node 3 in Fig. 1 to apply to join the cluster.
4: The MN controlling the number of nodes in the cluster sends a response message as a decision to add the IS node in the personal cluster.
5: If step 1 to 3 are not successful, the IS node with highest NC amongst all the IS neighbors changes to MNS, which is shown as node 3 in Fig. 1.
6: Other IS nodes follow *Step 1* until either joining a cluster or changing to MNS.

From the above steps, our cluster formation algorithms has these contributions: (1) the cluster formation is only performed among personal nodes as indicated by *Step 1*, which meets the *Requirement a*; (2) personal nodes with similar mobility pattern in the neighborhoods join to the same cluster, because personal nodes choose the MN with the lower M and join the cluster as described by *Step 2*, which meets the *Requirement d*; (3) existing cluster structure is not influenced by the new nodes starting in PN, because they always first try to join the existing personal clusters instead of forming new clusters as described in *Step 2 and 3*, which meets the *Requirement f*; (4) the MN controls the number of nodes in the cluster below U by sending acceptation messages to the IS nodes trying to join the cluster as stated in *Step 4*, which meets the *Requirement g*; (5) the node with highest NC in the local neighborhood can declare itself as a MN as node 3 in Fig. 1 as described in *Step 5*, so the MN selection considers the heterogeneity of nodes and meets the *Requirement e*.

3.3 Cluster Maintenance

The aims of the cluster maintenance algorithm are to increase the stability of the existing cluster structure in case of mobility in PN (*Requirement f*) and to control the

number of nodes (n) in each cluster (*Requirement g*) [3], [9]. To improve the stability of the cluster structure, re-clustering is carried out only in the following two cases:

1. A MN does not have enough resource (e.g. power) to act as a MN.
2. To control the n in each cluster, two clusters may merge into one cluster to increase the n in the cluster by "Cluster Merge" [9].

MNs are responsible for controlling n in each personal cluster. To reduce frequent cluster division and merging [9], there is no cluster division in our algorithm. Each IS node sends a request to apply to join a cluster and the MN will decide whether to accept it or not to control n below U. And to avoid small clusters, MNs checks n periodically to perform "Cluster Merge". Both cluster radius [3] ($<=k$) and cluster size ($L<n<U$) [9] are considered as requirements in "Cluster Merge". Based on these requirements, a MN (requesting MN) having a lower n selects a neighbor MN (selected MN) and sends a request to it. Considering these requirements, the selected MN may accept the requesting MN by sending a response. Then the requesting MN changes to CMS and all the CMs in the personal cluster join the selected cluster.

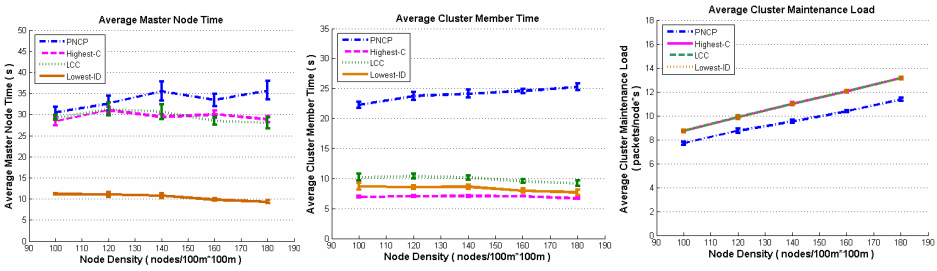
4 Performance Evaluation

PNCP is implemented in ns-2 [12] and is compared with other clustering algorithms for performance evaluation. The parameters of the simulation environments are set as shown in Table 1. Some performance metrics are proposed as the basis for evaluating the performance of the proposed clustering protocol: Average Master Node Time (AMNT), which is defined as the average period of time during which a master node plays a central controller role continuously; Average Cluster Member Time (ACMT)**;** Average Cluster Maintenance Load (ACML).

Table 1. Common set-up values for the simulations

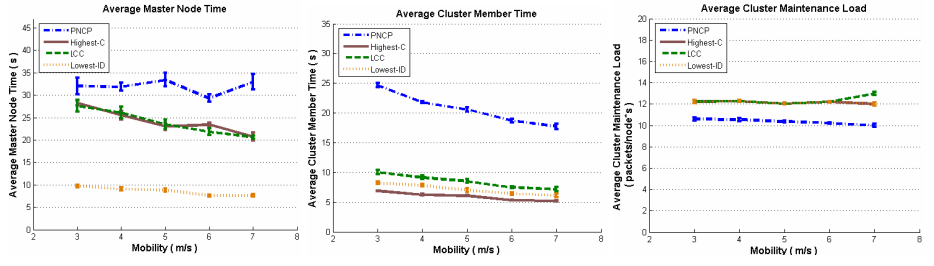
Properties	Value	Properties	Value
Terrain Dimension	100m*100m	Hello Interval	1s
Simulation Time	100s	Cluster Size Upper Limit (N)	30 nodes
MAC Layer Type	802.11	Cluster Size Lower Limit (L)	5 nodes
Weight Factor α_1, α_2, χ_1 and χ_2	1	Cluster Radius (k)	2 hops
Transmission range	15m		

The experiments are carried out to compare Highest_C, LCC, Lowest_ID and the proposed PNCP. Based on Fig. 2 (a), (b) and Fig. 3 (a), (b), PNCP has the highest AMNT and ACMT in all situations, which means that PNCP can offer a more stable cluster structure from the MN and CM perspective. Using *NC* gives this performance improvement, because the calculation of C considers the different contribution of each neighbor. Moreover, the power level (E) plays a decisive role on the other kinds of available resources.



(a) (b) (c)

Fig. 2. Simulation results vs. node density with maximum speed 3m/s



(a) (b) (c)

Fig. 3. Simulation results vs. node mobility with 90 nodes

From Fig. 2 (c) and Fig. 3 (c), PNCP has a lower ACML than the other clustering algorithms. The main reason is that the non-clustered state (NS) for a personal node to attach to its original personal cluster through a gateway. A NS node unicasts member messages to its one-hop gateway, which generates less traffic than a normal node that broadcasts member messages to personal neighbor nodes within *k*-hops.

Based on the simulations with different numbers of nodes and node mobility, PNCP provides some performance improvements: decreasing re-clustering, extending the average membership in the clusters, having a moderate number of nodes in each cluster, limiting cluster radius and generating less overhead. However, the study still has some limitations. Firstly, because of the limitation of ns-2, the simulation scenarios are not typical for PNs, however it is used here to study any ad hoc network simulation. Secondly, only fully functional nodes having complete networking functionality are considered in the simulation.

5 Conclusion

Clustering is the way ahead for organizing the heterogeneous nodes of PNs and in fact any ad hoc networks. We introduced a Personal Network Clustering Protocol (PNCP) for the distributed cluster formation and maintenance that are highly adaptive to the characteristics of PNs and overcome limitations of its predecessors in the literature.

PNCP meets the PN clustering requirements as explained in Section 1 and can form a more stable cluster structure with a lower overhead in most of the cases. In addition, PNCP considers heterogeneity in PNs, forms clusters with moderate number of nodes, and limits the cluster radius within *k*-hops. PNCP is a generic protocol which can be used to form clusters in any ad hoc networks. Future research in PNs can benefit from the cluster structure provided by PNCP. A flexible routing protocol exploiting the hierarchical structure formed by PNCP is needed to provide a better performance for PNs. Furthermore, a security method for PN clustering should be specified and analyzed along with the clustering. The field of PNs is rapidly growing and changing, while there are still many challenges that need to be met.

References

1. Ignas G. M. M. Niemegeers, Sonia M. Heemstra de Groot: Research issues in ad-hoc distributed personal networking. Wireless Personal Communications: An International Journal, Volume: 26, Issue: 2-3, Kluwer Academic Publishers (2003) 149–167
2. W. Lu (ed.): D1.2 Initial Architecture of Personal Networks. IOP GenCom Project QoS for Personal Network at Home Deliverable (2005)
3. J. Y. Yu and P. H. J. Chomg: A Survey of Clustering Schemes for Mobile Ad Hoc Networks. IEEE Communications Surveys and Tutorials, Vol. 7, No. 1 (2005) 32-48
4. M.R. Pearlman and Z.J. Haas: Determining the Optimal Configuration for the Zone Routing Protocol. IEEE JSAC, vol 17 (1999) 395-414
5. P. Basu, N. Khan, and T. D. C. Little: A Mobility Based Metric for Clustering in Mobile Ad Hoc Networks. In Proc. of Workshop on Wireless Networks and Mobile Computing (2001)
6. M. Gerla and J. T.C. Tai: Multi cluster, Mobile Multimedia Radio Networks. Wireless Networks 1 (1995) 255-265
7. C. C. Chiang, H. K. Wu, W. Liu and M. Gerla: Routing in Clustered Multihop, Mobile Wireless Networks with Fading Channel. In Proc. IEEE SICN (1996) 197-211
8. F.G. Gonzalez, J.S. Gonzalez: Connectivity based k-hop clustering in wireless networks. Proc. 35th Hawaii International Conference on System Sciences (2002)
9. T. Ohta, N. Murakami, R. Oda, and Y. Kakuda: An Improved Autonomous Clustering Scheme for Highly Mobile Large Ad Hoc Networks. Workshops on AHSP (2005) 655-660
10. J. Y. Yu, and P. H. J. Chong: 3hBAC (3-hop between adjacent clusterheads): A novel non-overlapping clustering algorithm for mobile ad hoc networks. Proc-IEEE PacRim 03, Canada (2003)
11. M. Chatterjee, Sajal K, D. Turgut: An on-demand Weighted Clustering Algorithm (WCA) for Ad hoc networks. Proceedings of IEEE GLOBECOM (2000) 1697 – 1701
12. Kevin Fall and Kannan Varadhan: The ns Manual. http://www.isi.edu/nsnam/ns/doc.pdf (2005)

Message Complexity Analysis of MANET Address Autoconfiguration Algorithms in Group Merging Case

Sang-Chul Kim

School of Computer Science, Kookmin University,
861-1, Chongnung-dong, Songbuk-gu, Seoul, 136-702 Korea
sckim7@kookmin.ac.kr

Abstract. This paper focuses on the derivation of the message complexity when two mobile ad hoc network (MANET) groups merge together, where the network groups already have been configured with IP addresses by using address autoconfiguration protocols (AAPs). The message complexity of the MANET group merging case (GMC) in Strong DAD, Weak DAD with proactive routing protocols (WDP), Weak DAD with on-demand routing protocols (WDO), and MANETconf has been derived respectively. In order to verify the derived bounds, analytical simulations that quantify the message complexity of the address autoconfiguration process based on the different conflict probabilities are conducted.

Keywords: Mobile Ad hoc Networks, Group Merge, Address Autoconfiguration, Message Complexity.

1 Introduction

Clustering (or grouping) of mobile nodes provides effective and efficient means to control routing and addressing in MANETs. In MANET, as the network grows (including more nodes) if hierarchical routing schemes are applied the control overhead is known to increase in a scalable fashion compared to using flat routing techniques. Due to this reason, several hierarchical routing protocols have been developed to enable scalable MANET routing solutions [1]. Due to the lack of any centralized control and possible node mobility in MANETs, many issues at the network, medium access, and physical layers currently remain as research topics since no counterparts in the wired networks or cellular networks can satisfy MANET requirements. One of the main criteria in determining efficiency in MANETs is the scalability of the control signaling. This issue becomes even more serious when MANET groups merge, as the addressing and new route establishment is required for multiple nodes simultaneously [2,3,4,5]. In mobile IPv6 networks, a mobile node can select its own IP address (using the subnet's prefix) but needs to obtain confirmation from the subnetwork before being permitted to use the chosen address. The confirmation process is based on a duplicated address detection (DAD) operation. The DAD operation is one of the

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 320–327, 2007.
© Springer-Verlag Berlin Heidelberg 2007

most important processes of address autoconfiguration. Currently, Weak DAD, Strong DAD, and MANETconf have been proposed as candidate algorithms for DAD address autoconfiguration [6, 7]. The broadcast storm problem introduced in [8] is a serious problem in MANET operations, and hence several algorithms are introduced in [9] to reduce the number of broadcast messages. The authors of [9] conclude that finding a minimum flood tree which gives the minimum number of forward nodes is proven to be a NP-complete problem, where the minimum flood tree is derived in [10]. In order to provide scalability, adaptability, and autonomy, Shen proposes the Cluster-based Topology Control (CLTC) algorithm [11] that uses a clustering strategy as well as a topology control algorithm with help of changing the transmission power. Shen uses the message complexity to statistically measure the performance of the CLTC protocol. The authors in [1] calculate the storage complexity and communication complexity to analyze the scalability of various MANET routing protocols and introduce the routing overhead of periodically updated Link State (LS) messages, which is known to follow the order of $O(N^2)$, where N indicates the number of nodes in a MANET. However, a message complexity analysis and comparison among the IP address autoconfiguration protocols for MANET GMC has not been conducted yet. Therefore, in this paper, the upper bounds of the message complexity of the IP address autoconfiguration protocols for MANETs are derived for the GMC. In order to verify the derived bounds, analytical simulations that quantify the message complexity of the address autoconfiguration process based on the different conflict probabilities are conducted. In addition, the acronyms of messages and nomenclatures of the retry count variables used in this paper are summarized in Table 1.

Table 1. Acronym table [*: variable]

Acronym	*Message*	*Acronym*	*Message*	*Acronym*	*Message*
AB	Abort	*AP*	Address Reply	*NQ*	Neighbor Query
AC	Address Cleanup	*AQ*	Address Request	*RR*	Route Reply
AD	Advertised	*IR*	Initiator Reply	*RQ*	Route Request
AE	Address Error	*IQ*	Initiator Request	*RT*	Requester Request
AL	Allocated	*LS*	Link State	m	DAD retry count limit*
AO	Allocation	*NR*	Neighbor Reply	n	retry count limit*

2 Message Complexity Analysis

A MANET can be represented as an undirected graph $G(\boldsymbol{V}, \boldsymbol{E})$ where $\boldsymbol{V}$ is a finite nonempty set of nodes, which can be represented as $\boldsymbol{V} = \{V_1^G, V_2^G, \cdots, V_W^G\}$ where $|\boldsymbol{V}| = W$ and $\boldsymbol{E}$ is a collection of pairs of distinct nodes from $\boldsymbol{V}$ that form a link, which can be represented as $\boldsymbol{E} = \{E_1^G, E_2^G, \cdots, E_W^G\}$ [12]. A connected, acyclic, undirected graph which contains all nodes is defined as a free tree. $\boldsymbol{V}$ can be partitioned into several subgraphs $\boldsymbol{V}_1, \boldsymbol{V}_2, \cdots, \boldsymbol{V}_k, \cdots, \boldsymbol{V}_n$ where each

partition subgraph is called as a free tree and $|\boldsymbol{V}_1| + |\boldsymbol{V}_2| + \cdots + |\boldsymbol{V}_n| = W$. A partitioned subgraph $\boldsymbol{V}_k$ is represented as a free tree $P(V, E)$, in which a node set V is represented as $V_1, V_2, \cdots, V_N$ and $|V|$ equals N containing all nodes in the partitioned subgraph $\boldsymbol{V}_k$, where $N \leq W$.

In this paper, the most common flooding method is used to broadcast an AQ message where every node retransmits an AQ message to its entire one-hop neighbors whenever it receives the first copy of the AQ message. Since each member node in a free tree will relay the AQ message initiated at node $\boldsymbol{V}_i$, the maximum number of nodes relaying an AQ message is N-1, where the rule of discarding duplicated messages at a node is adopted. Therefore, the maximum number of AQ messages broadcasted or relayed in the free tree is N, which can be represented as $O(N)$. The variable t is defined as the largest number of nodes in a communication path based on the routing tree, including the source node.

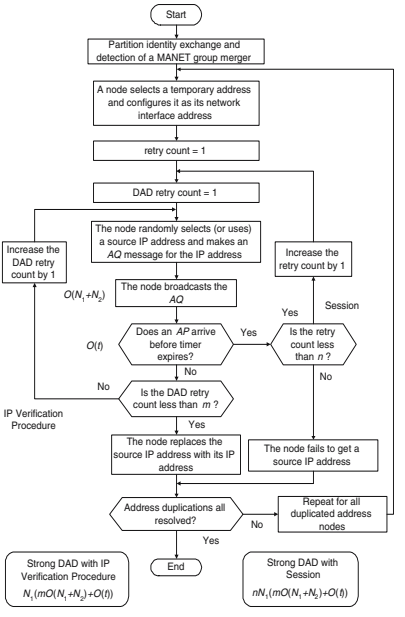
Definition 1. *For a MANET routing tree with t nodes in the maximum length path, $O(t)$ is the upper bound of the maximum number of unicasted or relayed AP messages when a node unicasts an AP message.*

In order to analyze the GMC, a scenario is considered where two MANET groups $\boldsymbol{V}_i$ and $\boldsymbol{V}_j$, where $|\boldsymbol{V}_i| = N_1$, $|\boldsymbol{V}_j| = N_2$, and $N_1 \leq N_2$, merge into each other and a node in the $\boldsymbol{V}_i$ finds an IP address that is duplicated in the $\boldsymbol{V}_j$ based on a routing message such as *Hello* received (Strong DAD), or based on the LS, RQ, or RR message received (Weak DAD). It is assumed that in order to include the fields of *Partition Identity*(PI) and the number of nodes (N_1 or N_2), the routing message is modified in each group for Strong DAD, or LS, RQ, or RR message is modified for Weak DAD. All nodes of a group know their group's PI (Lowest IP, UUID), where the *Universal Unique ID* (UUID) is the MAC address of the lowest IP address node. When the two nodes I and J associated in two different groups become neighboring nodes to each other, two nodes I and J detect the merger of two different MANET groups with help of the routing message or LS, RQ, or RR message.

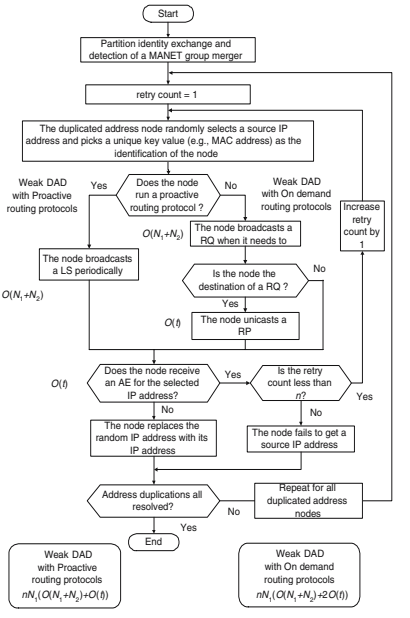
In order to analyze the GMC with Strong DAD, since the message complexity of the $\boldsymbol{V}_j$ is defined as $n(mO(N_1 + N_2) + O(t))$ as shown in Fig. 1(a), in the worst case, N_1 nodes in the $\boldsymbol{V}_i$ need to verify their IP addresses in a merged MANET. The message complexity of the GMC can be represented as $nN_1(mO(N_1 + N_2) + O(t))$, where each node in the $\boldsymbol{V}_i$ generates the message complexity of $n(mO(N_1 + N_2) + O(t))$, which concludes the following corollary.

Corollary 1. *$nN_1(mO(N_1 + N_2) + O(t))$ is the upper bound of the maximum number of broadcasted/relayed AQ messages and unicasted/relayed AP messages of the GMC with Strong DAD.*

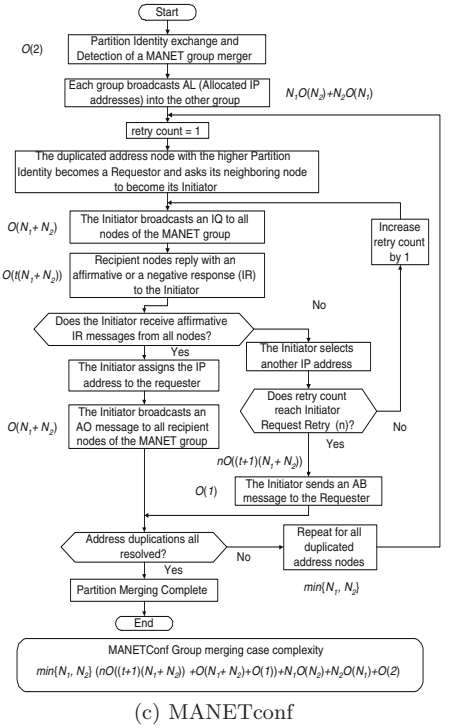
Since the message complexity in the $\boldsymbol{V}_i$ and $\boldsymbol{V}_j$ caused by a node in the $\boldsymbol{V}_i$ has been defined as $n(O(N_1+N_2)+O(t))$ and $n(O(N_1+N_2)+2O(t))$ as shown in Fig. 1(b), in the worst case, N_1 nodes in the $\boldsymbol{V}_i$ need to verify their IP addresses in a merged MANET. Therefore, the message complexity of the GMC can be represented as $N_1(n(O(N_1 + N_2) + O(t)))$ in WDP and $N_1(n(O(N_1 + N_2) + 2O(t)))$ in



(a) Strong DAD



(b) WDP and WDO



(c) MANETconf

Fig. 1. The flowcharts of Strong DAD, Weak DAD, and MANETconf operations

WDO where each node in the $\boldsymbol{V}_i$ generates the message complexity of $n(O(N_1 + N_2) + O(t))$ in WDP and $n(O(N_1 + N_2) + 2O(t))$ in WDO. Based on the results, the following corollary can be derived.

Corollary 2. *$nN_1(O(N_1 + N_2) + O(t))$ (WDP) and $nN_1(O(N_1 + N_2) + 2O(t))$ (WDO) are the upper bounds of the maximum number of broadcasted/relayed LS messages and unicasted/relayed AE messages of the GMC using WDP.*

In the GMC of MANETconf, based on the flowchart as shown in Fig. 1(c), when the two nodes I and J associated in two different groups, which are $\boldsymbol{V}_i$ and $\boldsymbol{V}_j$ respectively, become neighboring nodes to each other, they exchange their *Partition Identities*. The nodes I and J can detect the merger of two different groups when the two nodes (I and J of each group) exchange their AL sets of IP addresses, which must contain the group's PI. Since the AL message is composed of a list of IP addresses of a group, the size of AL message could be larger than the maximum transfer unit (MTU) permitted in the MANET. It is assumed that in the worst case, the required MTU is small where each AL message packet contains only a single IP address (and overhead of the upper layers). Therefore, in the upper bound case, an AL message from $\boldsymbol{V}_j$ is segmented into N_2 number of MTU sized messages and transmitted in $\boldsymbol{V}_i$. In addition, the AL message from $\boldsymbol{V}_j$ is segmented into N_1 number of MTU sized messages and transmitted in $\boldsymbol{V}_j$. The algorithm requires all nodes in $\boldsymbol{V}_i$ to broadcast the AL messages transferred from $\boldsymbol{V}_j$, and all nodes in $\boldsymbol{V}_i$ have to broadcast N_2 number of AL messages. As a result, the message complexity can be represented as $N_2O(N_1)$. Likewise, all nodes in $\boldsymbol{V}_j$ need to broadcast the AL messages transferred from $\boldsymbol{V}_i$, and therefore, all nodes in the $\boldsymbol{V}_j$ have to broadcast N_1 number of AL messages. As a result, the message complexity can be represented as $N_1O(N_2)$. Therefore, the message complexity due to broadcasting the AL messages in $\boldsymbol{V}_i$ and $\boldsymbol{V}_j$ can be represented as $N_1O(N_2) + N_2O(N_1)$. The duplicated address node with the higher PI will become the Requestor asking its neighboring node to become its Initiator. Among the duplicated addresses nodes, the node of the group that has the higher PI (i.e., comparing the lowest IP address of each group first, and if needed, also by comparing the UUID of each group) will become the Requestor which chooses one of its neighbors with a non-conflicting address as its Initiator to send an IQ message. Any nodes detecting conflicted IP addresses become Initiators, where each Initiator broadcasts an IQ message to all nodes of the group with the address of the Requester. The message complexity upper bound of broadcasting IQ messages can be represented as $O(N_1 + N_2)$ since the IQ message is broadcasted into the merged MANET. Recipient nodes will reply with an affirmative or a negative response (using IR message) to the Initiator. Therefore, the message complexity upper bound of unicasting IR messages can be represented as $O(t(N_1 + N_2))$ since all nodes $(N_1 + N_2)$ unicast the IR message and each IR message has the message complexity upper bound of $O(t)$ based on Definition 1. If the initiator receives positive IR messages from all recipient nodes, it broadcasts an AO message to all recipient nodes of the group. The message complexity upper bound of broadcasting the AO message can be represented as $O(N_1 + N_2)$. Therefore, the following corollary can be derived.

Corollary 3. *In an IP verification procedure of the GMC,* $O((t+1)(N_1+N_2))$ *is the upper bound of the maximum number of broadcasted or relayed IQ messages and unicasted or relayed IR messages when a node needs to verify its IP address in a MANET.*

If the initiator receives any negative IR messages from its recipient nodes, it selects another IP address and repeats the steps of broadcasting IQ and receiving IR messages until the retry count reaches the retry count limit (n). Therefore, the message complexity of broadcasting AO messages and receiving IR messages until the retry count is less than n can be represented as $n(O(N_1+N_2)+O(t(N_1+N_2)))$. After n times of repetition, if the initiator receives negative IR messages, it sends an AB message to the requestor. The message complexity of unicasting the AB message can be represented as $O(1)$. Based on the above results, the following corollary can be derived.

Corollary 4. *In a session of a GMC,* $nO((t+1)(N_1+N_2))+O(N_1+N_2)+O(1)$ *is the upper bound of the maximum number of broadcasted or relayed IQ and AO messages and unicasted or relayed IR and AB messages in MANETconf.*

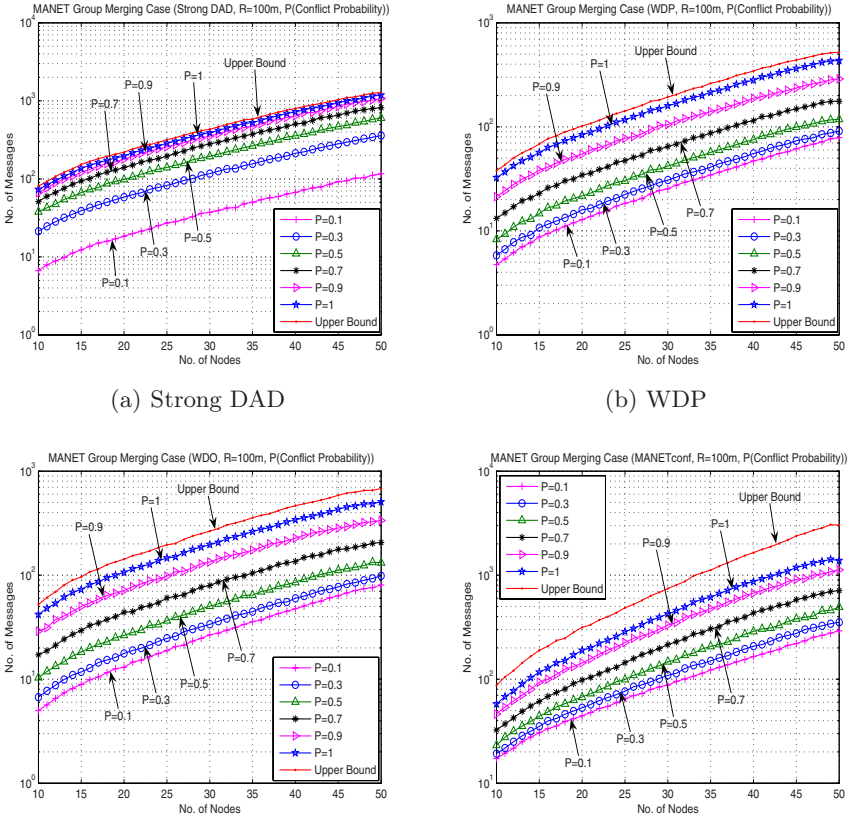
The above session procedure per a duplicated address node should be repeated until all duplicated address nodes are resolved. The repetition number of the session procedure in a merged MANET is $min(N_1, N_2)$. Based on the above results, the following corollary can be derived.

Corollary 5. *In resolving all duplicated addresses of a GMC with MANETconf,* $min(N_1, N_2)\ \{nO((t+1)(N_1+N_2))+O(N_1+N_2)+O(1)\}+O(2)+N_1O(N_2)+N_2O(N_1)$ *is the upper bound of the maximum number of messages.*

3 Numerical Results

In order to analyze the message complexity of each AAP, a computer simulator was developed where nodes are randomly distributed with uniform density in a network area of $1km^2$. A discrete-event simulator was developed in $Matlab$ in order to verify the various network topologies and to calculate the message complexity of each AAP. The random node generator and simulator performance was verified (for the numbers of nodes 100, 125, 150, and 175) so that the average number of nodes per cluster as well as several specs in the *adaptive dynamic backbone* (ADB) algorithm [11] matched with the results in [11], which was performed by $QualNet$, with less than a 1% difference for almost all cases. In our analysis, the conflict probability is defined as the probability in which the IP address that a node requests to use is already in use in the group. $Dijkstra$'s shortest path algorithm at each node is used to calculate the number of hops in unicasting or relaying an unicasted AP message from a destination node to a source node. In the Strong DAD, five is used for retry count limit (n) and three is used for DAD retry count limit (m). In the Weak DAD and MANETconf, five is used for retry count limit (n) and one is used for DAD retry count limit (m).

In addition, $100m$ is selected as the transmission range of nodes. The number of nodes is varied from 10 to 50. In Fig. 2, it can be observed that at the conflict probability of 0.5 and 0.7, WDP has the smallest message complexity and Strong DAD has the largest message complexity. In the range of 10 to 35 nodes at the conflict probability of 0.9, and also for the case of 10 and 25 nodes at the conflict probability of 1, WDP has the smallest message complexity and Strong DAD has the largest message complexity in the GMC. In addition, it can be calculated that with the increase of the conflict probability from 0.5 to 1, the maximum overhead percentage of the message complexity of WDO increases gradually from 28.41% to 35.96% and then decreases to 33.52 %, the maximum overhead percentage of the message complexity of MANETconf is decreased from 318.10% to 232.48% gradually, and the maximum overhead percentage of the message complexity of Strong DAD decreases rapidly from 408.38% to 172.29 %.



(a) Strong DAD (b) WDP

(c) WDO (d) MANETconf

Fig. 2. Message complexities of Strong DAD, WDP, WDO, and MANETconf

4 Conclusion

The main objective of this paper is to propose a novel method to perform a quantitative analysis of message complexity and to compare the message complexity among the MANET AAPs in the GMC. To conduct a quantitative analysis of message complexity, the analysis of the worst case scenario is conducted in this paper. By introducing the retry count limit (n) of a session in Strong DAD, the possibility of resulting in an infinite loop has been removed. By adapting the mechanism of the replying AE message, the Weak DAD is equipped to properly react when solving duplicated IP address situations. Based on the simulation results, when nominal n, m, t, N values and transmission range have been assigned with $p = 0.5$ and 0.7, the message complexity can be compared as follows: WDP $<$ WDO $<$ MANETconf $<$ Strong DAD. However, for the case where the conflict probability is 0.9 or 1, the message complexity of the MANET GMC can be compared as follows: WDP $<$ WDO $<$ Strong DAD $<$ MANETconf. The results of this paper provide a direct comparison of scalability of DAD schemes based on MANET group merging cases. The methodology applied in this paper can be used to analyze newly developed DAD schemes in the future, which is one of the objectives that lead to conducting this research.

References

1. X. Hong, K. Xu, M. Gerla: Scalable Routing Protocol for Mobile Ad Hoc Networks. IEEE Network, pp.11-21 (2002)
2. C.-C. Chiang and M. Gerla: Routing in Clustered Multihop Mobile Wireless Networks. Proc. Information Networking (ICOIN11), 3B-1.1-3B-1.9 (1997)
3. G. Pei, M. Gerla, X. Hong, C. C, Chiang: A Wireless Hierarchical Routing Protocol with Group Mobility. Proc. IEEE WCNC '99, New Orleans, LA (1999)
4. Z. J. Haas, M R. Pearlman, Prince Samar: The Zone Routing Protocol (ZRP) for Ad Hoc Networks. Internet Draft (2002) http://www.ietf.org/proceedings/02nov/I-D/draft-ietf-manet-zone-zrp-04.txt
5. M. Gerla, X. Hong, L. Ma, G. Pei: Landmark Routing Protocol (LANMAR) for Large Scale Ad Hoc Networks. Internet Draft (2002) http://www.ietf.org/proceedings/01dec/I-D/draft-ietf-manet-lanmar-02.txt
6. N. H. Vaidya: Weak duplicate address detection in mobile ad hoc networks. Proc. ACM MobiHoc, Lausanne Switzerland(2002) 206–216
7. S. Nesargi, R. Prakash: MANETconf: Configuration of hosts in a mobile ad hoc network. Proc. IEEE Infocom 2002, New York (2002)
8. S. Ni, Y. Tseng, Y. Chen, J. Sheu: The Broadcast Strom Problem in a Mobile Ad Hoc Network. Proc. ACM MobiCom (1999)
9. W. Lou, J. Wu: On reducing broadcast redundancy in Ad hoc wireless networks. IEEE Trans. on Mobile Computing, Vol. 1, No. 2 (2002) 111–122
10. H. Lim, C. Kim: Flooding in wireless ad hoc networks. Computer Comm. J., vol. 24, no.3-4 (2001) 353–363
11. C-. C. Shen, C. Srisathapornphat, R. L. Z. Huang, C. Jaikaeo, E. L. Lloyd: CLTC: A cluseter-based topology control framework for ad hoc networks. IEEE Trans. Mobile Computing, vol. 3, no.1 (2004) 18–32
12. J. Gross, J. Yellen. Graph Theory and Its Applications. CRC Press (1998)

A Robust Route Maintenance Scheme for Wireless Ad-Hoc Networks

Kwan-Woong Kim[1], Mike Myung-Ok Lee[2], ChangKug Kim[3], and Yong-Kab Kim[1]

[1] Div. of Electrical Electronic & Information Eng.,
Wonkwang Univ., Iksan, 570-749, South Korea
{watchbear,ykim}@wonkwang.ac.kr
[2] Murdoch University, South Street, Murdoch, Western Australia 6150, Australia
Mike.Lee@murdoch.edu.au
[3] Bioinformatics Div, National Institute of Agricultural Biotechnology, R.D.A.
225 Seodundong, Suwon, 441-707, Korea

Abstract. Ad hoc networks are dynamic networks that consist of mobile nodes. Nodes in Ad hoc networks are usually laptops, PDAs or mobile phones. These devices feature Bluetooth and/or IEEE 802.11 (WiFi) network interfaces and communicate in a decentralized manner. Due to characteristics of Ad-hoc networks, Mobility is a key feature of routing protocol design. In this study, we present an enhanced routing maintenance scheme that cope with topology changes pre-actively. The key feature of the proposed scheme is that switches next-hop node to alternative neighbor node before link breakage for preventing route failure. From extensive experiments by using NS2, the performance of the proposed scheme has been improved by comparison to AODV protocol.

Keywords: Wireless Ad-hoc Networks, Routing Protocols, Mobility, AODV.

1 Introduction

Wireless Ad-hoc network [1, 2] is a self-organized, dynamically changing multi-hop network. All mobile nodes in an ad-hoc network are capable of communicating with each other without the aid of any established infrastructure or centralized controller. Ad-hoc network is useful in many applications because they do not need any infrastructure support and has capability of self configuration. Sensor networks, disaster recovery, rescue and automated battlefields are examples of application environments.

The nodes have the responsibility of self-organizing so that the network is robust to the variations in network topology due to node mobility as well as the fluctuations of the signal quality in the wireless environment.

Compared to traditional routing protocols in wired networks, that of Ad Hoc networks required to cope with the high rate of topology changes.

This implies that the routing protocol should propagate topology changes and compute updated routes to the destination. Since wireless ad-hoc networks usually have limited bandwidth and battery power, their routing protocols should have low control overhead. Reactive or on-demand routing protocols have been developed for this reason. In an on-demand routing protocol, a node only maintains routes for in-use

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 328–335, 2007.
© Springer-Verlag Berlin Heidelberg 2007

destinations and does not pro-actively advertise routes. Rather, it queries for needed routes and offers routes in response to queries. Dynamic Source Routing (DSR), Ad-hoc On-demand Distance Vector Routing (AODV)[3], Lightweight Mobile Routing (LMR), Temporally Ordered Routing Algorithm (TORA), Associativity-Based Routing (ABR)[4], Signal Stability Routing (SSR) [5] are classified to on-demand scheme.

Table-driven protocols attempt to continuously update routes within the network so that when a packet needs to be forwarded, the route is already known and can immediately be used. The family of distance-vector or link-state algorithms is examples of table-driven schemes. There is Destination-Sequence Distance Vector (DSDV), Wireless Routing Protocol (WRP), Clusterhead Gateway Switch Routing (CGSR).[5]

Nodes in MANET may move freely and unpredictably, the path that packets traverse to its destination will be broken by link failure frequently. Link breakage caused by node mobility may degrade overall performance.

Tom Goff proposed preemptive-routing protocol that is an enhanced version of AODV and DSR [6]. Preemptive routing protocol measures receiving signal power to make decision whether launch re-route discovery before link breakage occurred by node mobility. Path recovery of this is similar with hand-off in cellular networks. Some on-demand protocols with multi-paths or backup routes have been proposed to improve the performance in ad-hoc networks. AODV-BR scheme improves AODV routing protocols by constructing a mesh structure and providing multiple alternate routes. The algorithm establishes the mesh and multi-path using the RREP of AODV, which does not transmit many control messages.[7][8]

In this work, we proposed a novel route maintenance scheme based on AODV that takes node mobility into consideration.

This paper structured as follows. In section 2, we present about related works and background. The proposed routing protocol based on AODV is presented in section 3. In section 3, Simulation results obtained by the proposed scheme are evaluated. Finally, section 4 presents conclusions and discussion.

2 Proposed Route Maintenance Scheme

Route maintenance in routing protocols plays a role of maintaining route connectivity and link breakage detection. In AODV, local route repair algorithm can be used for fast route recovery [3][9]. But most of existing routing algorithms have lack of ability route recover before link breakage occurred.

Our work is focus on routing maintenance to prevent route-failure which caused by node mobility and improve efficiency of routing protocol during forwarding packets. Our approach is quite different with other mobility support routing protocols.

To achieve this goal, Routing protocol should have ability of local route change from moving node to alternative node. The basic idea of the proposed scheme is based on our previous work [10].

The distance between two nodes is inverse proportional to the receiving signal strength *RxP* at the receiver in wireless networks [11]. If *RxP* is being lower, transmitting node moves far, else if *RxP* is increasing, then it moves near.

In previous studies, the method presented that estimates relative speeds of two nodes by measuring *RxP* variation. Prior to estimate relative speeds, the distance between two nodes must be known. But it is not easy to extract distance from *RxP*. Since *RxP* is composed of several factors such as transmitting power, antenna gain and channel loss factor [11], node might not have enough information.

In the proposed scheme, we use receiving signal variation function *V* for detecting node movements as follows.

$$V = RxP(t_1) - RxP(t_0) \tag{1}$$

If function *V* is negative value, it indicates two adjacent nodes becoming far, else if *V* is equal to zero, two nodes doesn't move or move same direction with the same speed. When value *V* is positive, two adjacent nodes move closer.

Process of the Proposed Scheme

In the proposed scheme, transmission range is divided into two zones, GREEN ZONE and RED ZONE as shown in figure 2. The next hop node locates in GREEN ZONE; it supposes to be safe state for data communication.

If node located in RED ZONE, the proposed scheme can be triggered for route re-construction.

Figure 1 illustrates the first phase of the proposed scheme. When intermediate node ('C') receives data packet from previous hop node 'A', it monitors receiving signal power *RxP* and signal strength variation *V*.

If *RxP* is under *RxTh* and *V(A)* is over than 0(movement detected), then node triggers local route change process to find alternative node among its neighbors. Looking for alternative nodes, node broadcasts HELP message to its one-hop neighbors. Where *RxTh* is receiving power threshold for RED ZONE and defined as follows.

$$RxTh = K \times RxP_{\min} \tag{2}$$

RxP_{min} is the minimum receivable power by network interface device (e.g. $3.65 \times 10^{-10} Watts$ in 802.11b [12]) and *K* is constant and set to 5.

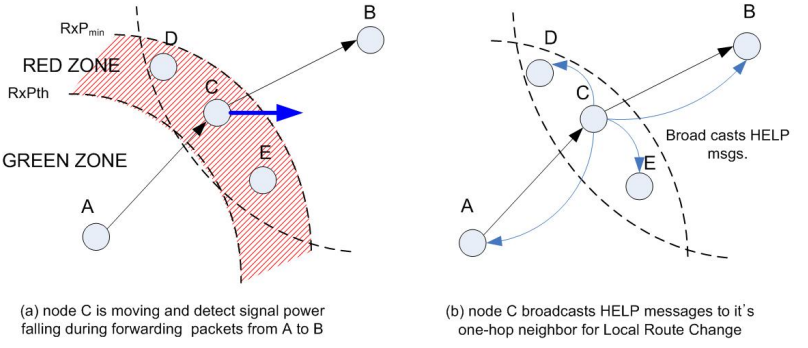


Fig. 1. The example of the proposed scheme: Broadcasts HELP message to its one-hop neighbor

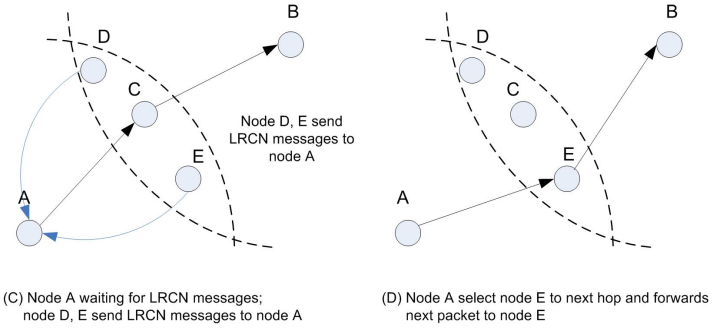


Fig. 2. The example of the proposed scheme: Previous-hop node receives LRCN message and switch next-hop node 'C' to new one

Type	Reserved	Hop Count
Prev-hop IP Address (32bits)		
Next-hop IP Address (32bits)		
Destination IP Address (32bits)		
Destination Sequence Number (32bits)		
V (32bits)		

(a) HELP message

Type	Reserved	Hop Count
Alternative node IP Address (32bits)		
Destination IP Address (32bits)		
V (32bits)		

(b) LRCN message

Fig. 3. New message formats for the proposed scheme

Subsequently, a node receives HELP packets, under condition that previous node 'A' and next hop node 'B' are belong to its neighbor. If signal variation *V(A)* is zero, node itself can be alternative of node C. Otherwise HELP message is ignored. Alternative node updates route information with HELP messages and sends LRCN (Local Route Change Notification) to previous node of HELP message. The previous hop node 'A' may receive HELP from its next hop node 'C'. The node 'A' set timer for waiting LRCN (Local Route Change Notification) messages. When LRCN message are received, previous-hop node cancel timer and updates next-hop address 'C' to source address of LRCN 'E'. When the timer expired, the node initiates local route repair [3] process to re-establish path to destination.

New message formats are shown in figure 3. V field is 32 bit floating point value.

Pseudo code of the proposed scheme procedures is shown bellows. To avoid unnecessary broadcasting HELP, node set the flag of precursor of route entry to 1. It indicates HELP sent already for the flow. Precursor list is a set of nodes that share the same route to reach final destination [3]. The flag of precursor initialized when route is updated.

Procedure of receiving DATA packet from node i

```
Compute V(i) by equation 1;
Forwarding DATA packet to the next hop;
If (RxP is less than RxPth and V < 0) {
        broadcasts HELP packet with previous node and
next-hop node;
End if
```

Procedure of receiving HELP msg from node j

```
If(prev-hop node is not me and next-hop node is not me)
    If (V(j) == 0 and both next-hop field and previous
     node field in HELP msg are my neighbor){
        Updates routing table with information of HELP.
       Send LRCN msg to prev-hop node of HELP msg.
     }
}
Else if (prev-hop field in HELP msg is my address){
     Launch LRCN timer for waiting LRCN msg.
}
```

Procedure of receiving LRCN msg from node k

```
If (destination address of LRCN is exist in routing
    table){
     Updates next-hop address of route entry with
     alternative address of LRCN message.
     Cancels LRCN timer
}
```

Procedure of LRCN timer expired

```
Launch local route repair
```

3 Performance Evaluation

In this section, a performance of the proposed routing protocol is evaluated using extensive simulations and compared its performance with AODV. NS2 simulator was used for experiments and the proposed scheme is implemented as part of AODV in NS2 [13]. The network model used for simulations consists of 100 mobile nodes in 1.0 km × 1.0 km area. The initial position of nodes is randomly chosen. Node pairs are randomly selected to generate CBR/UDP traffic. Channel bandwidth is 2 Mbps.

Each node uses IEEE 802.11 MAC protocol and the used Channel model is Wireless channel/Wireless Physical propagation model. Two-Ray Ground model is used for radio propagation model and transmission range and interference range of a mobile node is 250m and 550m. Traffic source are CBR (Constant Bit Rate) and 15 CBR sources generate UDP packet in every 0.1 sec. The size of UDP packet is 512 bytes. The simulation time is set to 200 seconds. The mean pause time of nodes is 10 seconds. Maximum speed of nodes varies from 5m/sec to 20 m/sec.

To avoid the bias of random number generation, we performed simulation 10 times under the same configuration. Table 1 shows parameter of energy model in NS2.

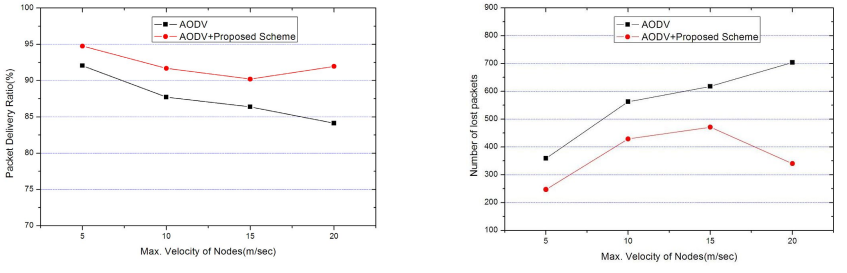
Table 1. Parameter of Energy Model in NS2

Attribute	Description	Value
-initialEnergy	Given energy for each node	**200 Joules**
-Grx, Gtx	Antenna Gain	**1**
-txPower	**Transmitting power in Watt**	**281.8mW**

Figure 4 plots end-to-end packet delivery ratio and number of lost packets respectively. As the maximum speed of nodes increase, more packets are dropped in the network by broken paths. In case of AODV, packet delivery ratio falls down significantly in high mobility situation. But AODV with proposed scheme keeps certain levels above 90%.

In all cases, the proposed scheme improves the number of received packets and reduces packet loss. The main reason of performance improvement is that the proposed scheme can change route to alternative node before the next-hop node move out of transmission range. It could reduce packet loss and route failure more efficiently in high mobility environments.

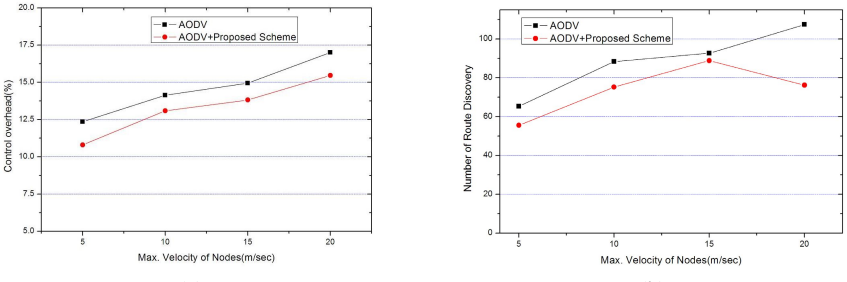
From these results, AODV combined with the proposed scheme can give quite positive effects for overall performance and efficiency of route discovery.



(a) (b)

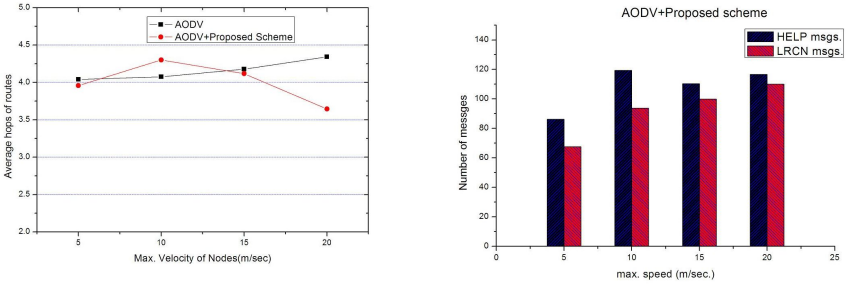
Fig. 4. (a) Packet delivery ratio. (b) Number of lost packets.

Performance comparison in control overhead and number of route discovery is shown in figure 5. In the most cases, AODV with proposed scheme reduces control message overhead and number of route discovery compared to AODV. It is obvious that local route change after route establishments can reduce probability of re-route discovery and control overhead efficiently.



(a) (b)

Fig. 5. (a) Comparison of control overhead. (b) Number of route discovery.



(a) (b)

Fig. 6. (a) Average hops of routes. (b) Number of new control messages.

Figure 6 (a) depicts average hops of routes and number of transmitted new messages shown in figure 6(b). In the proposed scheme, previous hop node launches local route repair process when it failed to receive LRCN messages. Therefore hop counts of route can be increase in some cases. As maximum speed of node increase, the more HELP and LRCN message generated as shown in figure 6 (b).

In general, the results are quite positive in the sense that the proposed scheme outperformed than AODV in terms of routing overhead and throughput. Using our technique, the proposed scheme may reduce re-route discovery as well as overall end-to-end throughput improvements over multi-hop ad-hoc networks.

4 Conclusion

Since the cost of detecting and re-establish broken path is high. The method for overcome mobility of node is one of main research issues in routing protocol design. In this paper, we presented a new route maintenance scheme for AODV using receiving signal variation. The main feature of the proposed scheme is capability of switching the next hop node to one of its available neighbour before the next hop node move out of transmission range. Additional messages are defined and the proposed scheme is implemented in network simulator NS2 for performance evaluation. From simulation results, the proposed scheme can reduce broken path, overhead of control messages, and improve end-to-end packet delivery ratio, which compared to AODV.

Acknowledgement. This paper was supported by Wonkwang University in 2006.

References

1. Perkins, C.E.: Ad Hoc Networking. Addison-Wesley, Upper Saddle River, NJ, USA, 1 (2001).
2. http://www.ietf.org/, IETF MANET Working Group.
3. Perkins, C.E., Royer, E.M., Das, S.R.: Ad-hoc on demand distance vector routing. IETF RFC3561, http://www.ietf.org/rfc/rfc 3561.txt, (2003)

4. Toh, C.K.: Associativity Based Routing for Ad Hoc Mobile Networks. Wireless Pers. Commun. J. Special Issue on Mobile Networking and Computing Systems, vol. 4, no. 2, 3 (1997)
5. Elizageth, M., Royer, Toh, C.K.: A Review of Current Routing Protocols for Ad Hoc Mobile Wireless Networks, IEEE personal Communications, 4 (1999)
6. Goff, T., Nael, B., et al.: Preemptive Routing in Ad-hoc Networks. ACM SICMOBILE. 6(2001) 43–52
7. Lee, S.J., Gerla, M.: AODV-BR: Backup routing in Ad Hoc networks. Proceedings of IEEE WCNC 2000. Chicago, IL, (2000).
8. Wang, Y.H., Chuang, C.C., Hsu, C.P., Chung, C.M.: Ad hoc on demand routing protocol setup with backup routes. Proceedings of ITRE 2003. International Conference on Information Technology Research and Education, 8 (2003) 137–141.
9. Kim, K.H., SEO, H.G.: The Effects of Local Repair Schemes in AODV-Based Ad Hoc Networks. IEICE TRANSACTIONS on Communications, Vol.E87-B No.9, 8(2004)2458-2466,
10. Brahma, M.K., Kim, W., Abouaissa A., Lorenz, P.: A Load-Balancing and Push-Out Scheme for Supporting QoS in MANETs. Telecommunication Systems Journal, Vol. 30, No.1-3, 10(2005)161-175
11. Anderson, J.B., Rappaport, T.S., Yoshida, S.: Propagation Measurements and Models for Wireless Communications Channels. IEEE Communication Magazine, 1(1995) 42-49
12. WaveLAN/PCMCIA Card User's Guide – Lucent Technologies
13. Network Simulator: NS2.29 available via website http://www.isi.edu/nsnam.ns/

Route Optimization with MAP-Based Enhancement in Mobile Networks

Jeonghoon Park, Tae-Jin Lee, and Hyunseung Choo*

School of Information and Communication Engineering
Sungkyunkwan University
440-746, Suwon, Korea +82-31-290-7145
{jhpark,tjlee,choo}@ece.skku.ac.kr

Abstract. The development of wireless network technology and user demands for mobility support have motivated the IETF to introduce mobile IP, mobile IPv6, and its extension, the network mobility (NEMO) basic support protocol. In the NEMO environment, mobile networks form a nested structure. Nested mobile networks based on NEMO basic support (NBS) protocol have the pinball routing problem because packets are routed to all home agents of the mobile routers using nested tunnelling. In this paper, a route optimization scheme is proposed which uses the mobility anchor point employed in hierarchical mobile IPv6, and modifies the binding update messages to minimize overhead in route optimization. We evaluate route optimization cost in terms of delay. The results demonstrate a minimum performance improvement of 30% and even shorter routing delay than NBS in non-optimized cases.

Keywords: Route optimization, routing protocol, mobile networks, pinball routing problem, and mobility anchor point.

1 Introduction

As a result of the development of Wireless network technology and demand for mobility support from users, the IETF has introduced mobile IP (MIP) [1], mobile IPv6 (MIPv6) [2], and its extension, the network mobility (NEMO) basic support protocol [3]. Route optimization, multi-homing, and security are studied actively in the field of mobile networks that use the NEMO basic support (NBS) protocol. In the NEMO environment, it is assumed that mobile networks can be nested. When the correspondent node (CN) sends a packet to the mobile node (MN), which is located at the nested mobile network, the packet has to visit the home agents (HAs) of all mobile routers (MRs). Further, the packet is tunneled by every HA, because NBS uses a bi-directional tunnel between the MR and HA. This describes the pinball routing problem.

When the packet is routed to all intermediate MR's HA, the routing path becomes too long and the packet size grows due to tunneling, resulting in network inefficiency. And the root-MR or MR's HA link result in a bottleneck for

* Corresponding author.

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 336–343, 2007.
© Springer-Verlag Berlin Heidelberg 2007

the aggregated traffic from/to all mobile network nodes. If the home network has failures, it cannot ensure the connectivity of mobile networks [4]. There have been many research proposals to solve this problem. The HMIP based route optimization method (HMIP-RO) [5] and reverse routing header (RRH) [6] are well-known schemes. HMIP-RO simply extends hierarchical mobile IPv6 (HMIPv6) to the NEMO environment to optimize routing. The RRH proposes an extension of the IPv6 routing header [7]. This scheme uses only one tunnel between the MR, which is the nesting MN and MR's HA.

In this paper, we propose a route optimization with MAP-based enhancement (ROME), to solve aforementioned problems in nested mobile networks. ROME uses a mobility anchor point (MAP), which was introduced by HMIPv6 [8], to optimize routing path. This scheme uses a MAP similar to HMIP-RO, but it reduces additional overhead when the MAP is applied to the NEMO environment, by modifying binding update (BU) messages. Hence we evaluate ROME with NBS, when mobile networks use a non-optimized route. When mobile networks optimize its route, the route optimization cost is compared with ROME, HMIP-RO, and RRH, in terms of delay. According to the results, ROME shows a minimum 30% performance improvement and shorter routing delay than NBS in non-optimized cases.

The subsequent sections of this paper are organized as follows. Section 2 describes a new mechanism for route optimization, named ROME. In Section 3, we evaluate the performance of the ROME scheme, through analytical modelling. This paper is concluded in Section 4.

2 The Proposed Scheme

In this section, we propose a new route optimization scheme named ROME. The CN and MN can find the optimal route using the MAP similar to HMIP-RO.

2.1 Motivation

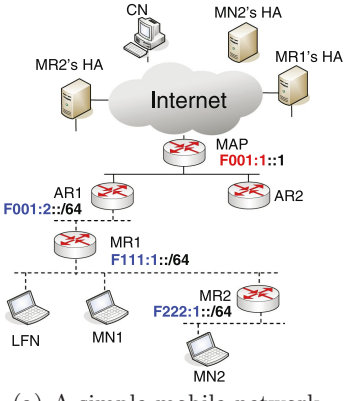
HMIP-RO simply applies the HMIPv6 to the NEMO environment and it has an advantage when mobile networks move within the subnet of the MAP. However, when mobile networks move out of the MAP, a BU storm occurs because all the MRs and MNs in moving mobile networks change their regional care-of address (RCoA) and send BU messages to their HAs. Further, when each MN transmit a packet to the CN, the packet is tunneled as in HMIPv6 whenever it passes through the intermediate MR. This actually induces an additional processing cost in the MRs and accordingly in the MAP.

ROME maintains the advantage of using MAP and resolves the problem of HMIP-RO. It modifies the BU message to avoid the BU storm. The MRs registers the RCoA, home address (HoA), and mobile network prefix (MNP) to their HAs, that is similar to HMIP-RO. Meanwhile, the MNs register the on-link care-of address (LCoA) and HoA to their HAs. When mobile networks move out of the subnet of MAP, only MRs register the RCoA to the HA. Hence, ROME reduces the number of BU messages as many as the number of MNs when mobile

networks move out of the subnet of MAP. To solve the nested tunnel problem, the MR that receives the router advertisement (RA) message with the MAP option adds the prefix of MAP to the permitted prefix filter range of the ingress filtering option. Accordingly, packets of RCoA source address can pass through the intermediate MRs and the MAP between the MN and the CN.

2.2 Rome

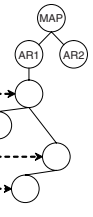
In ROME, we assume that the MAP employed in HMIPv6 is a router located in the route aggregation point and requires a certain level of processing ability. The MAP includes its address in the MAP option field of the RA message and propagates it. Similar to the HAs, the MAP manages the network mobility by registering the local BU messages that is received from the MRs and MNs to its binding cache entry.



(a) A simple mobile network.

MAP's Binding Cache Entry

	Home address	Care-of address	Mobile Network Prefix	P
MR1	F001:1::2	F001:2::2	F111:1::/64	→
MN1	F001:1::4	F111:1::4	-	→
MR2	F001:1::5	F111:1::5	F222:1::/64	→
MN2	F001:1::6	F222:1::6	-	→

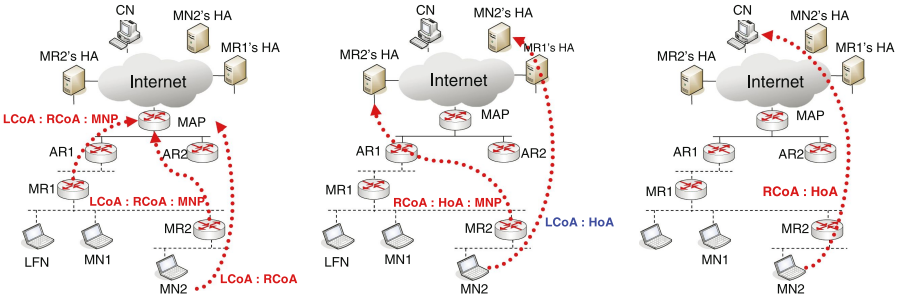


(b) MAP's binding cache entry.

Fig. 1. Basic networks configuration

Network Structure in MAP. The MAP recognizes the mobile network topology by combining BU messages. It manages this topology information using a tree structure, in order to search efficiently, as shown in figures 1 and 2. The BU messages from the MRs include the MNP of MR as represented in Fig. 2(a). So, The MAP defines these MRs as roots of subtrees as depicted in Fig. 1(b). The MRs and MNs based on this MNP are its children nodes. Based on the subtree, the entire network can be managed by the tree structure as shown in Fig. 1(b). The number of mobile networks using the RCoA with the prefix of the MAP is more than one and even they are overlapped. So, they cannot use duplicate address detection (DAD) [7] for unique address checking. In ROME, the MAP checks the address uniqueness when the MRs or MNs register the RCoA by local BU, and sends the result by binding acknowledgment (BA).

MR Consideration. The MR makes RCoA based on the prefix of the MAP and LCoA based on the prefix of the AR or the MNP of the parent MR. And, it



(a) BU for the MAP. (b) BU for the HA. (c) BU for the CN.

Fig. 2. BU messages

appends the prefix of the MAP to the permitted prefix filter range of the ingress filtering option in order for the packets with the RCoA as a source address pass the intermediate MRs and the MAP without tunneling between the MN and CN. After that, the MR registers the RCoA as HoA, and the LCoA as CoA, with the MNP to the MAP, using a local BU, as shown in Fig. 2(a). The MAP can recognize MNs inside the MR by MNP and parent MR by LCoA. Then, the MR registers the HoA and RCoA as CoA with the MNP to the HA using the BU, as shown in Fig. 2(b). The MR should not register again with the HA when it moves in the subnet of the MAP, because it registers the RCoA, which is the unaffected address in the subnet of the MAP. This reduces BU message cost.

MN Consideration. The MN makes the LCoA using the MNP contained in the RA of MR and the RCoA using the prefix of the MAP contained in the MAP option field. Then, the MN registers the RCoA as HoA and the LCoA as CoA using the local BU, similar to HMIPv6, as shown in Fig. 2(a). However, the MN registers, unlike HMIP-RO and HMIPv6, LCoA and HoA using BU similar to Mobile IPv6 as depicted in Fig. 2(b). Therefore, when mobile networks move out of the subnet of the MAP, the MN should not perform a BU to the HA. Therefore, a BU storm does not occur. The MN registers the RCoA and HoA to the CN using a BU as presented in Fig. 2(c). So, the MN and CN use the optimal route. If the MN moves in the subnet of the MAP, it should not perform a BU to the CN. This reduces movement registration cost. In addition, ROME guarantees the location privacy of the MNs, since the CNs do not know the LCoA of the MN.

3 Performance Evaluation

In this section, we evaluate ROME with existing schemes using two different mechanisms. In the first, we evaluate the non-optimized routing delay in ROME with NBS, because NBS doesn't use an optimal route. In the second, we evaluate route optimization cost in terms of delay between the MN and CN in ROME, with HMIP-RO and RRH. Table 1 shows the parameters and values used in performance evaluation.

Table 1. Parameters and Values

Parameter	Meaning	Value	Unit
i	Nesting level		
P	Propagation speed	2×10^8	m/sec
H_{AVG}	Average hop between nodes in the wired network	5	
D_{AVG}	Average distance of a hop in wired networks	10^4	m
D	Average distance of a hop in wireless networks	10	m
B_{WD}	Transmission speed in wired networks	10^8	bit/sec
B_{WL}	Transmission speed in wireless networks	5.4×10^7	bit/sec
S	Normal packet size	1500×8	bit
S_{TU}	Tunnel header size	320	bit
$S_{RH}[n]$	Routing header size with n addresses	$64 + 128n$	bit
T_S	Time to search the binding cache	10^{-4}	sec
T_{CH}	Time to change source or destination address	10^{-4}	sec
T_{RH}	Time to process routing header	5×10^{-4}	sec
T_{TU}	Time to process tunnel in entry or exit point	5×10^{-4}	sec

3.1 Non-optimized Route Delay

The nodes in mobile networks do not always use the optimal route. For example, the time for the return routability (RR) [2] and BU is considerable overhead when sending small amounts of data. Therefore, mobile networks use the basic protocol, i.e. NBS, and perform route optimization as occasion demands. In this subsection, we compare NBS and ROME when it uses a non-optimized route, in terms of routing delay.

In NBS, the packet from the CN must travel through the MN's HA and all of the MR's HA, and is encapsulated at several levels. However, in ROME, the packet can travel through only the MN's HA and the HA of the MR that belongs to the MN, and it is encapsulated at two levels. In general, the total delay of the packet consists of transmission delay, propagation delay, and processing delay. Represented as follows.

$$D_{Total} = D_{Trans} + D_{Prop} + D_{Proc}$$

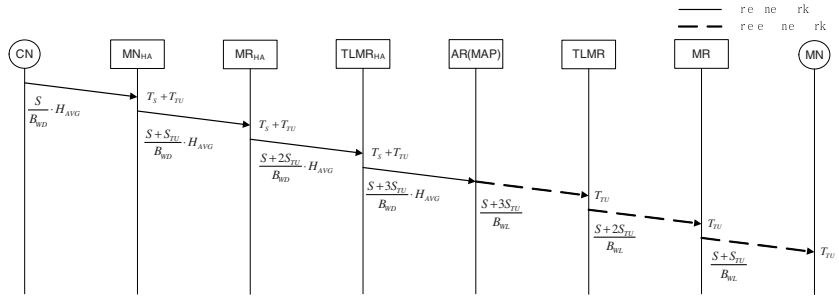
Fig. 3 represents the example of route delay in NBS and ROME. Transmission delay, propagation delay, and processing delay in NBS generalized by nesting level are derived as follows.

$$D_{NBS-Trans} = \frac{(i+2)S + \sum_{k=1}^{i+1} k \cdot S_{TU}}{B_{WD}} \cdot H_{AVG} + \frac{(i+1)S + \sum_{k=1}^{i+1} k \cdot S_{TU}}{B_{WL}},$$

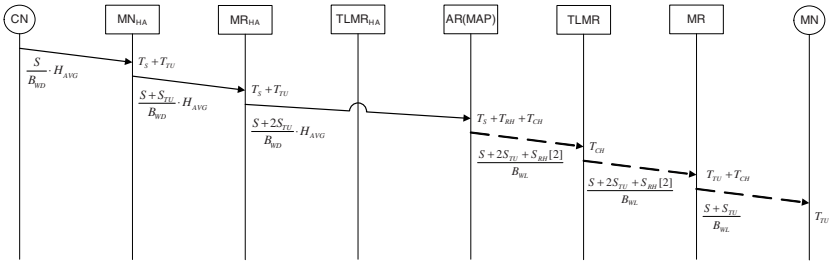
$$D_{NBS-Prop} = \frac{(i+2)D_{AVG} + (i+1)D}{P},$$

$$D_{NBS-Proc} = (i+1)T_S + 2(i+1)T_{TU}.$$

Transmission delay, propagation delay, and processing delay in ROME generalized by nesting level are derived as follows.



(a) Packet delay for the NBS.



(b) Packet delay for ROME.

Fig. 3. Packet delay for the NBS and ROME

$$D_{ROME-Trans} = \frac{3S+3S_{TU}}{B_{WD}} \cdot H_{AVG} + \frac{(i+1)S+(2i+1)S_{TU}+i \cdot S_{RH}[i]}{B_{WL}},$$

$$D_{ROME-Prop} = \frac{3D_{AVG}+(i+1)D}{P},$$

$$D_{ROME-Proc} = 3T_S+4T_{TU}+T_{RH}+(i+1)T_{CH}.$$

Fig. 4 represents the performance result of comparison between NBS and ROME. ROME shows lower performance than NBS, when the nesting level is 1, because the packet routing path is the same and the packet is encapsulated the same number of times, but ROME attaches the routing header for source routing in the MAP. ROME shows higher performance than NBS when the nesting level is 2 or greater. In particular, ROME shows greater than 60% performance improvement when the nesting level is 10.

3.2 Delay for Route Optimization

The route optimization schemes in the NEMO environment perform route optimization when the MN receives the encapsulated packet from its HA. In this subsection, we compare the route optimization cost of the HMIP-RO, RRH, and ROME in terms of delay.

The delay for route optimization in HMIP-RO is generalized by the nesting level derived as follows.

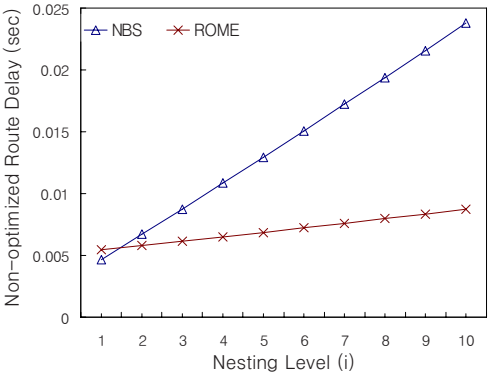


Fig. 4. Comparison of non-optimized route delay

$$ROD_{HMIP-Trans}[i] = \frac{4S+3S_{TU}}{B_{WD}} \cdot H_{AVG} + \frac{3(i+1)S+\left\{\sum_{k=1}^{i+1} k+(i+1)\right\} \cdot S_{TU}+2S_{RH}[i+1]}{B_{WL}},$$

$$ROD_{HMIP-Prop}[i] = \frac{4D_{AVG}+3(i+1)D}{P},$$

$$ROD_{HMIP-Proc}[i] = 3T_S+2(i+2)T_{TU}+2T_{RH}+2(i+1)T_{CH}.$$

Transmission delay, propagation delay, and processing delay for route optimization in RRH are generalized by nesting level, as follows.

$$ROD_{RRH-Trans}[i] = \frac{(i+6)S+2S_{RH}[i+1]+\left(\sum_{k=1}^{i+1} k+2\right)S_{TU}}{B_{WD}} \cdot H_{AVG} + \frac{3(i+1)S+\left(\sum_{k=1}^{i+1} k+2\right)S_{TU}+2i \cdot S_{RH}[i+1]}{B_{WL}},$$

$$ROD_{RRH-Prop}[i] = \frac{(i+6)D_{AVG}+3(i+1)D}{P},$$

$$ROD_{RRH-Proc}[i] = (i+2)T_S+2(i+3)T_{TU}+2T_{RH}+2(i-1)T_{CH}.$$

The transmission delay, propagation delay, and processing delay for route optimization in ROME generalized by nesting level are derived as follows.

$$ROD_{ROME-Trans}[i] = \frac{5S+3S_{TU}}{B_{WD}} \cdot H_{AVG} + \frac{3(i+1)S+(2i+1)S_{TU}+2(i+1)S_{RH}[i+1]}{B_{WL}},$$

$$ROD_{ROME-Prop}[i] = \frac{5D_{AVG}+3(i+1)D}{P},$$

$$ROD_{ROME-Proc}[i] = 4T_S+5T_{TU}+2T_{RH}+2(i+1)T_{CH}.$$

Fig. 5 shows the total route optimization delay for each scheme using Table 1. ROME shows at least 30% improvement over the other schemes.

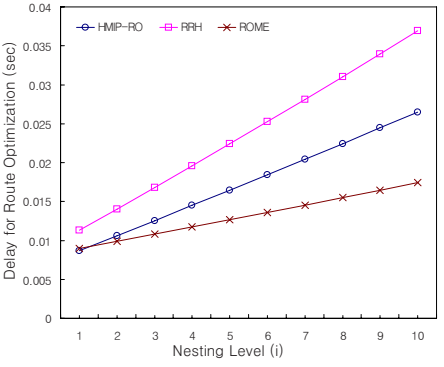


Fig. 5. Route optimization delay in various nesting level

4 Conclusion

In this paper, we propose a route optimization scheme based on the MAP, similar to HMIP-RO. ROME exhibits a shorter route length than the NBS, when mobile networks do not perform the route optimize procedure. The proposed scheme does not use tunneling between the MN and MAP, unlike HMIP-RO. Therefore, it reduces the overhead caused by tunneling. The BU storm is reduced by modifying the BU messages for the HA. The performance of the proposed scheme is better than existing schemes, HMIP-RO and RRH. In the future, we will study not only inter-NEMO route optimization, but also intra-NEMO route optimization, in mobile networks.

Acknowledgment. This research was supported by MIC, Korea under ITRC IITA-2006-(C1090-0603-0046)

References

1. Perkins, C.: IP Mobility Support for IPv4. RFC 3344, IETF (2002)
2. Johnson, D., Perkins, C., Arkko, J.: Mobility Support in IPv6. RFC 3775, IETF (2004)
3. Devarapalli, V., Wakikawa, R., Petrescu, A., Thubert, P.: Network Mobility (NEMO) Basic Support Protocol. RFC 3963, IETF (2005)
4. Ng, C., Thubert, P., Watari, M., Zhao, F.: Network Mobility Route Optimization Problem Statement. Internet draft, IETF (2006)
5. Ohnishi, H., Sakitani, K., Takagi, Y.: HMIP based Route Optimization Method in A Mobile Network. Internet draft, IETF (2003)
6. Thubert, P., Molteni, M.: IPv6 Reverse Routing Header and Its Application to Mobile Networks. Internet draft, IETF (2004)
7. Deering, S., Hinden, R.: Internet Protocol, Version 6 (IPv6). RFC 2460, IETF (1998)
8. Soliman, H., Castelluccia, C., El-Malki, K., Bellier, L.: Hierarchical Mobile IPv6 Mobility Management (HMIPv6). RFC 4140, IETF (2005)
9. IETF Network Mobility (NEMO) Working Group, IETF.

Performance Enhancement Schemes of OFDMA System for Broadband Wireless Access

Dong-Hyun Park, So-Young Yeo, Jee-Hoon Kim, Young-Hwan You, and Hyoung-Kyu Song

uT Communication Research Institute, Sejong University, Seoul, Korea
mith80@sdc.sejong.ac.kr, yeossoh@sdc.sejong.ac.kr, bid4u@sdc.sejong.ac.kr, yhyou@sejong.ac.kr, songhk@sejong.ac.kr

Abstract. Orthogonal frequency-division multiple access (OFDMA), which is a combination of orthogonal frequency-division multiplexing (OFDM) with frequency-division multiple access (FDMA), is regarded as a promising solution for enhancing the performance of interactive wireless systems in ubiquitous mobile communication environment. In such an application, this paper presents an investigation into improving the channel estimation scheme and of the effects of symbol timing misalignment when OFDMA is used as an access scheme. Under OFDMA uplink channel environments, appropriate symbol length of CAZAC sequences as a preamble could be utilized in accordance with the number of transmitting antenna and channel condition. The effect of the number of CAZAC sequences for channel estimation is also presented in terms of mean square error (MSE). Taking into account the effect of multiple access interference (MAI) introduced by a symbol timing misalignment, the symbol error rate (BER) and throughput performance are investigated for a typical OFDMA uplink scenario.

Keywords: OFDMA, channel estimation, MAI, CAZAC.

1 Introduction

Mobile users are demanding anywhere and anytime access to high-speed data real-and non-real time multimedia services from next-generation wireless systems. In accordance with the requirements of users, the future generations of broadband wireless communications will provide to subscribers the high quality of service (QoS) and bit rates by employing a variety of techniques. Also, they will support the future ubiquitous communications systems. In wireless multi-user environments, one of the reliable solutions for such communication systems is OFDMA technology. OFDMA, also referred to as Multiuser-OFDM, is being considered as a modulation and multiple access method for 4-th generation wireless networks. OFDMA is an extension of Orthogonal Frequency Division Multiplexing (OFDM), which is currently the modulation of choice for high speed data access systems such as IEEE 802.11a/g wireless LAN (WiFi) and IEEE 802.16a/d/e wireless broadband access systems (WiMAX) [1]-[3].

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 344–351, 2007.
© Springer-Verlag Berlin Heidelberg 2007

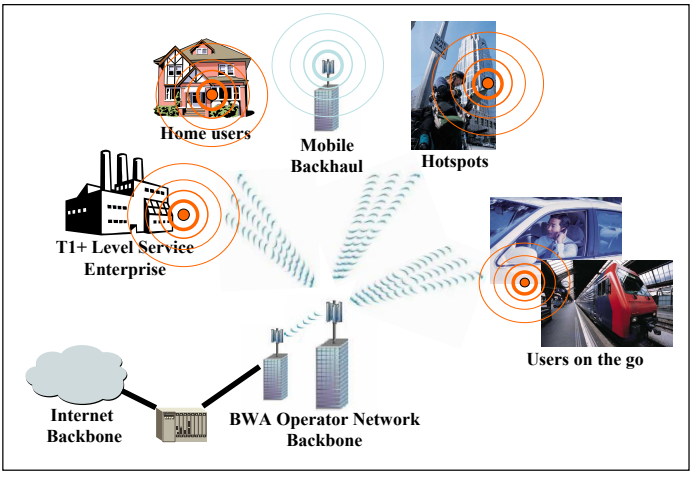


Fig. 1. Illustration of the reference model for interactive OFDMA system

The Evolution of OFDM to OFDMA completely preserves the robustness against multipath propagation and high bandwidth efficiency. However, in case of using multiple antenna for OFDM systems, recovering the transmitted signal is impossible without knowing channel coefficients because the signals from each multipath channel are overlapped. Therefore, channel estimation processing is a major key as recovering the corrupted signal [4]. In a typical OFDMA uplink scenario, moreover, multiple access interference (MAI) caused by symbol timing errors destroys the orthogonality among users [5]-[7]. So, this paper shows our investigations for improving the performance of channel estimation and symbol timing misalignment on OFDMA uplink. At the same time, we provide an algorithm which generates extended-CAZAC (E-CAZAC) sequences to overcome a limitation of the number of transmiting antennas and multipath components to maintain the orthogonality for both uplink and downlink of OFDMA systems. Besides, the Bit error rate (BER) and throughput performance of the interactive OFDMA uplink system for symbol timing misalignment is considered.

The outline the paper is organized as follows. In Section 2, symbol timing error model for OFDMA uplink is described. Section 3 gives the improving channel estimation scheme for both uplink and downlink OFDMA system. Section 4, simulation and numerical results show the mean square error (MSE) of channel estimation and BER performances of OFDMA uplink systems. Finally, the concluding remarks are given Section 5.

2 System Model

In this section we are concerned with the evaluation of the uplink subject to OFDMA system as shown in Fig. 1. The subcarrier frequencies from all users form a set of N orthogonal carriers by appropriate choice of the spacing, as it is

done in OFDM. As such, each tone in an OFDM symbol is used by a different uplink user. Thus, the users share the same bandwidth at the same time, however orthogonality is achieved by assigning distinct tones to distinct users. These tone assignment schemes consist of *interleaved tone assignment scheme* that are regularly interleaved across the overall set of N tones and *block tone assignment scheme* that are assigned to each user with disjoint blocks whose size is $\mathbf{K}_u$, where $\mathbf{K}_u$ is the number of subcarrier per each user. Usually, The interleaved scheme yields a worse signal to interference ratio (SIR) compared to the block scheme [8]. With the above block scheme in mind, the transmitted signal of the n-th subcarrier for the d-th user in the time domain can be expressed as

$$x_{n,d} = \frac{1}{N} \sum_{k=M_d(d-1)}^{M_d \cdot d-1} X_{k,l} \exp\left[-j2\pi \frac{kn}{N}\right] -N_G \leq n < N, \tag{1}$$

where k is a subcarrier index and $M_d = \frac{N}{N_u}$ is the number of subcarriers assigned to each user. Note that $x_{n,d}$ incorporates the GI of length $N_G \leq N$.
The transmitted signal by each user goes through multipath fading channel and the received signal can be expressed as

$$r_n = \frac{1}{N} \sum_{d=1}^{N_u} \sum_{k=M_d(d-1)}^{M_d \cdot d-1} H_{k,d} X_{k,d} \cdot \exp\left[-j2\pi n \frac{(k-\tau_d)}{N}\right] + W_n, \tag{2}$$

where W_n is the zero-mean Gaussian noise (AWGN) with two-sided power spectral density of $N_0/2$, τ_d is the timing error for the d-th user, and $H_{k,d}$ is the discrete frequency response $H_{k,d} = \alpha_{k,d} \cdot \exp(j\theta_{k,d})$. In the $H_{k,d}$, $\alpha_{k,d}$ is the independent identically distributed (i.i.d) Rayleigh random variable and $\theta_{k,d}$ is uniformly distributed random variable over $[-\pi, \pi]$.
Then, the reconstructed FFT output for the q-th user's l-th subcarrier is given by

$$R_{l,q} = \frac{1}{N} \sum_{n=0}^{N-1} \sum_{d=1}^{N_u} \sum_{k=M_d(d-1)}^{M_d \cdot d-1} H_{k,d} X_{k,d} \cdot \exp\left[-j2\pi n \frac{(k-l-\tau_d)}{N}\right] + W_n, \tag{3}$$

where $l \in S_q$ and $W_l = \sum_{n=0}^{N-1} exp[j(2\pi ln/N)]$.

In the equation (3), The MAI caused by symbol timing error τ_d of the multiple access users depends on the transmitted modulated symbol. So, it is difficult for the base station receiver to eliminate or mitigate the MAI. Also, we consider the power of the MAI. Let us assume that the data symbols $X_{k,d}$ which is the k-th output signal of the d-th user, on different subcarriers of different users are independent of each other, with a zero mean and the same average power P_a. Without loss of generality, we assume that the channel state is invariant during the observed symbol period.

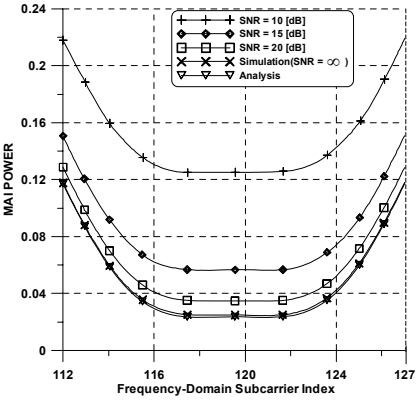


Fig. 2. Interference power plus noise versus frequence domain subcarrier index of the 8-th desired user for N=256 and K_u=16 (16QAM and $\tau_{n,8}$=65)

With the above assumptions in mind, we show that the power of the MAI [7] for k-th output signal of the d-th user is

$$E|MAI|^2 = \frac{1}{\pi^2} \sum_{n \neq d} \sum_{p=0}^{N-1} E|\alpha_{p,n}|^2 P_a \times \frac{1 - cos2\pi(p-k)(\tau_{n,d} - T_G)/T}{(p-k)^2}, \quad (4)$$

where $\tau_{n,d}$ is the symbol timing misalignment of the n-th user with respect to the d-th user at the receiver, the index p is subcarrier of n-th user. And $E|\alpha_{p,n}|^2$ is the average channel gain. T_G and T are the cyclic prefix and OFDM symbol duration including guard interval, respectively.

Fig. 2 shows the MAI power for SNR models versus frequency-domain subcarrier index of desired user (the 8-th user) for N=256 and K_u=16. In this figure, we set to the same symbol timing error ($\tau_{n,8} = 65$) for all users except the eighth user. The Simulation results were obtained by 16QAM signaling and multipath fading channel with respect to various SNR value, also analysis results were obtain through the average channel gain $E|\alpha_{p,n}|^2 = 1$. As we can see from this figure, we confirmed what subcarrier between adjacent subbands has higher MAI power. One popular solution to mitigate the MAI is to insert a guard band (GB) among two adjacent subbands and to use the receiver diversity in the base station. The improving performance of the above methods present the Section 4.

3 Improving Channel Estimation Using E-CAZAC

3.1 CAZAC Sequence

We introduce OFDMA system with multiple antenna using CAZAC preamble as system model. In section 2, in order to employ simulation, perfect channel estimation on each user is assumed. However, since channel condition has the

multipath fading features in practice, we need to estimate the channel coefficients. Also, one of the most dominant factors of the performance degradation is channel estimation error. In order to solve this problem, we achieved the a channel estimation technique using the constant-amplitude zero-autocorrelation (CAZAC) preamble which holds outstanding periodic autocorrelation property is one of the well-known algorithms [9]. The CAZAC preamble provide good and rapid signal acquisition performances even for low SNR conditions with 4-phase and 16-symbol-length. However, if the CAZAC preamble is adopted in multiple antennas system, its capability of channel estimation is limited by the number of transmit antennas (N_t) and multipath components (P) to maintain the orthogonality as follows [4]

$$1 \leq P \leq \frac{L}{N_t}, \tag{5}$$

where L presents the symbol-length of the CAZAC sequence. In order to overcome such problem, we provide an algorithm which generates extended-CAZAC (E-CAZAC).

3.2 Proposed Channel Estimation Method

The E-CAZAC sequences are obtained by zero-padded among the CAZAC sequences in order to maintain orthogonality when multiple antenna are used and multipath component are existed. The follow equation show a structure of the E-CAZAC:

$$e_{m \cdot L/4+n} = \begin{cases} j^{mn} & \text{for } n < 4 \\ 0 & \text{for } n \geq 4, \end{cases} \tag{6}$$

where $m \in 0, 1, 2, 3$, $n \in 0, 1, \cdots, L/4 - 1$. Remark that L must be times of 4.

Since such zero paddings do not provide any additional information about the power spectrum, flat power spectrum of CAZAC sequence is inherited. It shows that we can use the autocorrelation of CAZAC sequences. When E-CAZAC sequence is used, we can have more flexible capability of the channel estimation as varied L or N_t. In addition, One may think other CAZAC sequences with higher phase that better solution for multiple antennas system because all symbols are used for estimation. However, it provides the additional defect because longer sequences with higher phase causes the transmitter complexity and power consumption for transmission or correlation. Therefore, proposed scheme with conventional 4-phase CAZAC sequences can realize a multiple antennas OFDMA system with low hardware complexity and power consumption.

4 Performance Evaluation and Discussions

In this section, several simulation results which show the effect of each scheme are provided. To simulate the OFDMA system performance, a flat Rayleigh fading channel on each subcarrier is used and i.i.d fading among different subcarriers is assumed in the simulations. The entire bandwidth (BW) of 20 MHz is divided

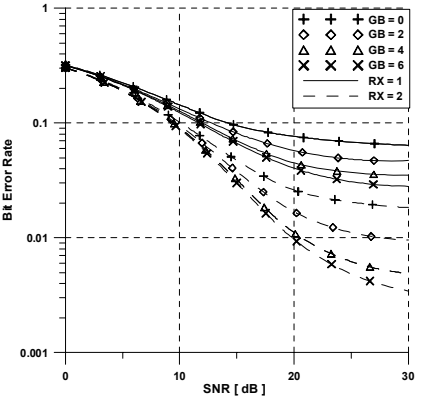


Fig. 3. BER performance of OFDMA uplink systems with respect to the number of GB for N=256 and K_u=16 (16QAM and $\tau_{n,d}$=65)

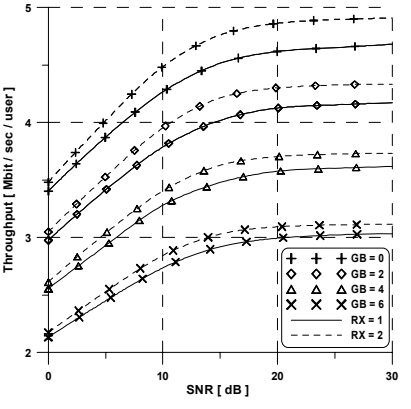


Fig. 4. Throughput performance of OFDMA uplink systems with respect to the number of GB for N=256 and K_u=16 (16QAM, $\tau_{n,d}$=65 and 20Mhz bandwidth)

into 256 subcarrier. And, the 8-th user is regarded as the desired user. In case of OFDMA uplink channel estimation, we evaluate the performance of proposed E-CAZAC sequnce in terms of MSE. Here, we adopt the least square (LS) estimator simulation.

Fig. 3 presents BER performances of OFDMA uplink systems with respect to the number of guard band. A performance enhancement can be observed with increase of the number of GB regardless of the number of receiver antennas. As expected, the BER for the existing MAI give a performance degradation since they destroy the orthogonality among users. However, As the number of guard band increase, throughput is found to decrease as shown in Fig. 4.

Fig. 5 displays the MSE performances of 32-symbol-length E-CAZAC sequence when 7 paths are presence. In the simulation, we can see that E-CAZAC

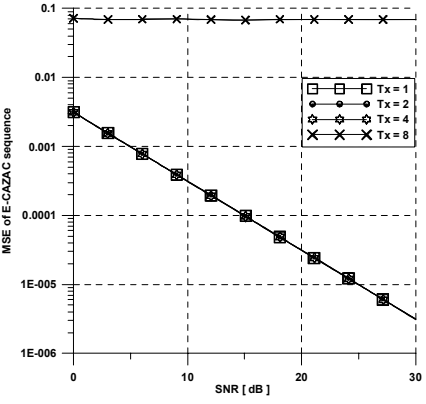


Fig. 5. The MSE performance of OFDMA uplink channel estimation using the E-CAZAC for 1, 2, 4 and 8 transmit antennas over 7-path Rayleigh faing channel

sequence performs completely when up to 4 transmit antennas are used as equation (7). In the other case, MSE performaces are very poor because the orthogonality of E-CAZAC sequence is broken. Therefore, we should notice this point. Also, if we want to use 8 transmit antennas, they will estimate the channel condition nicely as increase E-CAZAC sequence length. As expect, the MSE of CAZAC sequence gives the better performance by increasing the number of CAZAC sequence. However, we should use to be properly the symbol-length and number of CAZAC sequences restricted in conformity with the system and channel conditions.

5 Conclusions

OFDMA is a promising scheme for providing a ubiquitous environment in wireless channels. However, the both MAI problems and channel estimation problems in case of using the multiple antennas based on OFDMA system caused serious performance degradations. In this paper, we described the MAI reduction schemes in the OFDMA-based interactive wireless system. From the results presented above, we confirmed MAI reduction by using the guard band (GB) and received diversity. At the same time, by the considering system and channel conditions E-CAZAC preamble sequences can provide solutions of the more flexible channel estimation. The presented results are valid for OFDMA systems with reverse link.

Acknowledgement

This research was supported by the MIC (Ministry of Information and Communication), Korea, under the ITRC (Information Technology Research Center) support program supervised by the IITA (Institute of Information Technology Assessment).

References

1. ETSI ETS 301 958, Digital Video Broadcasting (DVB).: Interaction channel for digital terrestrial television (RCT) incorporating multiple access OFDM, ETSI, Tech. Rep., March 2002.
2. IEEE draft standard for local and metropolitan area network-part 16.: Air interface for fixed broadband wireless access systems - medium access control modifications and additional physical layer specifications for 2-11GHz, IEEE LAN MAN Standards Committee, 2002.
3. I. Koffman and V. Roman.: Broadband wireless access solutions based on OFDM access in IEEE802.16, IEEE Commun. Mag., vol. 40, pp. 96-103, April 2002.
4. Dong Jun Cho, Young Hwan You, and Hyoung Kyu Song.: Channel Estimation with Transmitter Diversity for High Rate WPAN Systems, IEICE Trans. Commun., vol. E87- B, no.11 Nov. 2004.
5. Y.-H. You, W.-G. Jeon, J.-W. Wee, Sang-Tae Kim, Intae Hwang, and H.-K. Song.: OFDMA Uplink Performance for Interactive Wireless Broadcasting, IEEE Trans. Broadcast., vol.51, no. 3, pp.383-388, 2005.
6. M. S. El-Tanany, Y. Wu, and L. Hazy.: OFDM uplink for interactive broadband wireless: analysis and simulation in the presence of carrier, clock an timing errors, IEEE Trans. Broadcast., vol. 47, no. 1, pp. 3-19, Mar. 2001.
7. M. Park, K. Ko, H. Yoo, and D. Hong.: Performance analysis of OFDMA uplink systems with symbol timing misalignment, IEEE Commun. Lett., vol. 7, no. 8, pp. 376-378, Aug. 2003.
8. A. Tonello, N. Laurenti, S. Pupolin.: On the effect of time and frequency offsets in the uplink of an asynchronous multi-user DMT OFDMA system, Proc. of Intemutioml Conference on Telecommunications 2000, Acapulco, Mexico, pp.614-618, May 22-25.2000.
9. R. C. Heimiller.: Phase Shift Pulse Codes with Good Periodic Correlation Properties, IRE Trans. Info. Theory IT-6, 254-257 October 1961.

Performance Analysis of Digital Wireless Networks with ARQ Schemes

Wuyi Yue[1] and Shunfu Jin[2]

[1] Department of Information Science and Systems Engineering
Konan University, Kobe 658-8501 Japan
yue@konan-u.ac.jp
[2] College of Information Science and Engineering
Yanshan University, Qinhuangdao 066004 China
jsf@ysu.edu.cn

Abstract. In this paper, we present a method to analyze the performance of high-reliability Internet systems in wireless environments with Automatic Repeat ReQuest (ARQ) schemes. Considering the setting up procedure of a data link in wireless networks, we build a Geom/G/1 queue model with a setup strategy to characterize the system operation, and analyze the probability distribution of the system to obtain the performance measures. Numerical results are given to evaluate the performance of different ARQ schemes in terms of the response time and the utility, and to show influence of the delay of the setup procedure and the round-trip-time on the system performance.

Keywords: ARQ, setup, wireless networks, performance analysis.

1 Introduction

Wireless networks represent a challenging and ever growing research area, supports for Internet services with excellent reliability in wireless networks are emerging requirements [1]. As a close-loop error control technique, Automatic Repeat ReQuest (ARQ) schemes have been shown to be very efficient and successful in wireless communication environments [2], [3]. In these systems, the interactions among the different network layers are very complex and their effects on the overall performance are not easy to accurately identify.

Queueing theory and Markov chains are used for the performance evaluation of ARQ schemes. [4] has indicated that it would be more accurate and efficient to use discrete-time queueing models than continuous-time queueing models when analyzing and designing digital transmitting systems.

The classical discrete-time queuing analyses have been presented in [4], [5]. Analysis of a discrete-time queueing model with setup strategy can be found in [6]. A Geom/G/1 queue model with a setup/close-delay/close-down strategy was built and analyzed by using an imbedded Markov chain in [7]. The same model using the approach of factorization principles with general vacations was analyzed in [8].

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 352–359, 2007.
© Springer-Verlag Berlin Heidelberg 2007

Extensive researches of advanced ARQ schemes as well as some performance analyses based on classical ARQ schemes have been conducted in [2], [3] and [9]. However, some simplifying assumptions considered in these studies do not hold in practice. For example: the setting up procedure of a data link was neglected and the round-trip-time was omitted. It is important to present influence of the delay of the setting up procedure and the round-trip-time on the system performance in such wireless communication networks.

To give actual system models and high quality performance evaluation, we relax these simplifications in this paper. Considering a memoryless session initiated by users, and taking into account the delay of the setting up procedure of a data link and the round-trip-time, we build a Geom/G/1 queue model with a setup strategy to characterize the system operation. Then we analyze the performance of the system with different ARQ schemes in terms of the response time of data frames and the utility ratio of the system. Based on numerical results, we evaluate the system performance and show influence of the delay of the setting up procedure and the round-trip-time on the system performance.

2 System Model

The system consists of a paired transmitter and receiver. When two adjacent users need to communicate each other, a data link must be set up by exchanging some control signals. When the data transmission finishes, the data link should be released. The system works as follows. And this process will be repeated.

(1) When a data frame arrives in the system, a setup period called U will be started, where U corresponds to a time period for setting up a data link.
(2) After the setup period U finishes, a busy period called Θ will begin. Here we define the busy period Θ to be a time period in which data frames are transmitted continuously until the buffer of the transmitter becomes empty.
(3) When there are no any data frames in the buffer of the transmitter to be transmitted, the data link will be released and the system will enter an idle period called I. A data frame arriving during I will make the system to enter a new setup period U again.

The time axis is divided into slots of fixed length. The length of a slot is defined as a time period from the instant that the first bit of a data frame is sent out to the instant that the last bit of the data frame is sent out.

A Geom/G/1 queue model with a setup strategy is presented to analyze the probability behavior of such a system. Considering the memoryless character of the user's initiated session, we assume that the input follows a Bernoulli process. Namely, the number of data frames arriving in a slot is assumed to be 1 with the probability λ, and 0 with the probability $1 - \lambda$.

We define a transmission period B called delivery delay as being the time period taken to successfully transmit a data frame. Namely, that is the time period from the instant for the first transmission of a data frame to the instant for the departure of the data frame from the transmitter buffer.

There are three kinds of basic ARQ schemes: Stop-and-Wait ARQ scheme, Go-Back-N ARQ scheme and Selective-Repeat ARQ scheme.

For all kinds of the ARQ schemes, the actual delivery of a data frame only occurs after the correct reception of all data frames with lower identifier, so we can assume that data frames arriving in the buffer with an infinite capacity are transmitted using a common data link, one by one, in a FIFO discipline.

3 Performance Analysis

Setup period U and transmission period B are assumed to be independent discrete-time random variables in slots, and are assumed to be generally distributed. The Probability Generation Function (P.G.F.) $U(z)$ and the average $E[U]$ of U, as well as the P.G.F. $B(z)$ and the average $E[B]$ of B are given as:

$$\begin{aligned} u_k &= P\{U = k, k \geq 1\}, \quad U(z) = \sum_{k=1}^{\infty} u_k z^k, \quad E[U] = \sum_{k=1}^{\infty} k u_k, \\ b_k &= P\{B = k, k \geq 1\}, \quad B(z) = \sum_{k=1}^{\infty} b_k z^k, \quad E[B] = \sum_{k=1}^{\infty} k b_k. \end{aligned} \tag{1}$$

We present the performance analysis of the system in a state of equilibrium, namely $\rho = \lambda E[B] < 1$.

3.1 Queueing Length and Waiting Time

In a late-arrival system with immediate access, we assume that data frame arrivals and departures occur only at the boundaries of slots. The arrival of a data frame during the slot t is assumed to occur at the instant t^-, and the departure of a data frame during the slot t is assumed to occur at the instant t.

Let $Q_n = Q\left(t_n^+\right)$ be the number of data frames in the system immediately after the nth data frame departed. Then $\{Q_n, n \geq 1\}$ forms an imbedded Markov chain. We define the state of the system by the number Q of the data frames in the system at the imbedded Markov points. Then we have that

$$Q_{n+1} = \begin{cases} Q_n - 1 + A_B, & Q_n \geq 1 \\ A_B + A_U, & Q_n = 0 \end{cases} \tag{2}$$

where A_U and A_B be the numbers of data frames arriving during U and B. From Eq. (2), we can give the P.G.F. $Q(z)$ and the average $E[Q]$ of Q as follows:

$$\begin{aligned} Q(z) &= \frac{(1-\rho)(1-z)B(1-\lambda+\lambda z)}{B(1-\lambda+\lambda z)-z} \cdot \frac{1-zU(1-\lambda+\lambda z)}{(1+pE[U])(1-z)}, \\ E[Q] &= \rho + \frac{B^{(2)}}{2(1-\rho)} + \frac{2\lambda E[U] + \lambda^2 U^{(2)}}{2(1+\lambda E[U])} \end{aligned} \tag{3}$$

where $U^{(2)}$ and $B^{(2)}$ are the second factorial moments of $U(z)$ and $B(z)$.

We denote by W the waiting time of a data frame. Under the FIFO discipline, the number of data frames left in the system immediately after the transmission of a data frame is identical to the sum of the number of data frames arriving during the waiting time W and the number of data frames arriving during the transmitting time B. Then the P.G.F. $Q(z)$ of Q can also be written as follows:

$$Q(z) = W\left(1-\lambda+\lambda z\right) B\left(1-\lambda+\lambda z\right). \tag{4}$$

By substituting $Q(z)$ of Eq. (3) to Eq. (4), we can obtain the P.G.F. $W(z)$ and the average $E[W]$ of W as follows:

$$\begin{aligned} W(z) &= \frac{(1-\rho)(1-z)}{\lambda B(z) - z + 1 - \lambda} \cdot \frac{\lambda - (z-1+\lambda)U(z)}{(1+\lambda E[U])(1-z)}, \\ E[W] &= \frac{\lambda B^{(2)}}{2(1-\rho)} + \frac{2E[U] + \lambda U^{(2)}}{2(1+\lambda E[U])}. \end{aligned} \tag{5}$$

3.2 Busy Period and Busy Cycle

We define the busy cycle called R as a time period from the instant in which a busy period is completed to the instant in which the next busy period ends. Let $E[R]$, $E[\Theta]$ and $E[I]$ be the averages of the busy cycle R, the busy period Θ and the idle period I, respectively. We have that

$$E[R] = E[U] + E[\Theta] + E[I] \tag{6}$$

where $E[U]$ is defined in Eq. (1), and $E[\Theta]$ and $E[I]$ will be given below.

Each data frame at the beginning of a busy period Θ will introduce a sub-busy period θ. All θ brought by the data frames at the beginning of the busy period Θ combine to make a total busy period Θ in the system. Therefore, the P.G.F. $\Theta(z)$ and the average $E[\Theta]$ of Θ are given as follows:

$$\Theta(z) = \theta(z)U(1-\lambda+\lambda\theta(z)), \quad E[\Theta] = \frac{(1+\lambda E[U])E[B]}{(1-\rho)}. \tag{7}$$

Considering the Bernoulli process, the average $E[I]$ of I is given as follows:

$$E[I] = \frac{1}{\lambda}. \tag{8}$$

Substituting Eqs. (7) and (8) to Eq. (6), the average $E[R]$ of R is given by

$$E[R] = \frac{1+\lambda E[U]}{\lambda(1-\rho)}. \tag{9}$$

4 Performance Measures for ARQ Schemes

Based on the analysis above, we present some performance measures as follows.

Response time T is defined as the total delay of a data frame. T is subdividedinto two parts. One is the waiting time W of this data frame, which is

the time spent in the buffer before its transmission. The other is corresponding to the transmission period B of this data frame. The average $E[T]$ of T is given by

$$E[T] = E[W] + E[B]. \tag{10}$$

Next, we define utility η as a ratio of the time period (one slot) in which a data frame is being transmitted correctly on a data link to the transmission period B of the data frame. Clearly, the utility η can be given by

$$\eta = \frac{1}{E[B]}. \tag{11}$$

To give the formulas for the performance measures of different kinds of ARQ schemes, the following assumptions and notions are introduced: (1) The transmissions of the ACK frame and the NACK frame are error free, and the lengths of the ACK frame and the NACK frame are omitted. (2) The rate of the transmission error is e $(0 \leq e \leq 1)$. Each data frame is correctly transmitted with probability $v = 1 - e$ $(0 \leq v \leq 1)$, and each data frame will be transmitted or retransmitted until correct reception is achieved. (3) The round-trip-time is assumed to be d slots as a system parameter.

Let N be the number of times of transmission needed for a data frame to be received correctly. Then the probability distribution and the P.G.F. $N(z)$ of N can be given as follows:

$$P\{N = n\} = (1-v)^{n-1}v, \quad N(z) = \sum_{n=1}^{\infty} P\{N = n\}z^n = \frac{vz}{1-(1-v)z} \tag{12}$$

where $n = 1, 2, ...$

4.1 Measures for Stop-and-Wait ARQ Scheme

In the system with Stop-and-Wait ARQ scheme, each transmission will take $1+d$ slots, no matter whether the transmission is correct or not. From Eq. (12), we can obtain the P.G.F. $B_{SW}(z)$, the average $E[B_{SW}]$ and the second factorial moment $B_{SW}^{(2)}$ of the transmission period B_{SW} for this case as follows:

$$B_{SW}(z) = N(z^{1+d}) = \frac{vz^{1+d}}{1-(1-v)z^{1+d}}, \tag{13}$$

$$E[B_{SW}] = \frac{1+d}{v}, \tag{14}$$

$$B_{SW}^{(2)} = \frac{(1+d)(vd+2(1-v)(1+d))}{v^2}. \tag{15}$$

Substituting Eqs. (14) and (15) to Eq. (10), we obtain the average response time $E[T_{SW}]$ as follows:

$$\begin{aligned} E[T_{SW}] &= E[W_{SW}] + E[B_{SW}] \\ &= \frac{\lambda\left((1+d)^2(2-v) + vd(1+d)\right)}{2(v-\lambda(1+d))v} + \frac{2E[U] + \lambda U^{(2)}}{2(1+\lambda E[U])} + \frac{1+d}{v} \end{aligned}$$

where the average waiting time $E[W_{SW}]$ of data frames is given by Eq. (5).

Substituting Eq. (14) to Eq. (11), we give the utility ratio η_{SW} as follows:

$$\eta_{SW} = \frac{1}{E[B_{SW}]} = \frac{v}{1+d}.$$

4.2 Measures for Go-Back-N ARQ Scheme

In the system with Go-Back-N ARQ scheme, each error transmission occupies $1+d$ slots, while the last correct transmission takes 1 slot. From Eq. (12), we can obtain the P.G.F. $B_{GBN}(z)$, the average $E[B_{GBN}]$ and the second factorial moment $B^{(2)}_{GBN}$ of the transmission period B_{GBN} for this case as follows:

$$B_{GBN}(z) = \frac{N(z^{1+d})}{z^d} = \frac{vz}{1-(1-v)z^{1+d}}, \tag{16}$$

$$E[B_{GBN}] = \frac{1+(1-v)d}{v}, \quad B^{(2)}_{GBN} = \frac{(1-v)(1+d)(2+2d-vd)}{v^2}. \tag{17}$$

By the same way, we can obtain the average response time $E[T_{GBN}]$ and the utility ratio η_{GBN} for the Go-Back-N ARQ scheme as follows:

$$\begin{aligned} E[T_{GBN}] &= E[W_{GBN}] + E[B_{GBN}] \\ &= \frac{\lambda(1-v)(1+d)(2+2d-vd)}{2(v-\lambda(1+(1-v)d))v} + \frac{2E[U]+\lambda U^{(2)}}{2(1+\lambda E[U])} + \frac{1+(1-v)d}{v}, \\ \eta_{GBN} &= \frac{1}{E[B_{GBN}]} = \frac{v}{1+(1-v)d}. \end{aligned}$$

where the average waiting time $E[W_{GBN}]$ of data frames is given by Eq. (5).

4.3 Measures for Selective-Repeat ARQ Scheme

In the system with Selective-Repeat ARQ scheme, each transmission, no matter whether it is correct or not, takes 1 slot. From Eq. (12), we can obtain the P.G.F. $B_{SR}(z)$, the average $E[B_{SR}]$ and the second factorial moment $B^{(2)}_{SR}$ of the transmission period B_{SR} for this case as follows:

$$B_{SR}(z) = N(z) = \frac{vz}{1-(1-v)z}, \quad E[B_{SR}] = \frac{1}{v}, \quad B^{(2)}_{SR} = \frac{2(1-v)}{v^2}. \tag{18}$$

We can also give the average response time $E[T_{SR}]$ and the utility ratio η_{SR} for the Selective-Repeat ARQ scheme as follows:

$$\begin{aligned} E[T_{SR}] &= E[W_{SR}] + E[B_{SR}] = \frac{\lambda(1-v)}{(v-\lambda)v} + \frac{2E[U]+\lambda U^{(2)}}{2(1+\lambda E[U])} + \frac{1}{v}, \\ \eta_{SR} &= \frac{1}{E[B_{SR}]} = v \end{aligned}$$

where the average waiting time $E[W_{SR}]$ of data frames is given by Eq. (5).

5 Numerical Results

In numerical results, considering the prevalent wireless network application, we let the transmission rate be 50 Mbps. To ensure the latest conflict signal be sensed by the transmitter before a data frame is completely sent out, we assume the size of a data frame to be 1250 bytes and the round-trip-time to be 0.1 ms. The setup period U is geometrically distributed with an average value of 0.2 ms.

By using these parameters, we show the average response time $E[T]$ as a function of the arrival rate λ (Data frames/Slot) with the rate of the transmission error $e = 0.1$ and the utility ratio η as a function of the rate of the transmission error for different kinds of ARQ schemes in Figs. 1-2, where the schemes of the Stop-and-Wait ARQ, Go-Back-N ARQ and Selective-Repeat ARQ are abbreviated as SW ARQ, GBN ARQ and SR ARQ, respectively.

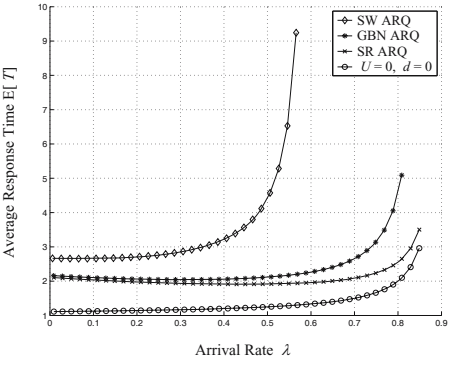


Fig. 1. Average response time $E[T]$

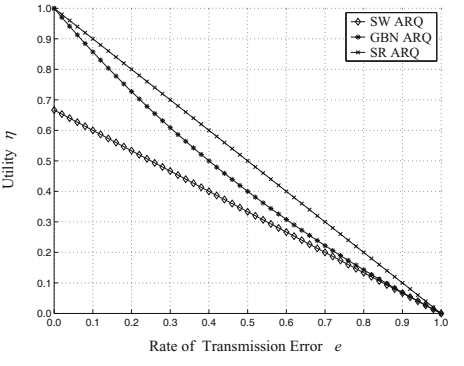


Fig. 2. Utility η

In Fig. 1, the curve of the lowest position shows the case without considering the delay of the setup procedure and the round-trip-time ($U = 0$ and $d = 0$) for all the ARQ schemes. Namely, this case is that the setup procedure and the round-trip-time are not considered in the performance analysis and evaluation of the system as in previous papers [2], [3], [9].

Other curves of Fig. 1 are to show how the average response time $E[T]$ changes depend upon the delay of the setup procedure and the round-trip-time. It can be observed that when the arrival rate λ is low, as λ increases, there is a bit increase of $E[T]$ in Stop-and-Wait ARQ scheme, while there is a slight decrease for Go-Back-N ARQ scheme and Selective-Repeat ARQ scheme. This is because that the setup procedure for a data link has a heavier impact on Go-Back-N ARQ scheme and Selective-Repeat ARQ scheme than on Stop-and-Wait ARQ scheme. As λ further increases, all the curves of these schemes increase suddenly, it is due to the heavy offered load ρ.

In Fig. 2, the utility ratio η for all the ARQ schemes without considering the delay of the setup procedure and the round-trip-time ($U = 0$ and $d = 0$) is just the same as the Selective-Repeat ARQ scheme. We can observe that the

lower the rate of the transmission error e is, the better the utility ratio η will be for all the cases. We can also know that among these three schemes, the system performance of Selective-Repeat ARQ scheme is better than other two schemes. However, we also notice that when the rate of the transmission error e is very high or very low, the utility η for Go-Back-N ARQ scheme and for Selective-Repeat ARQ scheme tends to be the same.

6 Conclusions

Performance analysis of high-reliability Internet systems in wireless environments with ARQ schemes was presented. Considering the setting up procedure of a data link, we built a Geom/G/1 queue model with a setup strategy to characterize the system operation. Taking into account the round-trip-time, we analyzed the stationary distribution of the system and gave the formulas of performance measures for different kinds of ARQ schemes. We presented numerical results to evaluate and compare the performance of the systems with different kinds of ARQ schemes in terms of the average response time of data frames and the utility ratio of the system, and to show influence of the delay of the setting up procedure and the round-trip-time on the system performance.

Acknowledgments. This work was supported in part by MEXT.ORC (2004-2008), Japan and in part by NSFC (No. 10671170) and MADIS, China.

References

1. Yue W. and Matsumoto Y.: Performance Analysis of Multi-Channel and Multi-Traffic on Wireless Communication Networks. Kluwer Academic Publishers (2002)
2. Badia L., Rossi M. and Zorzi M.: SR ARQ Packet Delay Statistics on Markov Channels in the Presence of Variable Arrival Rate. IEEE Transaction on Wireless Communication. **5** (2006) 1639-1644
3. Zheng H. and Viswanthan H.: Optimizing the ARQ Performance in Downlink Packet Data Systems With Scheduling. IEEE Transaction on Wireless Communication. **4** (2005) 495-506
4. Takagi H.: Queueing Analysis. Vol. 3: Discrete-Time Systems. North-Holland (1993)
5. Tian N. and Zhang G.: Vacation Queueing Models-Theory and Applications. Springer-Verlag (2006)
6. Tian N. and Zhang G.: The Discrete Time GI/Geo/1 Queue With Multiple Vacations. Queueing Systems. **40** (2002) 283-294
7. Jin S. and Tian N.: Performance Evaluation of Virtual Channel Switching System Based on Discrete Time Queue. Journal of China Institute of Communications. **25** (2004) 58-68 (in Chinese)
8. Jin S., Yue W. and Liu M.: Performance Analysis of SVC Based on Discrete Time Factorization Principle With General Vacations. Technical Report of IEICE. **102** (2005) 1-6
9. Hou F., Ho P., Xue M. and Zhang Y.: Performance Analysis of Differentiated ARQ Scheme for Video Transmission over Wireless Networks. International Workshop on Modeling Analysis and Simulation of Wireless and Mobile Systems (2005) 1-7

A Novel Frequency Offset Estimation Algorithm Using Differential Combining for OFDM-Based WLAN Systems*

Sangho Ahn[1], Sanghun Kim[1], Hyoung-Kee Choi[1], Sun Yong Kim[2], and Seokho Yoon[1]

[1] School of Information and Communication Engineering, Sungkyunkwan University, 300 Chunchun-dong, Jangan-gu, Suwon, Kyunggi-do, 440-746, Korea
{ash9252,hkchoi,ksh7150,syoon}@ece.skku.ac.kr
[2] Department of Electronics Engineering, Konkuk University, 1 Hwayang-dong, Gwangjin-gu, Seoul 143-701, Korea
kimsy@konkuk.ac.kr

Abstract. The timing offset is one of the main error sources in estimating the frequency offset in orthogonal frequency division multiplexing (OFDM)-based wireless local area network (WLAN) systems. Although some works have been done to mitigate the effect of the timing offset on the frequency offset estimation, most of the investigations require the knowledge of the timing offset range, which is not generally available in practical systems. In this paper, we propose a new frequency offset estimation algorithm using differential combining between two successive correlation samples, which does not require the knowledge of the timing offset range, and thus, is robust to the timing offset variation. The simulation results show that the proposed algorithm is not only robust to the timing offset variation, but also generally performs better than the conventional algorithm, in the case of the timing offset range being not known exactly.

Keywords: OFDM, WLAN, frequency offset estimation.

1 Introduction

Future wireless and mobile systems are envisioned to provide reliable and high-speed multimedia services to users. Among the wireless and mobile technologies proposed so far, OFDM technology has been attracting considerable research interest for wireless and mobile applications [1] and selected as a modulation scheme in wireless local area network (WLAN) standards, such as institute of electrical and electronics engineers (IEEE) 802.11a, high performance local area network type 2 (HiperLAN/2), and mobile multimedia access communication

* This research was supported by grant No. R01-2004-000-10690-0 from the Basic Research Program of the Korea Science & Engineering Foundation. Dr. Yoon is the corresponding author.

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 360–367, 2007.
© Springer-Verlag Berlin Heidelberg 2007

(MMAC) [2], [3]. This is because OFDM offers many advantages including high spectral efficiency and immunity to multipath fading and impulsive noise [1], [4].

However, OFDM-based systems are very sensitive to the frequency offset [5], which could destroy the orthogonal property among subcarriers, and consequently, deteriorate the overall performance of OFDM-based systems. Thus, the frequency offset estimation is one of the most important technical issues in OFDM-based systems [6], [7].

One of the main error sources in estimating the frequency offset in OFDM-based systems is the timing offset [4], [8]. Some techniques have been proposed to alleviate the effect of the timing offset on the frequency offset estimation [8]-[10]; however, most of the investigations require the knowledge of the timing offset range, which is not generally available in practical systems. Thus, it would be useful to develop an algorithm that does not require the knowledge of the timing offset range.

In this paper, we propose a new frequency offset estimation algorithm using differential combining between two successive correlation samples for OFDM-based WLAN systems. The timing offset causes a phase drift among the correlation samples used for frequency offset estimation, and thus, hinders the effective accumulation of the correlation samples, which is essential to obtain a large correlation value for the detection of the correct frequency offset estimate. In the proposed algorithm, we remove the phase drift due to the timing offset via differential combining between two successive correlation samples, and thus, can obtain a large correlation value for frequency offset estimation, regardless of the timing offset value. The simulation results show that the proposed algorithm is not only robust to the timing offset variation, but also performs better than the conventional algorithm on average, in the case of the timing offset range being not known exactly.

2 Signal Model

In OFDM-based systems, the transmitted symbol is generated by the inverse fast Fourier transform (IFFT) and expressed as

$$z(m) = \sum_{k=0}^{N-1} Z_k e^{j2\pi \frac{k}{T_s} \frac{T_s}{N} m}, \tag{1}$$

where m is the discrete time index, Z_k is a phase shift keying (PSK) or quadrature amplitude modulation (QAM) data symbol of the kth subcarrier, T_s is the symbol period, and N is the IFFT size.

For a zero-mean additive white Gaussian noise (AWGN) channel, the received signal can be expressed as

$$y(m) = z(m-\tau)e^{j2\pi \Delta f(m-\tau)/N} + w(m), \tag{2}$$

where Δf is the frequency offset normalized to the subcarrier spacing $\frac{1}{T_s}$, τ is the timing offset normalized to the sample interval T_s/N, and $w(m)$ is the AWGN. For the sake of simplicity, Δf is assumed to be an integer.

The receiver first demodulates the received OFDM symbol using the FFT operation, and then, yields the FFT output, Y_l, corresponding to the lth subcarrier as follows:

$$Y_l = Z_{l-\Delta f} e^{\frac{-j2\pi(l-\Delta f)\tau}{N}} + W_l, \tag{3}$$

where W_l is the FFT output of $w(m)$ corresponding to the lth subcarrier. From (3), we can see that the FFT output of the received OFDM symbol is cyclically shifted by the frequency offset and its phase is rotated by the timing offset.

3 The Effect of Timing Offset on Frequency Offset Estimation

In this paper, we consider the frequency offset estimation system employing a training sequence, as in [8]-[10]. Then, an estimate $\Delta\hat{f}$ of Δf can typically be obtained as

$$\Delta\hat{f} = \arg\max_{d} \left\{ \left| \sum_{k=0}^{N-1} Z_k^* Y_{(k+d)_N} \right| \right\}, \tag{4}$$

where Z_k is the locally generated training sequence, $*$ denotes the complex conjugate, d is the amount of cyclic shift, and $(\cdot)_N$ is the modulo-N operator. When Δf is estimated correctly, that is $\Delta\hat{f} = \Delta f$, the correlation value (normalized to $N|Z_k|^2$) in (4), if AWGN is not considered, becomes $\left| e^{\frac{-j\pi\tau(N-1)}{N}} \frac{\sin(\pi\tau)}{\sin(\pi\tau/N)} \right|$ and is plotted as a function of the timing offset τ, as shown in Fig. 1. From Fig. 1, we can clearly observe that the correlation value used for the frequency offset estimation is very sensitive to the variation of the timing offset, which implies

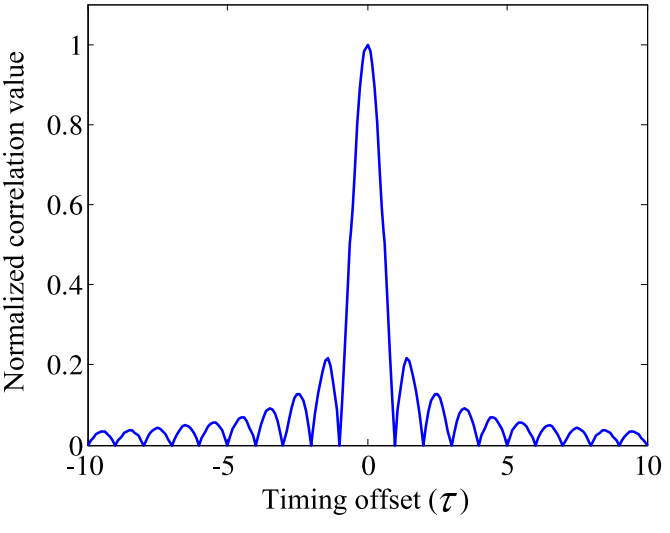


Fig. 1. Correlation value as a function of the timing offset ($N = 1024$)

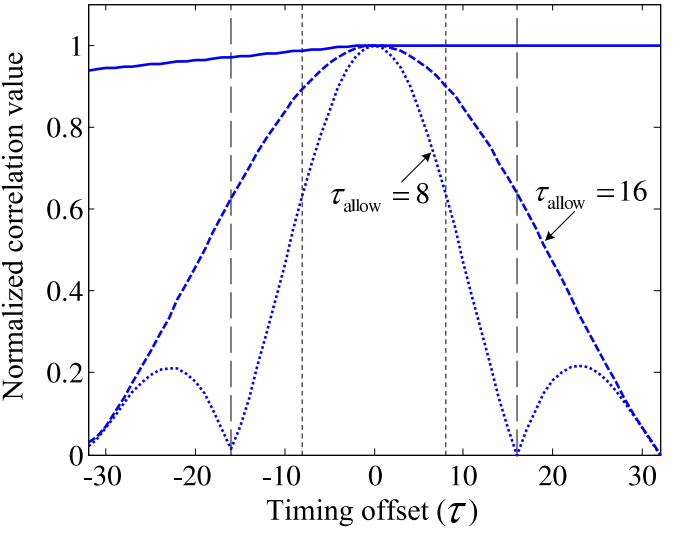


Fig. 2. Correlation value of the proposed (solid line) and conventional (dotted and dashed lines) algorithms ($N = 1024$, cyclic prefix = 100)

that the correlation value could be reduced significantly due to the timing offset even if the frequency offset is correctly estimated, resulting in considerable degradation in estimation performance (note that a large correlation value when $\Delta\hat{f} = \Delta f$, is essential for the detection of the correct frequency offset estimate).

In the conventional algorithm, the frequency offset estimation in the the presence of the timing offset is achieved using the coherence phase bandwidth (CPB), which is the maximum correlation range in which the correlation value increases monotonically, and depends on the timing offset range strongly. Denoting the CPB and allowed timing offset by BW_c and τ_{allow}, respectively, we get

$$BW_c \approx \frac{N}{2\tau_{allow}} \tag{5}$$

and using which we obtain the estimate $\Delta\hat{f}$ of Δf as

$$\Delta\hat{f} = \arg\max_{d} \left\{ \sum_{n=0}^{K-1} \left| \sum_{k=0}^{BW_c-1} Z^*_{k+nBW_c} Y_{(k+nBW_c+d)_N} \right| \right\}, \tag{6}$$

where K is the number of blocks divided by the CPB and τ_{allow} is equal to N/BW_c. As shown in (6), the conventional algorithm compensates for the correlation value reduction due to the timing offset by adding the absolute values of each partial correlation over BW_c. Fig. 2 shows the correlation value (normalized to $N|Z_k|^2$) in (6), where the dotted and dashed lines represent the cases of $\tau_{allow} = 8$ and 16, respectively. As shown in the figure, the correlation value decreases rapidly as the timing offset becomes larger than τ_{allow}, which results

in the significant performance degradation in estimating Δf. That is, the conventional algorithm requires the knowledge of the timing offset range to set a value of τ_{allow} (or BW_c) demanded for its proper operation.

4 Proposed Algorithm

To mitigate the effect of the timing offset on the frequency offset estimation, we perform the differential combining between two successive correlation samples, $Z_k^* Y_{(k+d)_N}$ and $Z_{k+1}^* Y_{(k+1+d)_N}$. Then, the differentially combined components $\left\{ \left(Z_k^* Y_{(k+d)_N} \right) \left(Z_{k+1}^* Y_{(k+1+d)_N} \right)^* \right\}_{k=1}^{N-1}$ become phase aligned, and thus, a large correlation value can be obtained by adding the components, regardless of the timing offset. It should be noted that each of the components $\left\{ \left(Z_k^* Y_{(k+d)_N} \right) \left(Z_{k+1}^* Y_{(k+1+d)_N} \right)^* \right\}_{k=1}^{N-1}$ is divided into real and imaginary parts. Hence, we take the envelope of $\sum_{k=0}^{N-1} \left(Z_k^* Y_{(k+d)_N} \right) \left(Z_{k+1}^* Y_{(k+1+d)_N} \right)^*$ to combine the divided parts, and finally, obtain the following frequency offset estimation algorithm:

$$\Delta \hat{f} = \arg\max_d \left\{ \left| \sum_{k=0}^{N-1} \left(Z_k^* Y_{(k+d)_N} \right) \left(Z_{k+1}^* Y_{(k+1+d)_N} \right)^* \right| \right\}, \tag{7}$$

From Fig. 2 in Section 3, we can see that the correlation value (normalized to $N \left| Z_k \right|^2 \left| Z_{k+1} \right|^2$) of the proposed algorithm is almost constant regardless of the timing offset value. It is also observed that the correlation value of the proposed algorithm slightly decreases when the timing offset value is negative, which is caused by the interference from the neighboring preamble including the cyclic prefix, as shown in Fig. 3, where CP is an abbreviation of the cyclic prefix. However, from Fig. 2, we can see that the correlation value of the proposed algorithm is still much larger than that of the conventional algorithm.

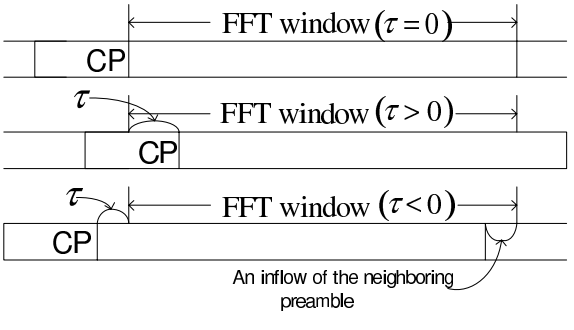


Fig. 3. Interference from the neighboring preamble when the timing offset value is negative

5 Performance Comparison

In this section, we compare the performance of the proposed algorithm with that of the conventional algorithm [10] in terms of the frequency offset estimation accuracy, in the presence of the timing offset. Simulation was performed on an AWGN channel with a fixed frequency offset 10 and an OFDM system with a guard interval (cyclic prefix) of 100 samples and 1024 subcarriers was considered. The simulation results were obtained with 10^3 Monte Carlo runs at each value of signal-to-noise ratio (SNR).

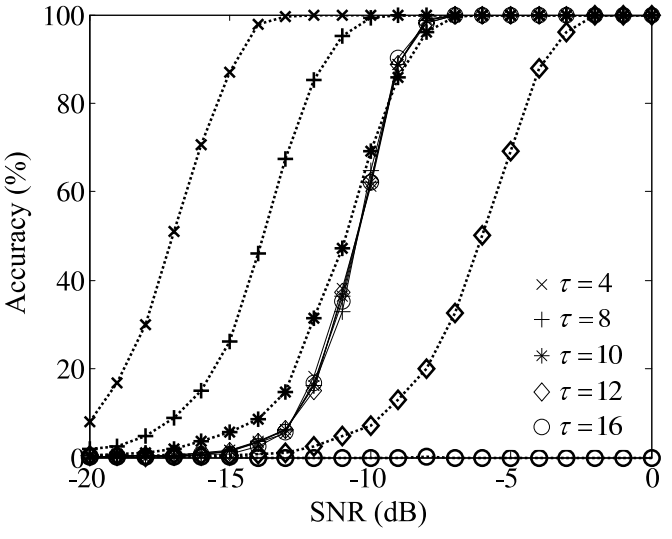


Fig. 4. Accuracy of the proposed (solid line) and conventional (dotted line) algorithms as a function of SNR when $\Delta f = 10, N = 1024, \tau_{allow} = 8$, and $\tau = 4, 8, 10, 12$, and 16

Figs. 4 and 5 show the accuracy of the proposed (solid line) and conventional (dotted line) algorithms as a function of SNR when $\tau_{allow} = 8$ and 16, respectively. From the figures, it is observed that the conventional algorithm performs better than the proposed algorithm when the timing offset τ is equal to or less than τ_{allow}; however, the performance of the conventional algorithm significantly degrades as τ gets larger than τ_{allow} and eventually becomes much worse than that of the proposed algorithm. From the result, we can see that the conventional algorithm cannot operate properly without the knowledge of the timing offset range demanded for setting τ_{allow}. On the other hand, it is seen that the proposed algorithm is much more robust to the timing offset variation and generally performs better than the conventional algorithm in practical systems, where the knowledge of the timing offset range is not available.

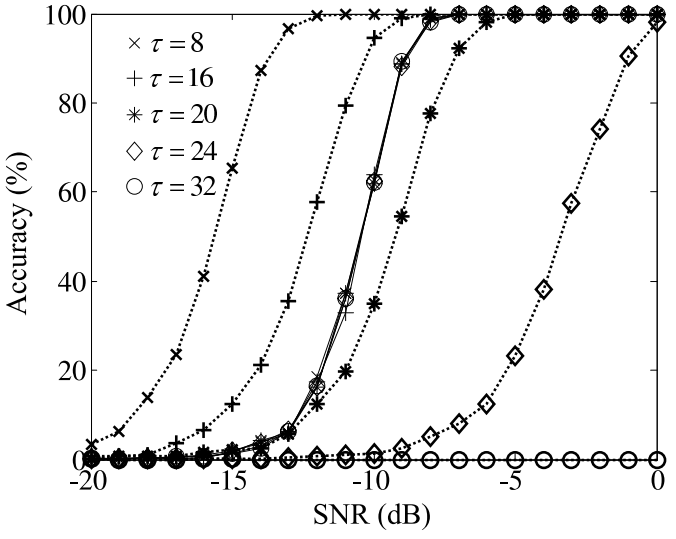


Fig. 5. Accuracy of the proposed (solid line) and conventional (dotted line) algorithms as a function of SNR when $\Delta f = 10$, $N = 1024$, $\tau_{allow} = 16$, and $\tau = 8, 16, 20, 24$ and 32

6 Conclusion

In this paper, we first investigated the effect of the timing offset on the frequency offset estimation, and then, proposed to use the differential combining between two successive correlation samples to mitigate the effect. From the comparison results, we observed that the performance of the conventional algorithm might significantly degrade and eventually become useless for the frequency offset estimation in the case where the knowledge of the timing offset range is not available. On the other hand, the proposed algorithm is very robust to the timing offset variation and generally outperforms the conventional algorithm, in the case of the timing offset range being not known exactly.

References

1. R. V. Nee and R. Prasad, *OFDM for Wireless Multimedia Communications*. London, England: Artech House, 2000.
2. IEEE Std. 802.11a-1999, "Wireless LAN MAC and PHY Specifications – High-Speed Physical Layer in the 5 GHz Band," *ISO/IEC 8802-11: 1999 (E) / Amd 1: 2000 (E)*, New York, NY: IEEE, 2000.
3. N. Prasad and A. Prasad, *WLAN Systems and Wireless IP for Next Generation Communications*. Boston, MA: Artech House, 2002.
4. J. A. C. Bingham, "Multicarrier modulation for data transmission: an idea whose time has come," *IEEE Commun. Mag.*, vol. 28, pp. 5-14, May 1990.
5. P. H. Moose, "A technique for orthogonal frequency division multiplexing frequency offset correction," *IEEE Trans. Commun.*, vol. 42, pp. 2908-2914, Oct. 1994.

6. B. Y. Prasetyo, F. Said, and A. H. Aghvami, "Fast burst synchronisation technique for OFDM-WLAN systems," *IEE Proceedings: Commun.*, vol. 147, pp. 292-298, Oct. 2000.
7. J. Li, G. Liu, and G. B. Giannakis, "Carrier frequency offset estimation for OFDM-based WLANs," *IEEE Signal Process. Lett.*, vol. 8, pp. 80-82, Mar. 2001.
8. H. Nogami and T. Nagashima, "A frequency and timing period acquisition technique for OFDM systems," *in Proc. IEEE PIRMC*, Toronto, Canada, pp. 1010-1015, Sep. 1995.
9. K. Bang, N. Cho, H. Jun, K. Kim, H. Park, and D. Hong, "A coarse frequency offset estimaion in an OFDM system using the concept of the coherence phase bandwidth," *IEEE Trans. Commun.*, vol. 49, pp. 1320-1324, Aug.2001
10. S. Kim, S. Yoon, H. -K. Choi, and S. Y. Kim, "A low complexity and robust frequency offset estimation algorithm for OFDM-based WLAN systems," *Springer-Verlag Lecture Notes in Compu. Sci.*, vol. 3992, pp. 961-968, May 2006.

Design and Performance Evaluation of High Efficient TCP for HBDP Networks

TaeJoon Park[1], ManKyu Park[2], JaeYong Lee[2,⋆], and ByungChul Kim[2]

[1] Electronics and Telecommunications Research Institute
161 Gajong-Dong, Yuseong-Gu, Daejeon, 305-350, Korea
tjpark@etri.re.kr
[2] Department of Infocom Engineering, Chungnam National University
220 Gung-Dong, Yuseong-Gu, Daejeon, 305-764, Korea
mkpark@ngn.cnu.ac.kr,{jyl,byckim}@cnu.ac.kr

Abstract. While the legacy TCP is the most commonly used and reliable transport protocol in the Internet, it is not suitable for massive data transfers in high bandwidth delay product networks. To cope with this problem, we propose high efficient TCP congestion control mechanism that can provide efficient data transfer in HBDP networks. When there is some available bandwidth and it satisfies a certain condition, the congestion window grows rapidly. Otherwise, it maintains a linear growth of the congestion avoidance phase similar to the legacy TCP congestion avoidance algorithm. Based on the relationship of the current and minimum round trip time, the proposed method selects between the linear and rapid growth phases of the congestion window update during the congestion avoidance period. To prevent packet loss during the exponential growth phase, proposed method uses not only the end-to-end delay information, but also the estimated bandwidth of the bottleneck node.

Keyword: TCP, and High Bandwidth Delay Product Networks.

1 Introduction

As TCP was designed to the reliability guaranteeing transfer protocol, it is most widely used in various network environments as the general transfer protocol. While the transmission control protocol(TCP) is the most commonly used reliable transport protocol, it is generally accepted that standard TCP is not suitable for bulk data transfers in high bandwidth delay product(HBDP) networks. Since the TCP congestion avoidance algorithm is not so dynamic, the packet drop rate needed to fill a Gigabit pipe using the current TCP protocol is beyond the limit of currently achievable fiber optic error rates. Therefore, many TCP congestion control mechanisms are proposed to solve the low efficiency problem in HBDP networks. However, it still has some problems as to the bandwidth efficiency, the RTT fairness convergence time, and etc.

⋆ Corresponding author.

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 368–375, 2007.
© Springer-Verlag Berlin Heidelberg 2007

We propose a modified congestion avoidance mechanism, HE-TCP, which controls the slow start threshold(*ssthresh*) with the relationship of current and last available bandwidth in HBDP networks : when the congestion window size is lesser than the current *ssthresh* size, the congestion window grows aggressively, and otherwise, it maintains linear increase similar to the legacy congestion avoidance algorithm. So, in spite of the compatibility about the legacy TCP, it does not have a meaning to classify according to slow start and congestion avoidance period. The proposed algorithm prevents large packet losses by adjusting congestion window size appropriately. Also, it can rapidly utilize the large available bandwidth by maintaining the *ssthresh* dynamically. The simulation results show that the proposed algorithm improves not only the utilization of the available bandwidth but also RTT fairness convergence time in high bandwidth delay product network environment.

The rest of the paper is organized as following: Section 2 describes the previous work in this area. Section 3 discusses the proposed mechanism and section 4 describes the simulation results followed by conclusions in section 5

2 Related Work

During the initial start up phase (slow start) of the traditional TCP, compliant to the RFC 2581 [1] IETF standards, exponentially increases the amount of transferred data until detecting a packet loss by a timeout or triple duplicate ACKs. When a loss is detected, the TCP halves the congestion window *cwnd*, which controls the number of packets to be sent without acknowledgement, and moves into the congestion avoidance phase. During the congestion avoidance phase, the TCP increases the *cwnd* by one packet per the *cwnd* number of received ACKs and halves the cwnd for a packet drop. Thus, we call the TCP's congestion control algorithms Additive Increase Multiplicative Decrease (AIMD) algorithms. However, the AIMD control of the traditional TCP is not dynamic enough to fill a big pipe for HBDP networks. For example, a standard TCP connection with 1500-byte packets, a 100 ms RTT and a steady-state throughput of 10 Gbps requires an average congestion window of 83,333 segments and a maximum packet drop rate of one congestion event for every 5,000,000,000 packets [2]. The average bit error rate of (2×10^{-14}) is needed to fully utilize the link in this environment; however it is almost impossible to realize the requirement for the current network technology.

Over the past few years, there has been much research efforts to solve the under utilization problem of traditional TCP in HBDP networks[2][3]. High Speed TCP(HS-TCP), recently proposed by Floyd, tries to improve the loss recovery time of standard TCP by changing the increase and decrease parameters of AIMD mechanism. The values for the additive increase range from 1 (standard TCP) to a high of 73 packets, and the range of the multiplicative decrease is from 0.5 (standard TCP) to a low of 0.09. Consequently, when a congestion event occurs over HBDP networks, HS-TCP does not drop back as much and adds more than one packet per RTT, thus recovering faster. However, HS-TCP has some

drawbacks for deployment in terms of convergence times, compatibility, fairness, and so on [4][5].

3 Congestion Control Mechanism of the HE-TCP

The proposed congestion control mechanism can be divided into three phases: SS, CA, and RI. The SS phase is the exponential growth phase; the CA phase is the additive increase phase; and the RI phase is the *ssthresh* controlled growth phase only activated in HBDP networks. Figure 1 shows an example evolution of the congestion window of the proposed HE-TCP. The proposed congestion control mechanism has provided a solution to prevent overshooting in the SS phase and a fast bandwidth reserving solution in the CA phase to switch over to the RI phase.

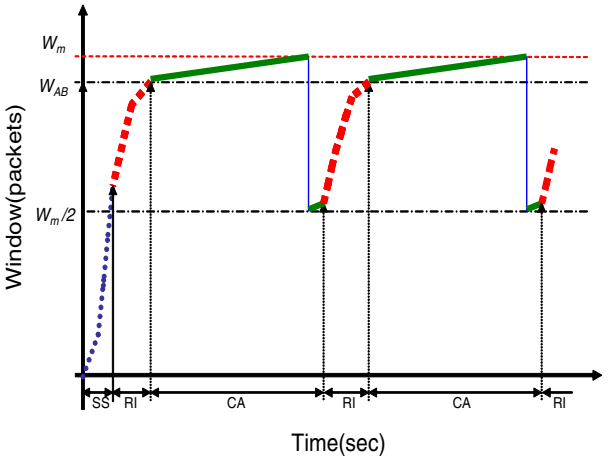


Fig. 1. Evolution example of window size including the startup phase in the proposed HE-TCP

3.1 Generalized Time-Sliding Window Estimator

To control the start time of RI, we used a generalized time-sliding window (GTSW) estimator which uses a time-varying coefficient and a exponentially-weighted moving average filter. The number of computation times per round of GTSW estimator is restricted to a predefined number because of the effect of packets clustering due to congestion and compression. Let t_k be the time instant at which the k_{th} ACK is received at the sender. Let $CurBW$ and k be the bandwidth share and the time-varying coefficient at time t_k, respectively. The GTSW estimator is defined as (1):

$$BW_k = (1 - \alpha_k) \cdot BW_{k-1} + \alpha_k \cdot CurBW \tag{1}$$

where $\alpha_k = \Delta t_k/(RTT_k+\Delta t_k)$, $\Delta t_k = t_k - t_{k-1}$, and $CurBW = (\sum_{t=t_{k-1}}^{t_k} d_t)/\Delta t_k$.

From (1), it follows that

$$\begin{aligned} BW_k &= (1 - \frac{\Delta t_k}{RTT_k + \Delta t_k}) \cdot BW_{k-1} + (\frac{\Delta t_k}{RTT_k + \Delta t_k}) \cdot CurBW \\ &= \frac{RTT_k \cdot BW_{k-1} + PacketSize \cdot PacketCnt}{RTT_k + \Delta t_k} \end{aligned} \tag{2}$$

where RTT_k is the smoothed round trip time at time t_k. There have been some previous works on TCP bandwidth measurement methods. From (2), in the case of Available Bandwidth Estimator (ABE)[6], *PacketCnt* and Δt_k are *cwnd* and RTT, respectively. Because ABE computes the optimum congestion window once every RTT, it is not dynamic enough to estimate the currently usable bandwidth within each round. The *PacketCnt* and Δt_k of Time Sliding Window (TSW)[7] Estimator are 1 and inter-packet arrival time, respectively. The TSW Estimator computes and updates the state variables upon each packet arrival which leads to the problem of much computation in HBDP networks. GTSW estimator uses a time-varying coefficient and a exponentially-weighted moving average (EWMA) filter. The GTSW estimator is very similar to the ABE and the TSW Estimator. But The *PacketCnt* and Δt_k of TSW Estimator are (*cwnd* / n). So the number of computation times per round of GTSW estimator is restricted to predefined number,n, because of the effect of packets clustering due to congestion and compression. Therefore, the GTSW estimator is appropriate under the circumstances of HBDP networks. HE-TCP sets the *ssthresh* to reflect its estimated bandwidth delay product as

$$SSthresh = \frac{BW_k \cdot BaseRTT}{PacketSize} \tag{3}$$

where *BaseRTT* is the minimum round trip time, and *PacketSize* is the packet size.

3.2 The SS Phase

The legacy TCP's SS phase is suitable for fast bandwidth reservations. To avoid a premature exit from the SS phase (Fig. 1) and to increase utilization, the initial *ssthresh* can be increased. However, a large initial *ssthresh* can cause slow start over-shooting problems and multiple packet losses, and further reduces the utilization. To compensate for the weak points, our mechanism switches the SS phase to the RI phase and limits the maximum size of the *cwnd* by resetting *ssthresh* to the estimated available bandwidth of the bottleneck node. Limiting *ssthresh* to the available bandwidth can limit the fairness in TCPs that use losses for congestion verification. Therefore, in order to estimate the fair share of available bandwidth, bandwidth estimation is computed using a GTSW estimation scheme which has both adaptive gain and adaptive sampling. Usually, most problems in the delay-based congestion control methods, such as

```
When an ACK for a new packet arrives;
if( cwnd < 10 )    SampleInterval = 1;
else               SampleInterval = (int) cwnd    10;

if( AckCnt    SampleInterval  ){
    DelTime = now - PreTime;
    T_ratio = DelTime    ( BaseRTT + DelTime );
    CurBW = ( PacketSize    AckCnt )    DelTime;
    BW = ( 1.0 - T_ratio )    PreBW + T_ratio    CurBW;
    ssthresh  = (int) ( BW    BaseRTT    PacketSize );
    PreBW = BW;
    PreTime = now;
    AckCnt = 1;
} else {
    AckCnt += 1;
}

if( cwnd < ssthresh )     cwnd += 1;
else                      cwnd += 1/cwnd;

When triple-duplicate ACKs arrive;
  cwnd = cwnd    2;
  PreBW = 0;
  ssthresh = 0;
```

Fig. 2. Pseudo-code for the proposed congestion control

TCP-vegas, are due to the reduction of *cwnd* depending on the delay. However, the proposed method only decides the *cwnd* switching point from rapid growth to additive increment and allows bandwidth competition using congestion control, as in the legacy TCP; thus, it minimizes problems in the legacy *cwnd* control that uses delay-based methods.

3.3 The CA Phase

In the TCP Reno, the congestion avoidance phase, CA (Fig. 1), starts when the congestion window exceeds the *ssthresh* or a packet loss is detected with triple duplicate ACKs. The start condition of the CA phase in the HE-TCP is the same as that of the TCP-reno. However, to manage the under-utilization problem during the CA phase in HBDP networks, we adopt an rapid increase of congestion window in the CA phase, called the RI phase. In the HE-TCP, if the *cwnd* is smaller than the estimated band-width BW, *ssthresh*, the CA phase switches to the RI phase to drastically increase the congestion window and utilize the larger available bandwidth. If the *cwnd* exceeds the estimated bandwidth BW during the RI phase, the phase switches back to the CA phase again and the *cwnd* increases linearly until a packet loss is detected. In the event of a packet loss, the congestion widow size is halved for fairness. Limiting

ssthresh to the estimated bandwidth, depending on past measurements, can limit the fairness in TCPs that use losses for congestion verification. Therefore, the variables related to the band-width estimation, BW_k and BW_{k-1}, are initialized to 1 to maintain the fairness.

3.4 The RI Phase

The RI phase is similar to the initial slow start phase of the HE-TCP; however, the start time of the RI is controlled by the estimated bandwidth BW of the connection, and yet it is very helpful to solve the fairness problem. As BW approaches to maximum available bandwidth, the increase rate of *cwnd* is rapidly decreased to make a soft landing avoiding slow start overshoot problems. The exit condition of the RI phase in the HE-TCP is the same as that of the SS phase. Figure 2 shows the pseudo-code of the proposed mechanism. Only when the congestion window is larger than that of the ordinary small BDP networks, the proposed mechanism is used; otherwise, the normal TCP algorithm is used to maintain backward compatibility with the legacy TCP.

4 Simulation Study

We performed simulations using the network simulator ns-2 [8] for a network with a bottleneck bandwidth of 800 Mbps and RTT values of 200 ms. The queue size at each bottleneck node is 50% of the bandwidth delay product of the bottleneck link. For convenience, the window size is measured in number of packets with a packet size of 1000 bytes. Drop tail queues are used in the routers.

Figure 3 shows the congestion window variations of the three TCP variants with a loss rate of 10^{-6}. In spite of severe packet loss events, the congestion window of the HE-TCP quickly recovers the link capacity and the performance

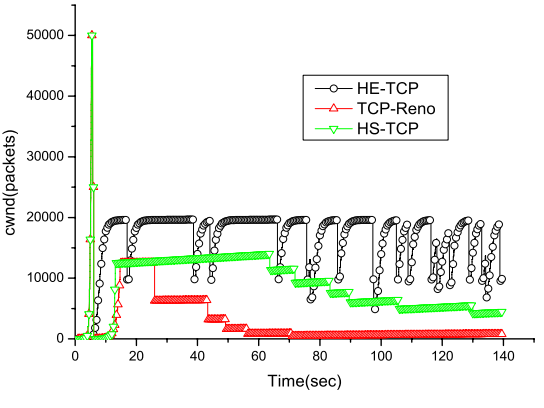


Fig. 3. Congestion window behavior of the HE-TCP, HS-TCP, and TCP Reno mechanisms with a loss rate of 10^{-6}

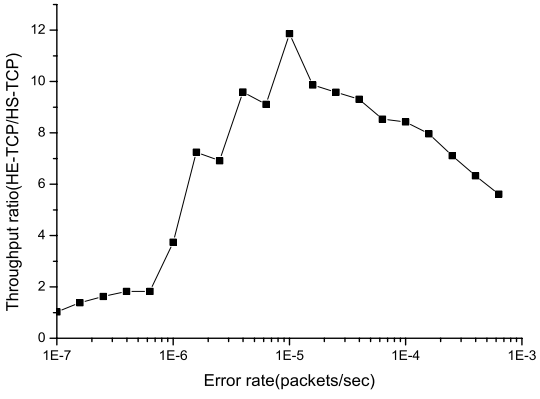


Fig. 4. Throughput ratio of the HE-TCP and HS-TCP mechanisms

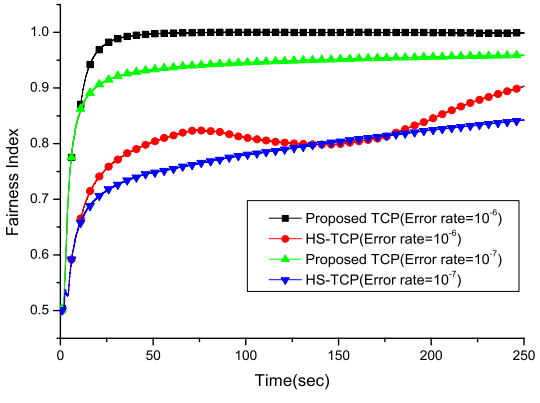


Fig. 5. Fairness comparison of the HE-TCP and HS-TCP mechanisms

is almost maintained at the available bandwidth. However, the performances of the HS-TCP and TCP-Reno are limited by the slow linear increase of the congestion window and the large loss probability.

The performance of the proposed HE-TCP and fairness has been compared with HS-TCP. The parameters for the HS-TCP are set at 31 for low windows, 83,000 for high windows, and 10^{-7} for high p.

In Fig. 4, the performance of the HS-TCP is compared with that of the HE-TCP in terms of packet loss rate. In both cases, the difference is minimal with an almost zero drop rate (less than 10^{-7}), since both use 100% of the available bandwidth. However, with a reasonably high loss rate, greater than 3×10^{-6}, the throughput of the HE-TCP is better than that of the HS-TCP.

Figure 5 shows the fairness comparison between the HE-TCP and HS-TCP when the packet loss rate is 10^{-6} and 10^{-7} with two flows. To show the fair share

distribution across the connections, we use Jain's Fairness Index as defined in [9]. When the throughputs of all flows are equal, the fairness index becomes 1. In both cases, HE-TCP's fairness index converges to 1 faster than the HS-TCP. Hence, the HE-TCP shows improved fairness even for cases where the performance difference is minimal due to a low packet loss rate.

5 Conclusion

We proposed a modified TCP congestion control mechanism, the HE-TCP, with an rapid increase phase during the congestion avoidance state without overshooting in HBDP networks. In addition, we evaluated the the performance of the proposed HE-TCP.

To evaluate the proposed HE-TCP, it was compared with the HS-TCP, which is the representative TCP for HBDP networks. The simulation results showed that the proposed mechanism improves fairness even when the performance difference is minimal due to a low loss rate. When the loss rate increases, the proposed method was proven to outperform other methods as well. The proposed HE-TCP can solve the TCP under-utilization problem in HBDP networks. The proposed algorithm can be easily implemented in sender TCPs. So, the proposed HE-TCP can be a promising transport protocol for large data transfer applications in HBDP networks.

References

1. IETF RFC 2581: TCP Congestion Control (1999)
2. IETF RFC 3649: HighSpeed TCP for large congestion windows (2003)
3. Kim, S., Park, S., Moon, J., Lee, H.: A low-crosstalk design of 1.25 Gbps optical triplexer module for FTTH systems. ETRI Journal, Vol. 28, no. 1. (2006) 9-16
4. Wang, R., Pau, G., Yamada, K., Sanadidi, M.Y., Gerla, M.: TCP startup performance in large bandwidth delay networks. Proceeding of IEEE INFOCOM 2004, Vol. 2. (2004) 796-805
5. Mascolo, S., Casetti, C., Gerla, M., Sanadidi, M.Y., Wang, R.: TCP Westwood: Bandwidth estimation for enhanced transport over wireless links. Proceeding of ACM/IEEE Mobi-Com 2001 (2001)
6. Xu, K., Tian, Y., Ansari, N.: TCP-Jersey for Wireless IP Communications. IEEE J. Select. Areas Comulications, Volume 22., (2004) 747-756
7. David D. C., Wenjia F.,: Explicit allocation of best effort packet delivery service. In: IEEE/ACM Trans. Networking, Vol. 6. (1998) 362-373
8. The network simulator ns-2. Available: http://www.isi.edu/nsnam/ns/
9. Jain, R.: The art of computer systems performance analysis: techniques for experimental design, measurement, simulation and modeling. New York, John Wiley & Sons (1991)

A Reliable Transmission Strategy in Unreliable Wireless Sensor Networks*

Zhendong Wu and Shanping Li

College of Computer Science, Zhejiang University, Hangzhou, China
zhendongwu@hotmail.com, shan@cs.zju.edu.cn

Abstract. Recent empirical and theoretical studies have shown that wireless links in low-power sensor networks are unreliable. Obviously, it is not a good choice for sensor networks that transmitting data unreliably, which will bring many problems to the networks, such as uncertainty, performance-decline, and so on. In this paper, a novel transmission strategy for reliable transmitting over unreliable links is proposed. Based on the observation that the broadcast nature of the wireless communications has the potential that allows all neighbors receive and process the packets simultaneously, we suggest that using a nodes-set, instead of one node, to receive and relay packets. The key idea is that receiving packets simultaneously and relaying packets competitively. Theoretical and experimental analysis demonstrates that the new strategy can provide reliable transmitting over unreliable links meanwhile improves the performance of networks significantly.

Keywords: Sensor Networks, Reliable Transmission, Nodes cooperation.

1 Introduction

Recent empirical and theoretical studies [1] ~ [4] have shown that wireless links in real sensor networks are unreliable. Particularly, in dense deployments, a large number of links in the sensor network (even higher than 50%) can be unreliable. The unreliability is derived from many factors such as multipath fading, unreliable sensor nodes and stochastic interference. It seems that the unreliability is one of the inherent properties of low-power wireless links.

Obviously, reliable transmission is hoped for communications of networks. In order to overcome the unreliable links in network layer, some efforts ([5] [6] [7]) have been put in defining metrics to characterize the energy efficiency of communication. Based on these metrics, the communication over unreliable links could be optimized. These methods all focus on the single link. However, the performance improvement through the optimization of single link is limited. For example, there are two links (link1 and link2) from node A to node B. In time1, link1 is better than link2. It means link1 has higher PRR (packet reception rate) than link2. But, in time2, link2 is better than link1. It can't be predicted which link would be better in one time. So when you want to send packet from node A to node B, you can't exactly choose the optimal link to send. It can

* This paper is supported by National Natural Science Foundation of China (No. 60473052).

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 376–384, 2007.
© Springer-Verlag Berlin Heidelberg 2007

be said that it is difficult to get satisfactory transmission performance over unreliable links just only using single link optimization.

In this paper, a reliable transmitting strategy, called *Set Transfer*, is proposed. It can realize reliable transmission over unreliable links meanwhile improves the energy efficiency of networks significantly. The key idea is that using multi nodes receive packets simultaneously and relay packets competitively. The energy efficiency of *Set Transfer* is analyzed theoretically and experimentally.

The rest of the paper is organized as follows. Section 2 is related works. Section 3 proposes the reliable transmitting strategy *Set Transfer*. In Section 4, the energy efficiency of *Set Transfer* is analyzed. Simulation and conclusion are in Section 5, 6.

2 Related Works

Although the ability of single node is limited, the ability of multi-nodes cooperation may be huge. There have been some researches on sensor nodes cooperation in physical layer [8] ~ [11]. Multipath fading is a fundamental phenomenon which makes reliable wireless transmission difficult in wireless communication. In order to reducing the effective error rate in a multipath fading channel, some diversity techniques such as multiple-antenna [8], relay channel model [11], are proposed. The main idea of diversity techniques is that sending one signal through different channels and combining signals from different channels at the terminal. Diversity can achieve signal gains. A. Sendonaris at el. [9] [10] proposed a new form of spatial diversity technique which achieve diversity gains through the cooperation of in-cell users (mobile - base station communication). Unfortunately, this approach sends many duplicate packets, decreasing the energy efficiency of networks. Diversity-techniques studies are now concentrated on physical layer.

Some unreliable-link models [2] [3] [4] have been proposed. It can help us further understand the realistic link layer of sensor networks. According to these models, there are three distinct reception regions in wireless sensor links: connected, transitional, and disconnected. The transitional region is often large in size, and generally has high-various reception rates. M. Zuniga at el. [4] derived expressions for reception rate as a function of distance for different settings.

Inspired by these researches, we study unreliable-links cooperation in network layer. Based on the unreliable-link models [2] [3] [4], a novel cooperation approach is proposed which can be said a new form of diversity technique.

3 The Reliable Transmitting Strategy: *Set Transfer*

3.1 The Preliminaries

In this part, two assumptions and a profile about *Set Transfer* are presented.

A. Assumptions

1. The low-power wireless sensor links are unreliable.
2. The broadcast mechanism such as CSMA-CA is used in link layer protocols.

CSMA-CA mechanism has been widely used in wireless communications, for example, IEEE standard 802.15.4 [12] and IEEE standard 802.11 [13]. It also works well in ad hoc and sensor networks because of its simple and efficiency.

B. The Profile of Set Transfer

The *Set Transfer* uses a node-set, instead of one node, to receive and relay packets. A node-set receives the packets simultaneously (called *Set Receiving*) and relays it competitively (called *Set Relaying*). As Fig.1 shows, the source node sends the packets to set A1, and set A1 uses *Set Receiving* to receive packets and uses *Set Relaying* to relay packets to set A2, and A2 receives and relays the packets to A3, and so on. Finally, through *n* hops, packets arrived at the set A*n*, which includes the destination node. Then any node in the set A*n* transmits the packets to the destination node through broadcasting.

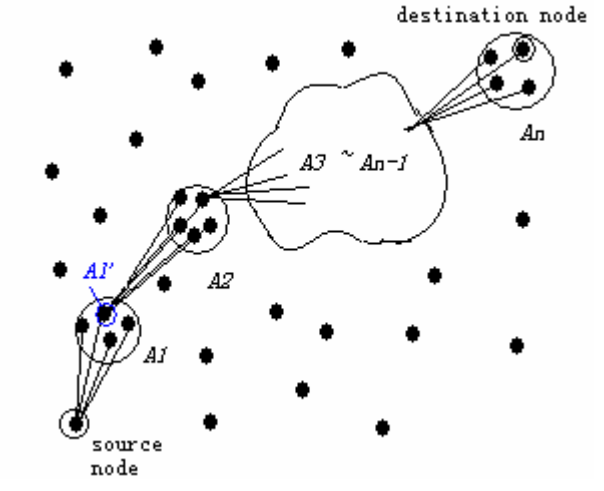


Fig. 1. Set Transfer

3.2 Set Receiving

According to the broadcast assumption, the wireless sensor networks have the potential that allows for multi-nodes receive and process packets simultaneously. Thus, instead of using one node to receive packets, we suggest using a nodes-set to receive packets. It is called *Set Receiving*. For example, in Fig.1, the packets sent from source node can be received by node A1' or by nodes-set A1. If nodes-set A1 is used, it means that each node in the set A1 receives packets independently, and if any node receives packets correctly, the receiving-process of set A1 is completed correctly. Here set A1 could be seen an integrity which receives and relays packets as one node. In part 2.3, the integrity of nodes-set will be demonstrated.

According to unreliable links assumption, the wireless sensor links are unreliable. The *Pr* is used (packet reception rate, $0 < Pr < 1$) to denote the reliability of links. The higher the P*r* is, the more reliable the link is. It can be assumed that each link in a

nodes-set is independent with its own *Pri* ($1 < i < n$). Then, the packet reception rate *Pset* of *Set Receiving* is:

$$Pset = 1 \quad (1 \quad Pr1) \quad 1 \quad Pr2 \quad 1 \quad Prn \tag{1}$$

$$\geq 1 \quad (1 \quad Prmin)^n \tag{2}$$

Where $Prmin = \min Pri$, $1 < i < n$. The n is the power of the nodes-set. From expression (1), the reliability of *Set Receiving* is characterized by a non-decreasing, strictly concave function. If *Prmin* is 0.7 and n = 6, then the packet error rate can be reduced from 10^{1} to 10^{4}. It can be said that the *Set Receiving* is reliable.

3.3 Set Relaying

In part 3.2, *Set Receiving* uses a nodes-set to receive packets, and some nodes in the set will correctly receive the same packets in the same time. However, in multi-hop transmission, a nodes-set how to relaying packets is still a question. If all nodes in nodes-set all relay packets to the next hop, the blast of packets replication will occur, just like flooding mechanism, which wastes too many energy and bandwidth of networks. In this part, *Set Relaying* scheme is introduced which guarantees that, in one nodes-set, only one node relays the packets.

Some works on wireless sensor networks [2] [3] [4] proposed some link models based on recent empirical studies [1] [2]. These models revealed that wireless sensor links have three distinct reception regions with different distance between two nodes: connected (normally 0m ~ 11m), transitional (normally 11m ~ 32m), and disconnected (> 32m). The transitional region is often large in size, and generally has high-various PRR (packet receive rate) which distributed in 10%~90%; the connected region is often small in size, and generally has high PRR. Fig.2 shows the analytical link loss model cited from M. Zuniga *et al* [4]. If consider the unreliable sensor devices and stochastic interference factors, the size of the transitional region will further increase.

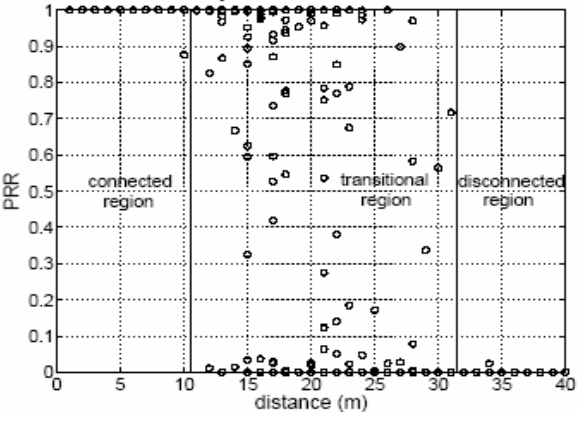


Fig. 2. Analytical PRR vs Distance

For the efficiency of transmission, data transmitting usually occurs in transitional region. According to broadcast assumption and above analyze, we get two observations as follows.

1. The broadcast nature of the wireless medium allows sensor nodes detecting their neighbor's transmission.
2. There are three distinct reception regions in wireless sensor links: connected, transitional, and disconnected.

Based on the two observations, the *Set Relaying* can be described as follows:

1. *One nodes-set should be lies in one connected region*.
2. W*hen a nodes-set receive packets, it stochastically select one node from the nodes that receive packets successfully to relay the packets with broadcast mechanism which informs other nodes the packets' relaying. (Competitive scheme)*

Rule1 guarantees that the relayed packets can be detected reliably. Rule2 guarantees that only one node in the set relays the packets to the next-hop nodes-set.

Combining *Set Receiving* and *Set Relaying*, we get an end-to-end transmission approach *Set Transfer*, which can realize reliable transmission over unreliable links.

4 Performance Analysis

Section 3 has demonstrated that *Set Transfer* can provide reliable transmission over unreliable links. This section will demonstrate that *Set Transfer* can improve the energy efficiency of wireless sensor networks, which is an important design issue.

To evaluate the energy efficiency of different transmission strategies, the metric Energy Per Packet (*EPP*) is used, which represents the average energy consumption for each delivered packet from the source node to the destination node. Without loss of generality, each packet length could be assumed equivalent. The *EPP* of a sensor network is comprised as follows:

$$EPP = E_{maintain}\,(E_{elec}) + E_{RF} + E_{process} \tag{3}$$

- $E_{maintain}\ (E_{elec})$: the amount of energy required to maintain the sensor nodes working, namely the energy consumption of circuits.
- E_{RF}: the amount of energy required to transmit packets in wireless environment.
- $E_{process}$: the amount of energy required to process packets in the sensor nodes.

In the design of modern processors, the power consumption $E_{process}$ can be made negligible compared to transmit powers [14]. So it can be deduced that

$$E_{RF} >> E_{process} \tag{4}$$

E_{elec} is the power that supports sensor nodes in working state which can not be ignored compared to transmit powers. On the other hand, *Set Transfer* can work with any MAC layer protocols as long as it accords with the broadcast assumption, such as IEEE 802.11, 802.15.4 and some new medium access control (MAC) protocols designed for WSN [15] [16]. It guarantees that the number of working nodes in two different transfer strategies is equal. According to above discussion, we have

$$E_{elec\text{-}set} = E_{elec\text{-}single} \tag{5}$$

Where $E_{elec\text{-}set}$ denotes the E_{elec} of *Set Transfer*, and $E_{elec\text{-}single}$ denotes the E_{elec} of *Single Link Transfer*.

It is important that *Set Transfer's* one nodes-set only use one node to relay packets (see *Set Relaying* 3.3), which guarantees that

$$E_{RF\text{-}set} = E_{RF\text{-}single} \tag{6}$$

First consider the links are reliable, according to (3) ~ (6), we get

$$\begin{aligned} EPP_set_reliable &= E_{elec\text{-}set} + E_{RF\text{-}set} + n\,E_{process} \\ &= E_{elec\text{-}single} + E_{RF\text{-}single} + n\,E_{process} \\ &\approx EPP_single_reliable \end{aligned} \tag{7}$$

Where n is the average number of nodes-set.

Second consider the links are unreliable, re-transmission is needed. As section 2 mentioned, the PRR of *Single Link Transfer* is *Pr*, and *Set Transfer* is *Pset*. Then through computing the expectation of transmissions, we get

$$EPP_set_unreliable = \frac{1}{pset}\, EPP_set_reliable \tag{8}$$

$$EPP_single_unreliable = \frac{1}{pr}\, EPP_single_reliable \tag{9}$$

According to equations (7) (8) (9), we have

$$EPP_set \approx \frac{p_r}{p_{set}}\, EPP_single \tag{10}$$

If the p_r is 0.7 and $p_{set} \approx 1$, then $EPP_set \approx 0.7 EPP_single$. Obviously, *Set Transfer* improved the energy efficiency of the networks with unreliable links.

5 Simulations and Comparisons

In this section, we first study the reliability of *Set Transfer* through numerical experiments, and second study the performance of *Set Transfer* through simulations.

In numerical experiments, each link is assigned a value of P*r* (0.1~0.6: 30%; 0.6~0.9: 70%), and then study the packet lost rate with scales in the number of nodes. It can be seen from Fig.3 that the curve about Set Receiving has the shape of log functions, and six or more nodes could have reliable one-hop transmission in this unreliable environment. In Fig.4, the number of working sensors in receiving set is fixed at 3 and the P*r* is changed. As Fig.4 shows, 3 nodes Set Receiving improve the transmission's reliability significantly. The curve of Set Receiving changes smoothly, and when the P*r* is 0.9~1, the 3 nodes Set Receiving nearly do not lost any packets.

In the simulations, 250 nodes are randomly deployed in a 200m×200m square area. The radio model ([4]) with three distinct reception regions: connected, transitional and disconnected is used. To simplify the simulation, each reception region has a random PRR (packet reception rate), such as connected (0~11m): 0.9~1, transitional

(11m~32m): 0.5~0.7 30%, 0.7~0.9 70%, transitional (32~40m): 0.1~0.6 70%, 0.6~0.9 30%, disconnected (>40m): 0. Four transmission mechanisms are compared: ETX transmission in [7]; Optimized GF transmission in [6]; Lazy loss detection transmission in [5]; *Set Transfer* transmission as proposed in this paper.

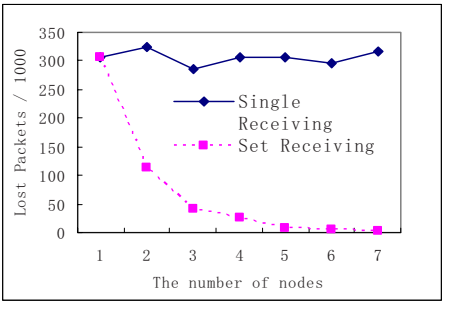


Fig. 3. Packet Lost Rate with different number of nodes

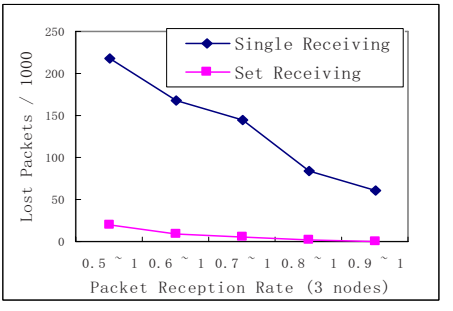


Fig. 4. Packet Lost Rate with different PRR

Fig.5 shows the energy efficiency of various transmission strategies. The average consumed energy of transmitting 1000 data packets successfully is observed. The end-to-end distance is described by hop counts. If just use original ETX metric to choose the next hop node, each hop distance will be too short to acquire good energy efficiency. So ETX strategy is slightly adjust like this: using ETX metric to choose the next hop node in a set of nodes. The power consumptions of each node in receiving, transmitting modes are 14.4mW and 36mW respectively [17]. From Fig.5, *Set Transfer* improves the energy efficiency of networks significantly. It is mainly because Set Transfer can provide reliable transmission which saves retransmission energy. In simulation, the results of each test of the ETX, Optimized GF and Lazy detection are changed acutely. For example, the E_ETX (6hops) changed from 806.3W S to 1498.6W S, although the average consumed energy is 932.9 W S. The link's unreliability results in this acutely changing, and single link methods all have this problem. However, the Set Transfer has no this problem. The E_Set (6hops) changed from

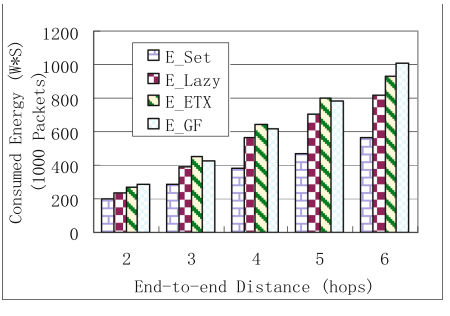


Fig. 5. The end-to-end consumed energy

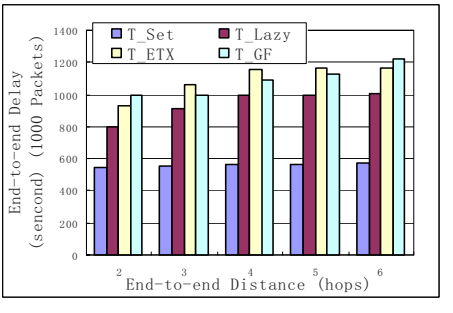


Fig. 6. The end-to-end delay (second)

553.008 W S to 576.776 W S, and the average consumed energy is 561.336 W S. This phenomenon further reflects that the links Set Transfer provided is reliable and the energy Set Transfer consumed is smooth.

Fig.6 shows the average end-to-end delay time of various transmission strategies. From Fig.6, the delay time of *Set Transfer* is obviously shorter than single link strategies. Through providing more reliable transmission than single link strategies, *Set Transfer* can effectively reduce the packet loss which will cause cascading effects, such as ACK delay, retransmitting, and so on.

6 Conclusion

The unreliable wireless links in real sensor networks will decrease the performance of networks seriously. In this paper, through studying the characteristics of the wireless links deeply, we proposed a multi-nodes cooperation strategy *Set Transfer*, which provides reliable transmission over unreliable links meanwhile improve the energy efficiency of networks. Further theoretical and experimental analysis gives us some insights on the improvements in energy efficiency that cooperation can bring to networks when the underlying links are unreliable.

References

1. Zhao, J., Govindan, R.: Understanding Packet Delivery Performance in Dense Wireless Sensor Networks. In ACM Sensys(2003)
2. Woo, A., Tong, T., Culler, D.: Taming the Underlying Issues for Reliable Multhop Routing in Sensor Networks. In ACM SenSys(2003)
3. Zhou, G., He, T., Krishnamurthy, S., Stankovic, J.: Impact of radio irregularity on wireless sensor networks. In MobiSys(2004)
4. Zuniga, M., Krishnamachari, B.: Analyzing the transitional region in low power wireless links, In *IEEE SECON*(2004)
5. Cao, Q., *et al*.: Efficiency Centric Communication Model for Wireless Sensor Networks. In *IEEE Infocom*(2006)
6. Seada, K., *et al*.: Energy efficient forwarding strategies for geographic routing in lossy wireless sensor networks, In *ACM Sensys*(2004)
7. Couto, D.D., *et al.*: A high-throughput path metric for multi-hop wireless routing. In *ACM Mobicom*(2003)
8. Alamouti, S.M.: A Simple Transmit Diversity Technique for Wireless Communications. IEEE Journal on Select Areas in Communications, Vol. 16, No. 8, 8(1998)
9. Sendonaris, A., Erkip, E., Aazhang, B.: User Cooperation Diversity-Part I: System Description. IEEE Transactions on Communications, Vol. 51, No. 11(2003)
10. Sendonaris, A., Erkip, E., Aazhang, B.: User Cooperation Diversity-Part II: Implementation Aspects and Performance Analysis. IEEE Transactions on Communications, Vol. 51, No. 11 (2003)
11. Nicholas Laneman, J., Tse, D.N.C., Wornell, G.W.: Cooperative Diversity in Wireless Networks: Efficient Protocols and Outage Behavior. IEEE Transactions on Information Theory, Vol. 50, No. 12(2004)
12. IEEE standard 802.15.4 2003.
13. IEEE 802.11. Wireless LAN medium access control (MAC) and physical layer (PHY) specifications(1999)

14. Stephany, R., *et al.*: A 200MHz 32b 0.5W CMOS RISC microprocessor. In Proc. 1998 IEEE Int. Solid-State Circuits Conf., 4(1998)238–239
15. Ye, W., Heidemann, J., Estrin, D.: Medium Access Control With Coordinated Adaptive Sleeping for Wireless Sensor Networks. IEEE/ACM Transactions on networking, Vol. 12, No.3,6(2004)
16. Ferrara, D., *et al.*: MACRO: An integrated MAC/routing protocol for geographic forwarding in wireless sensor networks. In IEEE Infocom (2005)1770-1781.
17. ASH Transceiver TR3000 Data Sheet [Online]. Available: http://www.rfm.com/

Genetic Algorithmic Topology Control for Two-Tiered Wireless Sensor Networks*

Donghwan Lee[1], Wonjun Lee[1], and Joongheon Kim[2]

[1] Department of Computer Science and Engineering, Korea University, Seoul, Korea
wlee@korea.ac.kr
[2] Digital Media Research Lab., LG Electronics, Seoul, Korea

Abstract. This paper proposes an optimized topology control scheme (GA-OTC) based on a genetic algorithm for clustering-based hierarchical wireless sensor networks (WSNs). By using a genetic algorithm, we can obtain optimal solutions to multiple objective functions according to the two criteria of both balanced energy consumption and minimized total energy consumption of cluster heads while the entire WSNs field is covered by clusters. Through performance evaluation studies, we show that *GA-OTC* achieves desirable properties.

Keywords: GA-OTC, WSNs, multiple objective functions.

1 Introduction

In recent years, many researches on wireless sensor networks (WSNs) have become one of the active research topics in wireless networking technologies [1]. Sensor nodes (SNs), the main components of WSNs, are deployed over the networked sensing fields, and perform specific tasks with the data processing, event sensing, and data communicating [1]. Due to their limited power source, energy consumption has become the most critical factor when designing WSNs protocols. Facing these challenges, several schemes to prolong the lifetime of WSNs, including clustering-based topology control, have been investigated [2]-[9]. By using an efficient clustering-based topology control algorithm, several advantages can be obtained, i.e., (1) improving network connectivity and capacity, (2) improving spatial reuse, and (3) mitigating the MAC-level medium contention [2]-[4]. The clustering-based WSNs can be classified into two types [3]: i.e., homogeneous one and heterogeneous one. In homogeneous clustering-based WSNs [5][6], by all SNs have same architectures, the algorithm for cluster head (CH) election is required because the CH must be chosen among SNs. In this structure, each SN has the function for clustering which can be a burden to the hardware constrained SNs. On the other hand, in non-homogeneous clustering-based WSNs [7]-[9], different types of nodes are used for CH-role. Therefore a node which has a functionality of the CH has more computing power than

* This work was supported by ITRC Project supervised by ITTA-2005 (C1090-0501-0019).

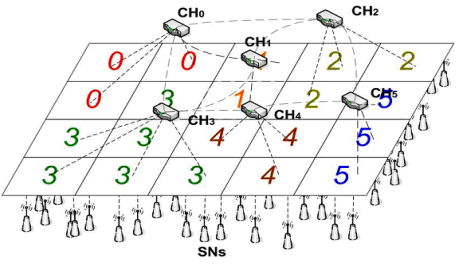
Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 385–392, 2007.
© Springer-Verlag Berlin Heidelberg 2007

SNs which is not the CHs. The non-homogeneous clustering-based WSNs, therefore, can achieve more energy-efficiency. In this paper, the non-homogeneous clustering-based WSNs is considered as a reference network model. In WSNs, balance energy consumption and minimized total energy consumption are required among CHs to prolong network lifetime [4][9]. To achieve the desired properties, the method of regulating cluster radii is effective as shown in [4][9]. As a method of regulating cluster radii, genetic algorithmic approach is used in this paper. By using a genetic algorithm, the most notable optimization method, we can achieve effective performance in regulating cluster radii to obtain optimal solutions.

The remainder of this paper is organized as follows. Section 2 shows related work. Section 3 describes our proposed algorithm. Section 4 shows performance results. Finally, section 5 concludes this paper and shows future work direction.

2 Related Work

Clustering-based WSNs can be classified into two types [3]: i.e., homogeneous clustering-based WSNs and hierarchical clustering-based WSNs. In homogeneous clustering-based WSNs, the algorithm for cluster head election is required because the CH must be chosen among SNs. Therefore, SNs must have the functionality for clustering and controlling their clusters. This functionality of SNs must be a burden to the hardware constrained SNs. On the other hand, in hierarchical clustering-based WSNs [3][8][9], different types of SNs are used. A mobile node with the functionality of CH has more computing power than a SN which is not the CH. The hierarchical clustering-based WSNs, therefore, can achieve more energy-efficiency. There are numerous topology control algorithms for hierarchical clustering-based WSNs [7]-[9]. The proposed algorithm in [7] uses three kinds of SNs for energy-efficiency. However the proposed scheme in [7] must maintain global information. It can be a burden to CHs. Therefore, decentralized scheme is required in resource-constrained WSNs. The proposed scheme in [8] has a two-tiered WSNs architecture. In the upper tier, there are CHs and in the lower tier, typical SNs. In [8], however, the radii of clusters are fixed. In this system, the performance of scheme depends on each radius of the cluster. That is to say, it can be a waste of energy when a cluster has a larger radius than required. On the other hand, if a cluster has a smaller radius than required, it cannot guarantee perfect coverage, means there is no area uncontrolled by a CH in the networked sensing field. Therefore, maintaining the optimal cluster radius is important. However, it is impossible for [8], which has a fixed radius system, to regulate the cluster radius. On the other hand, our previous work, i.e., LLC [9] can regulate the cluster radius for energy efficiency in WSNs. However in LLC, there exists a maximum radius problem which can lead energy-inefficiency in some network topology due to the assumption of the circularity of cluster radius. With genetic algorithmic approach, the proposed scheme in this paper, we can overcome maximum radius problem and achieve better performance.



0	0	1	1	1	1	1
0	0	0	1	1	1	1
0	0	1	1	1	1	3
0	2	2	1	1	3	3
2	2	2	2	3	3	3
2	2	2	2	3	3	3
2	2	2	2	2	3	3

0	0	0	0	1	1	1
0	0	0	1	1	1	1
0	0	0	3	1	1	1
0	0	2	3	3	1	1
2	2	2	2	3	3	1
2	2	2	2	3	3	3
2	2	2	2	3	3	3

Fig. 1. Examples of the grid which is assigned to 3 CHs and initial population (4 CHs)

3 Genetic Algorithmic Optimized Topology Control (*GA-OTC*)

The aim of *GA-OTC* is is to propose an algorithm which covers all SNs and considers energy consumption by a genetic algorithm. *GA-OTC* balances and minimizes energy consumption of CHs. A sensor field which needs covering with CHs is segmented by a grid form. SNs are located in any point of each grid cell. Once a grid cell is assigned to CH, CH communicates with SNs located in a grid cell as shown in Fig. 1. The group of grid cells assigned to a CH is defined as a clustering area.

3.1 Energy Model of a Wireless Device

To describe *GA-OTC*, the energy model of a CH can be formulated as follows.

$$E = E_{communicating} + E_{processing} \tag{1}$$

Energy consumption on data communicating. Assuming an $1/d^2$ path loss, where d is the distance between a CH and an SN, the energy consumed when transmitting and receiving data are Eq. (2) and Eq. (3).

$$E_{tx} = (\alpha_t + \alpha_{amp} \cdot d^2) \cdot n \tag{2}$$

$$E_{rx} = \alpha_r \cdot n \tag{3}$$

α_t, α_{amp}, and α_r are the energy/bit consumed by transmitter, transmitter op-amp, and receiver. The total energy consumption on data communicating is composed by E_{tx} and E_{rx} as shown in Eq. (4).

$$E_{communicating} = E_{tx} + E_{rx} \tag{4}$$

Energy consumption on data processing. In general, the energy consumption on data processing is significantly smaller than the energy consumption on data communication. Therefore the energy consumption can be ignored.

3.2 Problem Formulation

We define a clustering set that a sensor j belongs to the clustering set of CH, i.e., C_i, for all sensors. Using an energy model for a CH denoted in section 3.2, the communication energy (CE) between C_i and j, denoted as CE_{j,C_i}, is calculated as follows.

$$CE_{j,C_i} = E_{tx} + E_{rx} = (\alpha_t + \alpha_{amp} \cdot d^2_{j \to C_i}) \cdot n + \alpha_r \cdot n \tag{5}$$

where $d_{j \to C_i}$ is the distance between j and C_i. Energy consumption of each CH by every SN of its clustering set is calculated by adding up energy consumption of all sensors in each clustering set, denoted as EC_{C_i}.

$$EC_{C_i} = \sum_{j \in C_i} CE_{j,C_i} \tag{6}$$

Thus, the variance and average of energy consumption of CHs can be shown as Eq. (7) and Eq. (8).

$$\mathscr{V}(C) = \frac{1}{r} \sum_{i=0}^{r} (EC_{C_i} - \mathscr{A}(C)) \tag{7}$$

$$\mathscr{A}(C) = \frac{1}{r} \sum_{i=0}^{r} EC_{C_i} \tag{8}$$

where r is the number of CHs. C is a vector that consists of set variables $(C_1, C_2, \ldots\ldots, C_r)$. $\mathscr{V}$ and $\mathscr{A}$ are functions of C which denote variance and average of energy consumption. Let us denote all possible solution space of C to Ω, now we can construct objective functions for minimizing both variance and average of energy consumption of CHs.

$$\arg\min_{C \in \Omega} \mathscr{V}(C) \tag{9}$$

$$\arg\min_{C \in \Omega} \mathscr{A}(C) \tag{10}$$

3.3 Topology Control Procedure

Initial population. Each P chromosome is set as follows. To assign grid cells to each CH continuously, we use probabilistic iteration based in grassfire concept [12]. In this paper, we use propagation probability of grid cells instead of threshold of grassfire concept. Probabilistic propagation starts in every initial cell, where CH is located, and continues until all the cells are assigned to CHs as shown in Fig. 1.

Fitness computation. The fitness computation is designed with genetic algorithm. The proposed scheme in this paper uses simple additive weighting method [13]. The objective function of multi-objective problem consists of the sum of weighted objective functions. The fitness computation function by using simple additive weighting method is shown in Eq. (13), which is derived by Eq. (11) and Eq. (12).

$$f_v = \frac{1}{\sigma} = \frac{1}{\sqrt{\mathscr{V}}} = \frac{1}{\sqrt{\frac{1}{r}\sum_{i=0}^{r}(EC_{C_i} - \mathscr{A}(C))}} \tag{11}$$

$$f_a = \frac{1}{\mathscr{A}} = \frac{1}{\sqrt{\frac{1}{r}\sum_{i=0}^{r}(EC_{C_i})}} \tag{12}$$

$$F = w_v \cdot f_v + w_a \cdot f_a \tag{13}$$

where w_v and w_a are the weight values for determining which to be optimized. The summation of weight values has to be one. We use simple additive method described above instead of the pareto method, another well-known scheme which determines optimal solution. The pareto method uses dominent-nondominent relationships between solutions. However, this method is not appropriate to use to determine which objective to optimize.

Selection. In this paper, we use roulette wheel selection, a common technique that implements the proportional selection strategy. In roulette wheel selection scheme which resembles survival of the fittest in nature, the chance to be selected for the reproduction of a chromosome is determined by its ratio of fitness value.

Crossover. The crossover process is a probabilistic process which exchanges information between two parent chromosomes to generate two child chromosomes. We use the crossover process which child chromosomes inherit one section from each of parent chromosomes partitioned by random partitioning. This crossover process might violate basic assignment principle defined as the continuous cell assignment, which is assumed in initial population. However, if the fitness of such a chromosome is low, it is dismissed in selection process. In this way, crossover process helps algorithm converge into optimal solution as it guarantees the variety of populations.

Mutation. The mutation process in this paper is computed by exchanging of boundary cells assigned to each CHs.

Optimal solution and terminal criterion. The process of fitness computation, selection, crossover, and mutation is executed for a maximum number of iterations. We use an elitist strategy to preserve the best parent chromosome. Therefore a parent chromosome with the highest fitness is always copied into the next generation.

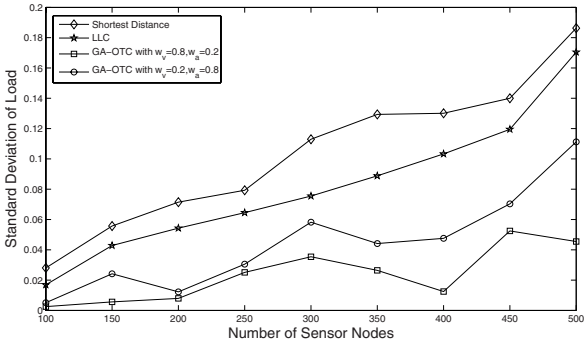


Fig. 2. Standard deviation of load with varying density (5 CHs)

4 Performance Evaluation

The performance evaluation of *GA-OTC* algorithm is shown in the aspect of (1) balanced energy consumption among CHs and (2) minimized total energy consumption. We plot the results obtained from our simulation in Fig. 2 and 3. To evaluate our algorithm, we compare the results of the proposed scheme with shortest distance based algorithm and Low-Energy Localized Clustering (LLC) algorithm [9]. In shortest distance based algorithm a CH includes an SN in its cluster if the distance between them is the minimum. LLC algorithm is the algorithm which regulates the cluster radius for energy efficiency.

4.1 Balanced Energy Consumption

The graph shown in Fig. 2. indicates that *GA-OTC* algorithm achieves clearly lower standard deviation with increase in density. Results demonstrate that our approach *GA-OTC* with w_v=0.8, w_a=0.2 and *GA-OTC* with w_v=0.2, w_a=0.8 is better than shortest distance based algorithm and LLC algorithm. In LLC, it causes ineffective radius problem compared to our algorithm. Because LLC selects maximum radius to cover all sensor fields and its' clustering radii of CHs are fixed to circle.

4.2 Minimized Total Energy Consumption

Average energy consumption on a CH will be minimum when shortest distance based algorithm is used but the load will not be balanced. We try to minimize the average energy consumption as good as shortest distance based algorithm with increase in CHs. As shown in Fig. 3, *GA-OTC* with w_v=0.2, w_a=0.8 performs better than LLC and *GA-OTC* with w_v=0.8, w_a=0.2. *GA-OTC* with w_v=0.8, w_a=0.2 performs worse than LLC and *GA-OTC* with w_v=0.2, w_a=0.8 since

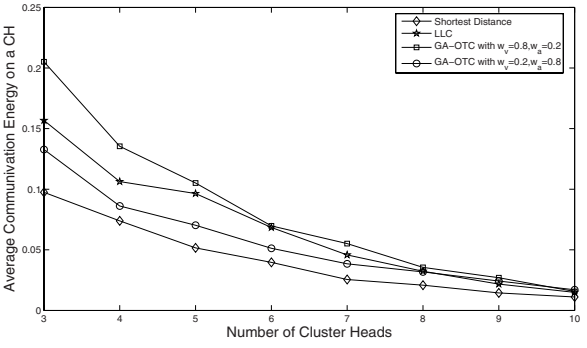


Fig. 3. Average energy consumption on a CH (300 SNs)

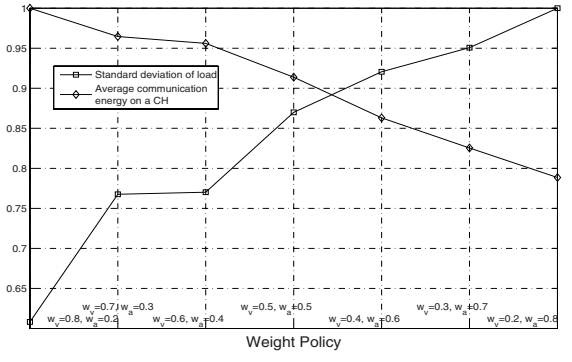


Fig. 4. Standard deviation of load and average communication energy on a CH with varying weight policy (5 CHs, 300 SNs, normalized with respect to the highest value)

weight factors of it is focused on minimizing the variance rather than average of energy consumption. Two demonstrated results are based on the aspect of two weight policies. When w_v=0.2, w_a=0.8, weight of minimized total energy consumption is greater than weight of balanced energy consumption, results indicate that the average is better than w_v=0.2, w_a=0.8 and vice versa. The graph shown in Fig. 4. clearly describes the tendency of the standard deviation and average communication energy reflected by various weight policies. The values on the counter domain are normalized with respect to the highest value. This weight policy can be implemented to optimize various environments of WSNs by changing weight factors. For example, in some WSNs, the survival of every node can be the most important mission. In these cases, we can achieve longer expected life time of SN that first dies by giving more weight value to weight factor w_v. In case that total life time of WSNs is more important, the bigger value is assigned to weight factor w_a rather than w_v.

5 Conclusion and Future Work

A genetic algorithmic topology control scheme for WSNs, named *GA-OTC*, is proposed to balance and minimize energy consumption. A genetic algorithm is used to search for the optimal clustering set of SNs. In addition, in order to solve a multi-objective problem, weight factors are used for two objective functions, i.e., balanced energy consumption and minimized total energy consumption. By using weight factors, *GA-OTC* can determine the objectives to optimize. Our novel scheme shows better performance on balancing and minimizing energy consumption than shortest distance based algorithm and LLC algorithm. As for future work, we consider applying another meta-heuristic such as simulated annealing and tabu search to the proposed optimized topology control scheme.

References

1. Akyildiz, I., Su, W., Sankarasubramaniam, Y., Cayirci, E.: Wireless Sensor Networks: A Survey. Computer Networks, Vol. 38. Elsevier (2002) 393-422
2. Mhatre, V., Rosenberg, C.: Design Guidelines for Wireless Sensor Networks: Communication, Clustering and Aggregation. Ad Hoc Networks, Vol. 2. Elsevier (2004) 45-63
3. Mhatre, V., Rosenberg, C.: Homogeneous vs Heterogeneous Clustered Sensor Networks: A Comparative Study. In Proc. IEEE Int'l. Conference on Communication (2004)
4. Kim, J.,Choi, J., Lee, W.: Energy-Aware Distributed Topology Control for Coverage-Time Optimization in Clustering-Based Heterogeneous Sensor Networks. In Proc. IEEE Vehicular Technology Conference (2006)
5. Heinzelman, W., Chandrakasan, A., Balakrishnan, H.: An Application-Specific Protocol Architecture for Wireless Microsensor Networks. Trans. Wireless Comm. Vol. 1, IEEE (2002) 660-670
6. Younis, O., Fahmy, S.: HEED: A Hybrid, Energy-Efficient, Distributed Clustering Approach for Ad Hoc Sensor Networks. Trans. Mobile Computing. Vol. 3. IEEE (2004) 366-379
7. Gupta, G., Younis, M.: Load-Balanced Clustering of Wireless Sensor Networks. In Proc. IEEE Int'l. Conference on Communication (2003)
8. Pan, J., Hou, Y., Cai, L., Shi, Y., Shen, S.: Topology Control for Wireless Sensor Networks. In Proc. ACM Int'l Conference on Mobile Computing and Networking (2003)
9. Kim, J., Kim, E., Kim, S., Kim, D., Lee, W.: Low-Energy Localized Clustering: An Adaptive Cluster Radius Configuration Scheme for Topology Control in Wireless Sensor Networks. In Proc. In Proc. IEEE Vehicular Technology Conference (2005)
10. Min, R., Bhardwaj, M., Cho, S., Sinha, A., Shih, E., Wang, A., Chandrakasan, A.: An Architecture for a Power-Aware Distributed Microsensor Node. In Proc. IEEE Workshop on Signal Processing Systems (2000)
11. Davis, L.: Handbook of Genetic Algorithms. Van Nostrand Reinhold, NY (1991)
12. Blum, H.: Biological Shape and Visual Science. Journal of Theoretical Biology, Part I, Vol. 38. Elsevier (1973) 205-287
13. Hwang, C., Yoon, K.: Multiple Attribute Decision Making: Methods and Applications. Lecture Notes in Economics and Mathematical Systems, Springer-Verlag, Berlin (1981)

A Delay Sensitive Feedback Control Data Aggregation Approach in Wireless Sensor Network*

Peng Shao-liang, Li Shan-shan, Peng Yu-xing, Zhu Pei-dong, and Xiao Nong

School of Computer, National University of Defense Technology, Changsha, China
{pengshaoliang,ssli,pyx,pdz,xn}@nudt.edu.cn

Abstract. In wireless sensor networks (WSN) real-time data delivery is a challenging issue, especially when bursty events happen and many emergent packets appear. The cluster-based data aggregation has emerged as a basic and general technique in sensor networks. However, there is a trade-off issue between aggregation waiting (AW) delay and aggregation accuracy. Therefore, we propose a Delay Sensitive Feedback Control (DSFC) model to solve the trade-off issue. Our goal is to decrease the AW delay on the condition of meeting the desired aggregation accuracy for emergent packets when emergent event occurs. Simulation results verify the performance of DSFC, and demonstrate their attraction.

Keywords: Wireless Sensor Networks, Real-Time, Feedback Control.

1 Introduction

Wireless sensor networks can be used for many mission-critical applications such as target tracking in battlefields, nuclear leak alert, habitat temperature monitoring in forests and hazardous environment exploration [1]. Although these applications for sensor network remain diverse, one commonality they all share is the need for a timely delivery of sensory data. Without a real-time communication service, applications cannot react to the changes of the environment quickly enough to be effective. Therefore, in this paper we propose a comprehensive mechanism to meet real-time constraint of packets generated in crisis states.

Data aggregation has emerged as a basic tenet in sensor networks. The key idea is to combine the data coming from different sensors – eliminating redundancy, minimizing the number of transmissions and thus saving energy. Data aggregation can also provide a rich, multi–dimensional view of the environment being monitored [2]. However, nowadays many works on data aggregation are based on cluster topology [3]. All vicinal clustering members aggregate data at cluster head. Clustering–based aggregation has high energy efficiency due to intra-cluster data redundancy.

Although clustering benefits from data aggregation, which may not be the only reason for using clusters. Clustering makes the system scalable. Instead of having a

* Supported by the National Natural Science Foundation of China (Grant No. 60433040), the National Basic Research Program of China under Grant No 2002CB312105 (973), and the National Basic Research Program of China under Grant No 2006CB303000 (973).

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 393–400, 2007.
© Springer-Verlag Berlin Heidelberg 2007

centralized control over thousands of nodes, or having a distributed protocol that operates over thousands of nodes, it is better to organize nodes into smaller clusters, and assign the responsibility of MAC and routing in each cluster to a single cluster head node. So we propose a bursty real-time mechanism which is cluster-based.

However, there are problems for cluster-based data aggregation: higher latency and less accuracy, especially when a lot of emergent packets are generated within a short period. We try to find the optimal point for the trade-off between aggregation delay and accuracy. That is to decrease delay as much as possible within the accuracy bound. In this paper, a distinguished feedback control model is proposed to adaptively aggregate data in a delay sensitive manner while not damage the gathering accuracy. Our goal is to decrease the AW delay on the condition of meeting the desired aggregation accuracy for emergent packets when emergent event occurs.

The rest of this paper is organized as follows. Section 2 briefly describes some related work. In Section 3, we discuss the issues of delay/accuracy trade-off and propose a Delay Sensitive Feedback Control (DSFC) model to solve this problem. Several simulations described in Section 4 verify the obtained theoretical results. The paper ends with conclusion and directions for the future in Section 5.

2 Related Work

Interactions between sensor networks and the physical world require real-time services to play an important role. Several real-time methods have been proposed for sensor and ad hoc networks. SPEED [4] protocol is designed to provide soft end-to-end deadline guarantees for real-time packets in sensor networks. MM-SPEED [5] extends SPEED to support different delivery velocities and levels of reliability.

However, previous approaches are efficient only when most traffic requires low reporting rates, which is not the case for many sensor networks especially with bursty events. Also, it is not adaptive to dynamics of data collection. Besides the real-time protocol design, several researches approaches focus on the delay analysis for data aggregation. In [6], Athanassios et al. only focus on data aggregation: an energy–accuracy trade-off, but they can not mention aggregation delay. However, all the prior work has been mainly concerned with the benefits of aggregation and has only superficially acknowledged the delay/accuracy trade-off present.

A hierarchical network organization will not only facilitate data aggregation, but will facilitate the formation of uniform, robust backbones, and give an opportunity for nodes to sleep while not sensing or transmitting. Many clustering protocols have been investigated [7]. However, prior work has rarely been concerned with the delay of aggregation. In order to build an E2E quick path for packets of emergent events, we will put forward a bursty real-time mechanism which is ground on clustering.

The E2E delay of emergent packet is crucial when bursty events happen. The delay is mainly composed of (i) intra-cluster data aggregation (ii) inter-cluster routing. We will take into account the delay of intra-cluster data aggregation because the latter is routing issue, which has been investigated in many literatures.

3 Delay Sensitive Feedback Control in Intra-cluster Data Aggregation

In this Section, we first briefly describe priority based packets classification and scheduling method because latter some mechanisms are based on it. Then, in Section 3.2, we analyze the delay due to data aggregation. Finally, in Section 3.3 we discuss the issues of delay/accuracy trade-off and propose a Delay Sensitive Feedback Control (DSFC) model to solve it.

3.1 Priority Based Packets Classification and Scheduling

Considering the characteristic of bursty events in wireless sensor networks, we will borrow the idea of QoS from Internet which has been proposed using two different methods –the Differentiated services (DiffServ) and the Integrated services (IntServ). We can classify the packets into two classes: emergent packet and regular packet, according to the semantic characteristics of packets. Packets classification is implemented by setting a bit in the TOS field. Cluster head assigns the TDMA slots for each node in the region according to the application of every source node, creates a TDMA schedule for its nodes and broadcasts it every round aggregation. TDMA also can be modified to allocated more and earlier slots to emergent packets.

3.2 Delay Due to Data Aggregation

Although data aggregation results in fewer transmissions, there is a tradeoff - potentially greater delay because data from different sources may have to be held back at a cluster head in order to be aggregated with data coming from other sources. Fewer transmissions should be achieved at the cost of delay, especially when bursty events happen and many emergent packets appear.

Generally there are three factors affecting the delay of intra-cluster data aggregation: **(i)Aggregation Routing (AR)**, (ii) **Aggregation Function (AF),** (iii) **Aggregation Waiting (AW)**. Earlier data from different sources may have to wait at a cluster head in order to be aggregated with subsequent data coming from other sources. The problem of AW delay will be aggravated especially when emergent events occur. The delay of AW is the most important bottleneck of data aggregation. In this paper, we will place emphasis on the issue of AW delay and consider how to deal with the accuracy/aggregation tradeoff.

3.3 Trade-Off Between AW Delay and Aggregation Accuracy

3.3.1 Problem Statement

In order to addressing the delay bottleneck of aggregation waiting, we must take into account the loss of aggregation accuracy. However, many nodes in the same cluster are located at the close geographical area, their gatherings are correlated and accordingly some accuracy loss is acceptable. We can consider to aggregate partial data to reduce the AW delay on the condition of meeting the desired aggregation accuracy. Therefore, there is a trade-off issue between AW delay and aggregation accuracy.

The trade-off issue can be regarded as a problem of Multiobjective Optimization. AW delay and aggregation accuracy are two main objects but they come into conflict with each other. Consequently, we use Optimization Method to solve this trade-off problem. Optimization Method is one of the most effective and prompt means to deal with the trade-off issue. The AW delay of emergent packet is crucial when bursty events happen, so we will aim at the goal of reducing AW delay. In detail, we will put forward a Delay Sensitive Feedback Control (DSFC) model [8], which can decrease the AW delay on the condition of meeting the desired aggregation accuracy.

3.3.2 Delay Sensitive Feedback Control Model

We assume an ideal TDMA strategy used as our intra-cluster MAC protocol. Every source node will ask for sending packets to cluster head at a certain reporting rate before every round aggregation. Cluster head can execute a centric control every round according to the request of every source node, and assign the TDMA slots for each intra-cluster node. Each cluster head then use a Priority based Packet Scheduling (discussed in section 3.2) strategy based on TDMA to allocate time slots. Thus, emergent packets can obtain more and earlier time slots. Each source node sends data to its cluster head according to the specified TDMA schedule.

In this section, we present a Delay Sensitive Feedback Control (DSFC) model to solve the problems on delay/accuracy trade-off. We can use the DSFC model to adaptively aggregate partial data to reduce the AW delay on the condition of meeting the desired aggregation accuracy. We begin with the following definitions.

Definition1. Aggregation Scale (AS): The ratio of emergent packets arrived at cluster head to total produced emergent packets in a round of data aggregation. The calculation of AS is as followed:

$$AS(i) = \frac{N_{arrived}(i)}{N_{total}(i)} \qquad \text{i is round number, i=1,2,3,4….} \tag{1}$$

$N_{total}(i)$ is the total number of produced emergent packets in the ith round aggregation. $N_{arrived}(i)$ is the factual number of emergent packets that arrive at cluster head and need to be aggregated in the ith round. Obviously, $AS \le 1$. $N_{total}(i)$ can be calculated by cluster head because all source nodes must request to send packets at a certain reporting rate when the source nodes sense an emergent event. Cluster head can execute a centric control every round according to the request of every source node. $AS(i)$ is used to control the aggregation accuracy and AW delay. Figure 1 shows the feedback controller, and it can control the parameter of **Aggregation Function (AF)**. The concrete definition of AF is specific to the application of WSN, and can be expressed as $AF(i \quad AS(i))$. For example, $AF(i,1)$ means that cluster head carry out full data aggregation until all data of source node reach cluster head. The aggregation accuracy of full data aggregation is the highest but the AW delay is the longest.

Definition2. Error (E): The error of partial data aggregation. It can be calculated as followed:

$$E(i) = \left| AF(i,1) - AF(i,\ AS(i)\ ) \right| \tag{2}$$

E(i) is the difference between partial aggregation and full aggregation which means of data aggregation accuracy at i round. It can be used to adjust the AS (i+1) at (i+1) round aggregation. The Closed-loop Feedback Control model is illustrated as equation (3):

$$AS(i+1) = AS(i) + \alpha(\ E(i) \quad E_0\ ) \tag{3}$$

α is the controller coefficient and can be well approached by some off-line experiments. E_0 is the minimal error that is related to specific applications of WSN.

The DSFC architecture is shown in figure 1:

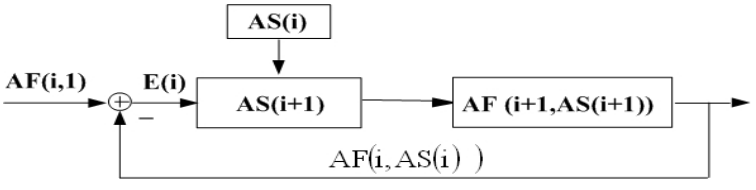


Fig. 1. The Architecture of DSFC Model

Our goal is transmit the emergent packets as fast as possible when emergent events occur. Emergent packets are partially aggregated in the proportion of $AS(i+1)$. The DSFC can adaptively decrease the AW delay at the cost of accuracy. We can continuously reduce the AS, namely $AS(i+1)$, as long as the aggregation accuracy meets the requirement of applications. The DSFC system tends to be steady until $E(i)$ is close to E_0.

The total time of the ith round data aggregation $AT(i)$ is composed of three parts, which is illustrated as equation (4):

$$AT(i) = S + C + \sum_{j=1}^{N_{arrived}} T_j \tag{4}$$

S: The synchronization time of intra-cluster nodes

C: The intercommunion time that cluster head executes a centric control and harmonization before the source data arrive at cluster head. The course includes source node requesting phase and cluster head assigning time slots phase.

T_j: The time slot of packet j. T_j is assigned by cluster head.

3.3.3 Stable State

Under the adjustment of DSFC, the system eventually converges at a stable state. In order to further decrease the AW delay and full aggregation overhead, full aggregation

is unnecessary in every aggregation round. Partial data aggregation can be executed with a fixed AS for a period of time.

4 Performance Evaluation

4.1 Simulation Settings

We simulate DSFC and EPCR using GloMoSim [9], a scalable discrete-event simulator developed by UCLA. This software provides a high fidelity simulation for wireless communication with detailed propagation, radio and MAC layers. Table 1 illustrates the detailed configuration for our simulation. The communication parameters are chosen in reference to the Berkeley mote specification.

Table 1. Simulation settings

Routing	Shortest Path
MAC Layer	TSMA (Intra-Cluster), 802.11 (Inter-Cluster)
Radio Layer	RADIO-ACCNOISE
Propagation Model	TWO-RAY
Bandwidth	200Kb/s
Payload Size	32 Byte
TERRAIN	(300m, 300m)
Node Number	300
Radio Range	30~50m
Clustering Method	HEED [10]
Cluster Radius	25 m

We use a simulation model based on GlomoSim and conduct extensive experiments to evaluate its performance. To make dense network with a strong connectivity and have more choice to construct cluster and paths, we deploy 300 nodes with a radio range of 30m to 50m. A channel error rate of 10% is uniformly distributed around the network (With max ARQ of 802.11).Our works is not based on a certain routing protocol, so we select the Shortest Path routing protocol and long-hop standpoint. In our evaluation, we present the following set of results: (1) DSFC model to adaptively control Aggregation Scale, (2) AW delay and error under different scenarios, (3) aggregation communication energy consumption. All experiments are repeated 16 times with different random seeds.

4.2 Performance of DSFC

First, we examine the performance of intra-cluster DSFC model. In figure 2 we present the DSFC control model to adaptively aggregate partial emergent packets to reduce the AW delay at the three different scenarios. DSFC can effectively and directly control the AF to converge at a certain value according to E(i) and the data correlation. Figure 3 illustrates the AW delay of different scenarios. The emergent

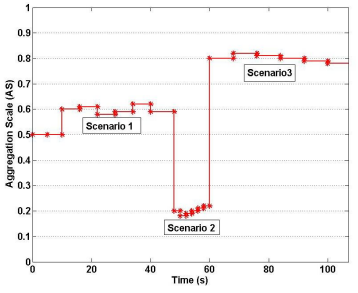


Fig. 2. Control process of DSFC model

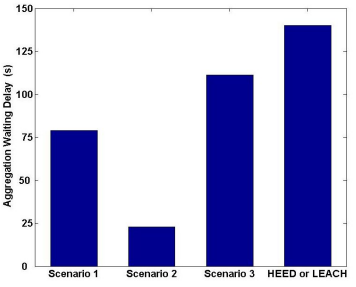


Fig. 3. AW delay of Different Scenarios

packets data aggregation based clustering generally aggregate all intra-cluster data. So, cluster head has to wait for a long time in order to be aggregated all data coming from other sources. We can see clearly in Figure 4 the AF delay of HEED and LEACH [10] is longer than that in scenario1, 2, and 3.

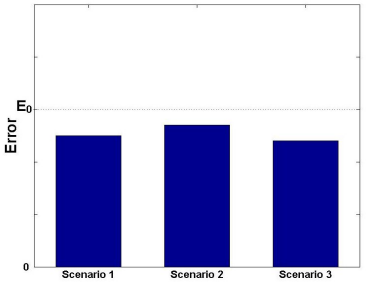


Fig. 4. Error of Different Scenarios

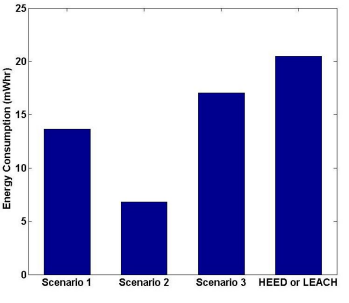


Fig. 5. Intra-cluster Energy Consumption

Figure 4 shows the error of partial data aggregation under different scenarios. All the errors of 3 scenarios are all less than the minimal error, E_0. Figure 3 and Figure 4 illuminate DSFC can effectively reduce the AW delay on the condition of meeting the desired aggregation accuracy.

Under energy constraints, it is vital for sensor nodes to minimize energy consumption in radio communication to extend the lifetime of sensor networks. From the results shown in Figure 5, we argue that partial aggregation based DSFC tends to reduce the transmission of emergent packets in cluster, thus saving energy. Figure 5 shows fewer intra-cluster transmissions in scenario1, 2, and 3 consume less energy than HEED and LEACH [10], which always carry out the traditional full aggregation.

5 Conclusion

We have proposed a DSFC model to solve the trade-off issue between AW delay and aggregation accuracy. Simulations demonstrate DSFC model can adaptively aggregate

partial data to reduce the AW delay on the condition of meeting the desired aggregation accuracy. The costs of energy consumption in DSFC are also very low.

References

1. LEWIS, F.L.: Wireless Sensor Networks. Smart Environments:Technologies, Protocol and Applications(2004)
2. Enachescu, M., Goel, A., Govindan, R., Motwani, R.: Scale-Free Aggregation in Sensor Networks, Theoretical Computer Science: Special Issue on Algorithmic Aspects of Wireless Sensor Networks, Vol. 344, No. 1, 11(2005)15-29
3. Younis, O., Fahmy, S.: Distributed Clustering in Ad-hoc Sensor Networks: A Hybrid, Energy-Efficient Approach. Proceedings of IEEE INFOCOM. Hong Kong(2004)
4. He, T., Stankovic, J.A., Lu, C., Abdelzaher, T.: SPEED: A stateless protocol for real-time communication in sensor networks. In Proc. of the 23rd International Confernece on Distributed Computing Systems (ICDCS-23), Providence, RI, USA,5(2003)
5. Felemban, E., Lee, C.G., Ekici, E., Boder, R., Vural, S.: Probabilistic QoS guarantee in reliability and timeliness domains in wireless sensor networks. In Proceedings of 24th Annual Joint Conference of the IEEE Computer and Communications (INFOCOM), 3(2005)2646–2657
6. Boulis, A., Ganeriwal, S., Srivastava, M.: Aggregation in Sensor Networks: An Energy-Accuracy Trade-off. Proc. of IEEE SANPA, 5(2003)1117
7. Heinzelman, W.B., et al.: An Application-Specific Protocol Architecture for Wireless Microsensor Networks. IEEE Transactions on Wireless Communications.Vol.1, No.4, 10(2002)18
8. Peng S.L., Li S.S., Liao X.K., Peng Y.X., et al.: Feedback Control with Prediction for Thread Allocation in Pipeline Architecture Web Server. In ICDCN, 2006 to be published.
9. Zeng, X., Bagrodia, R., GloMoSim, M.G.: a Library for Parallel Simulation of Large-scale Wireless Networks. In Proceedings of the 12th Workshop on Parallel and Distributed Simulations -- PADS '98, Banff, Alberta, 5(1998)
10. Younis, O., Fahmy, S.: Distributed Clustering in Ad-hoc Sensor Networks: A Hybrid, Energy-Efficient Approach.Proceedings of IEEE INFOCOM. Hong Kong, 3(2004)

A Low Power Real-Time Scheduling Scheme for the Wireless Sensor Network*

Mikyung Kang[1] and Junghoon Lee[2,**]

[1] Information Sciences Institute East, University of Southern California
[2] Dept. of Computer Science and Statistics, Cheju National University
mkkang@east.isi.edu, jhlee@cheju.ac.kr

Abstract. This paper proposes a low power real-time scheduling scheme that minimizes power consumption and reassigns polling order on the wireless sensor networks. The proposed scheme aims at enhancing the power performance based on the maximum constant power method along with the additional auxiliary schedule which can switch the primary one when a specific channel falls in a bad state or time slot is unused. Built on top of the EDF scheduling policy, the proposed scheme also tests whether an unused slot can be reclaimed without violating the constraints of subsequent real-time streams. To improve the probability of reclaimed bandwidth, it also rearranges the polling order. The simulation results show that the proposed scheme is able to enhance success ratio and it also provides minimized power consumption.

Keywords: Real-time scheduling, low power scheme, EDF policy, reclaiming scheme, wireless sensor network.

1 Introduction

Battery-powered portable systems have been widely used in many applications, such as mobile computing, wireless communications, information appliances, wearable computing as well as various industrial and military applications.As systems become more complex and incorporate more functionality, they become more power-hungry[1]. Thus, reducing energy consumption and extending battery lifespan have become a critical aspect of designing battery-powered systems. Low energy consumption is today an increasingly important design requirement for digital systems, with impact on operating time, on system cost, and, of no lesser importance, on the environment.

Reducing power and energy dissipation has long been addressed by several research groups, at different abstraction levels[2]. As a consequence, reduction of the power consumption of processors, especially in OS (Operating System) level, is important for the power-efficient design of such systems. As a method

* This research was supported by the MIC, Korea, under the ITRC support program supervised by the IITA (IITA-2006-C1090-0603-0040).

** The corresponding author.

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 401–408, 2007.
© Springer-Verlag Berlin Heidelberg 2007

to reduce power consumption of processors in OS level, VSP (Variable Speed Processor) can change its speed by varying the clock frequency along with the supply voltage when the required performance on the processor is lower than the maximum.

Wireless sensor networks use battery-operated computing and sensing devices[3]. Sensor network nodes have very limited battery life; moreover once deployed, a sensor network may be left unattended for its entire operational lifetime. This is due to the fact that sensor networks may be deployed in wide, remote, unaccessible areas.The energy-constrained nature of sensor networks calls for protocols that have energy efficiency as a primary design goal. These characteristics of sensor networks and applications motivate a MAC (Medium Access Control) in almost every way: energy conservation and self-configuration are primary goals, while per-node fairness and latency are less important.

EDF (Earliest Deadline First), which dynamically picks a message that has the closest deadline among the pending ones, is known to be the most common scheduling policy for real-time message streams[4]. WLAN can also exploit EDF for message scheduling in various ways on the assumption that the communication pattern of sensory data is fixed and known *in priori*. However, it is not clear how well this algorithms will work for the wireless sensor networks, since wireless channels are subject to unpredictable location-dependent and bursty errors, which make a real-time application fail to send or receive some of its real-time packets[5]. Also, the design of energy-efficient protocols has become an area of intense research. This paper addresses these problems based on the maximum constant power method along with the additional auxiliary schedule which can switch the primary one when a specific channel falls in a bad state or time slot is unused.

The rest of this paper is organized as follows: After issuing the problem in Section 1, Section 2 introduces some related works, Section 3 explains background and basic assumptions. Section 4 describes the proposed low power real-time scheduling scheme in detail, and Section 5 shows the result of performance measurement. Finally, Section 6 summarizes and concludes this paper.

2 Related Works

This section introduces previous works regarding power-aware MAC protocols. PAMAS (Power Aware Multi-Access Protocol with Signalling for Ad Hoc Networks) is one of the earliest contention-based protocols to address power efficiency in channel access[6]. It saves energy by attempting to avoid over-hearing among neighboring nodes. To achieve this, PAMAS uses out-of channel signaling.

The IEEE 802.11 DCF is a contention-based protocol based on CSMA/CA (Carrier Sensing Multiple Access with Collision Avoidance)[7]. IEEE 802.11 performs both physical and virtual carrier sensing. Virtual carrier sensing is achieved by sending information about duration of each frame in the headers which is used by stations as an indication of how long the channel will be busy. After this time is elapsed, stations can sense the channel again. In order to solve the

"hidden terminal" problem and avoid data frame collisions, the RTS-CTS handshake is used. Two power management mechanisms are supported: *active* and *power-saving* (PS).

The Sensor MAC protocol, or S-MAC, was developed with power savings as one of its design goals[3]. It also falls into the contention-based protocol category but achieves energy efficiency by making use of low-power radio mode. Nodes alternate between periodic sleep and listen periods. Listen periods are split into synchronization and data periods. During synchronization periods, nodes broadcast their sleeping schedule, and, based on the information received from neighbors, they adjust their schedule so that they all sleep at the same time. During data periods, a node with data to send will contend for the medium (RTS-CTS exchange). Unlike PAMAS, it only uses in-channel signaling. Finally, S-MAC applies message passing to reduce contention latency for sensor-network applications.

T-MAC, a contention-based MAC protocol for wireless sensor network, is another notable example[8]. Applications for these networks have some characteristics that can be exploited to reduce energy consumption by introducing an active/sleep duty cycle. To handle load variations in time and location T-MAC introduces an *adaptive* duty cycle by dynamically ending the active part of it. This reduces the amount of energy wasted on idle listening, in which nodes wait for potentially incoming messages, while still maintaining a reasonable throughput.

TRAMA (TRaffic-Adaptive Medium Access) protocol is introduced for energy-efficient collision-free channel access in wireless sensor networks[9]. It reduces energy consumption by ensuring that unicast, multicast, and broadcast transmissions have no collisions and by allowing nodes to switch to a low-power, idle state whenever they are not transmitting or receiving. TRAMA assumes that time is slotted and uses a distributed election scheme based on information about the traffic at each node to determine which node can transmit at a particular time slot.

3 Background and Basic Assumptions

3.1 WLAN

WLAN divides its time axis into CFP (Contention Free Period) and CP (Contention Period), which are mapped into PCF (Point Coordination Function) and DCF (Distributed Coordination Function), respectively. To provide the deterministic access to each node during CFP, AP (Access Point) polls each node according to the predefined order, and only the polled node can transmit its frame. In the DCF interval, every node including AP contends the medium via the CSMA/CA protocol. The AP periodically initiates CFP by broadcasting a *Beacon* frame that has the precedence in transmission via shorter IFS (Inter-Frame Space).

Each station is associated with a channel which has either of two states, namely, *error state* and *error-free state* at any time instant. A channel is defined between each mobile and the AP, and it can be modeled as a Gilbert channel[10]. We can denote the transition probability from state *good* to state *bad* by p

and the probability from state *bad* to state *good* by q. The pair of p and q representing a range of channel conditions has been obtained by using the trace-based channel estimation. The average error probability and the average length of a burst of errors are derived as $\frac{p}{p+q}$ and $\frac{1}{q}$, respectively. A packet is received correctly if the channel remains in state *good* for the whole duration of packet transmission. Otherwise, it is received in error. Channels between the AP and respective stations are independent of one another in their error characteristics.

3.2 DVS Algorithms

DVS (Dynamic Voltage Scaling), which adjusts the supply voltage and clock frequency dynamically, is an effective low-power design technique for embedded real-time systems[11]. Since the energy consumption E of CMOS circuits has a quadratic dependency on the supply voltage, lowering the supply voltage is one of the most effective ways of reducing the energy consumption.

$$E \propto C_L \cdot N_{cycle} \cdot V_{DD}^2, \tag{1}$$

where E, C_L, N_{cycle} and V_{DD} are dynamic power consumption, load capacitance of CMOS circuit, the number of executed cycle, and supply voltage, respectively.

For hard real-time systems, there are two types of voltage scheduling approaches depending on the voltage scaling granularity: intra-task DVS (IntraDVS) and inter-task DVS (InterDVS) [2]. The intra-task DVS algorithms adjust the voltage within an individual task boundary, while inter-task DVS algorithms determine the voltage on a task-by-task basis at each scheduling point. Based on inter-task DVS algorithms for periodic hard real-time systems, slack estimation techniques are as follows[11] : Based on the periodicity and WCET (Worst Case Execution Time) of tasks, we can estimate statically given slack times and exploit those slack times to lower the clock speed, i.e., the worst case processor utilization can be estimated and the clock speed can be adjusted based on that as in the maximum constant speed heuristic. Since the arrival time of tasks are known a priori, when a single task is active, its execution can be extended to the earliest time of the next task with the lowered clock speed and voltage as in the stretching to NTA (Next Task Arrival) technique[1].

While the above techniques exploit static information, other techniques utilize dynamic information such as workload variation of tasks. The execution time of each task is usually less than its WCET, and the actual processor utilization during run time is usually lower than the worst case processor utilization.Thus, when a task completes its execution much earlier than its WCET, the expected utilization can be recalculated based on the actual execution time of completed task, and the clock speed can be adjusted based on that as in utilization updating. Also, when a higher-priority task completes its execution earlier than its WCET, the following lower-priority tasks can use the slack time, and the clock speed can be lowered based on the slack time as in priority-based slack stealing[12].

4 Proposed Scheduling Scheme

4.1 System Model

This paper exploits the contention-free TDMA style access policy as in [5], for the real-time guarantee, as the contention resolution via packet collisions consumes the precious communication energy. As the real-time guarantee cannot be provided without developing a deterministic access schedule, the network time is divided into a series of equally sized slots to eliminate the unpredictability stemmed from access contention. Accordingly, the allocation scheme assigns each slot to real-time streams so as to meet their time constraints. The slot equals the basic unit of wireless data transmission and the other non-real-time traffic is also segmented to fit the slot size. Therefore, a preemption occurs only at the slot boundary. This network access can be implemented by making AP poll each station according to the predefined schedule during the CFP.

The traffic of sensory data is typically *isochronous* (or synchronous), consisting of message streams that are generated by their sources on a continuing basis and delivered to their respective destinations also on a continuing basis[4]. This paper follows the general real-time message model which has n streams, namely, S_1, S_2, ..., S_n, and each S_i generates a message less than WCET C_i at each beginning of its period P_i. Each packet must be delivered to its destination within D_i time unit from its generation or arrival at the source, otherwise, the packet is considered to be lost. Generally, D_i coincides with P_i to make the transmission complete before the generation of the next message. A task set is called feasible if deadline of each task is satisfied at all times.

4.2 Scheduling Scheme

Assume all tasks are released simultaneously at time 0. A typical EDF schedule, which assumes that tasks run at their WCETs (C_i), is shown in Fig. 1.(a). The IC of RF module can be manufactured by using CMOS technology. If the power of the RF module is lowered by half or if the RF module with half performance is used meaning that C_i is doubled, the EDF schedule becomes as shown in Fig. 1.(b). For example, if stream C were to take a little longer to complete, stream A would miss its deadline at time 24. When some task instances are completed earlier than their WCETs, there are more idle intervals as shown in Fig. 1.(c).

For dynamic-priority scheduling, a task set is feasible if and only if the utilization, U, is less than or equal to 1. Thus, *power scaling factor*, η, is straightforward to compute because it is equal to the utilization, given by $\eta = \sum \frac{C_i}{P_i} = \frac{1}{6} + \frac{1}{8} + \frac{2}{12} = 0.458$. After all, the maximum constant power(V) can be lowered up to 0.458% and Fig. 1 shows the case whose maximum power is $\frac{V}{2}$. To reduce the power consumption considering error characteristics, this paper proposes an auxiliary schedule, X_{1t}, as shown in Fig. 1.(b). After power scaling, the auxiliary schedule helps the runtime scheduler to pick alternative stream according to the auxiliary schedule as shown in Fig. 2.(a), when it cannot poll the stream in the primary schedule. To build an auxiliary schedule, AP generates both arrival time

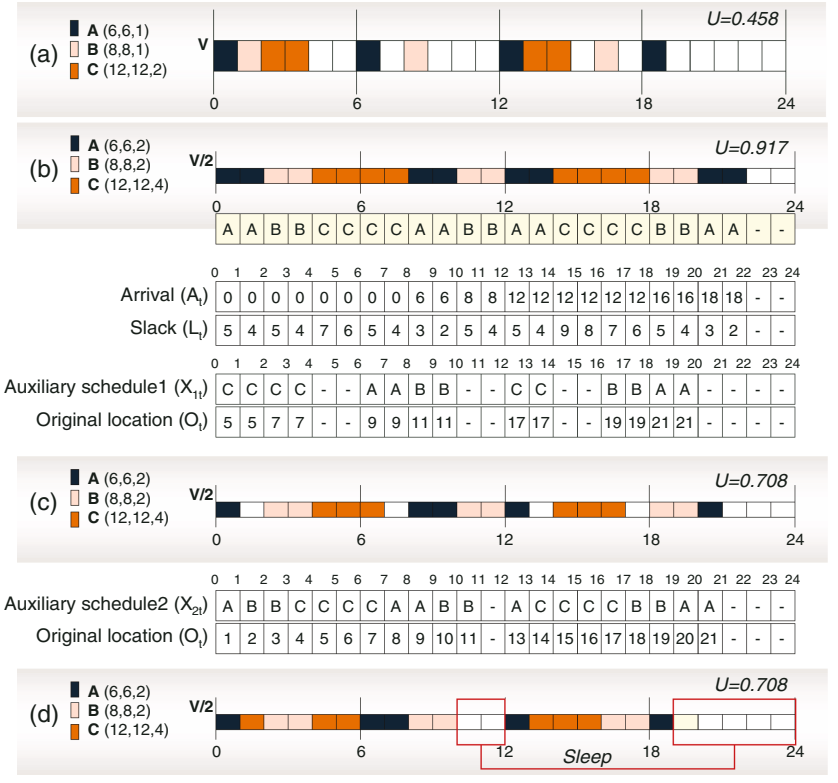


Fig. 1. A schedule for the example task set

```
for all t in the planning cycle
    t' = t + L_t
    while (t' > t)
        if (A_t' ≤ t)
            select E_t' as X_1t
        t' ← t' − 1
    end while
end for
```

(a) Auxiliary schedule

```
t' = t + 1
while (t' <= t + E_t)
    if (E_t'! = NULL)
        select E_t as E_t' ; break;
    else
        t' ← t' + 1
end while
```

(b) Secondary auxiliary schedule

Fig. 2. Auxiliary schedule

and slack time of each slot in the primary schedule based on EDF policy. By E_t, we denote the stream AP will poll at time t. If E_t is M, A_t and L_t denote the arrival time and the slack time of M, respectively. Slack means the redundant time to the completion of a given task or message. If the status of E_t is in state *bad* or E_t is empty, the proposed scheme selects auxiliary schedule (X_{1t}). None the less, if time slot is still empty, this scheme also selects secondary auxiliary

schedule (X_{2t}) as shown in Fig. 2.(b). Fig. 1.(d) shows the final allocation. In the result, the power consumption is reduced by minimizing the maximum power and the time of continuous power-down(sleep) mode is increased by collecting unused slots into one side.

5 Performance Measurement

This section measures the performance of the proposed scheduling scheme in terms of the success ratio according to the packet error rate as well as the ratio of BCET (Best Case Execution Time) to WCET using the event simulator. The success ratio means the ratio of the number of successfully transmitted packets within their deadlines to the total number of generated packets. Because the statistics of the actual execution times of instances of the tasks are not available, it is assumed that the execution time of each instance of a task is drawn from a random Gaussian distribution. Then, the BCET is varied from 10% to 100% of the WCET for each task. For the target stream sets, we fixed the length of a planning cycle to 24 as well as the number of streams to 3, and generated every possible stream sets whose utilization ranges from 0.2 to 1.0. We measured how many errors can be recovered and power-down mode can be allocated by the auxiliary polling schedule, for the packet error rate of 0.0 through 0.5. In MPEDF policy, a slack time of a task is estimated using the maximum constant power and stretching-to-NTA methods[1].

Fig. 3 plots the measured result and compares the success ratio. As shown in this figure, the performance gap gets larger with a higher packet error rate, by up to 8.6 %. The proposedEDF scheme can widen the gap between those curves, that is, considerably relieves the problem of error-prone WLAN and poor utilization of PCF operation. Fig. 4 plots the power consumption ratio according to the ratio of BCET/WCET. In the experiment, power consumption variable is aligned to the non-power-down mode(EDF). It is certain that the improvement gets larger as the error rate increases and BCET/WCET rate decreases. This is due to the fact that the recovered portion gets larger and the probability of power-down increases.

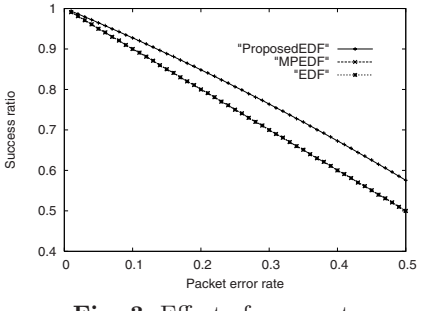


Fig. 3. Effect of error rate

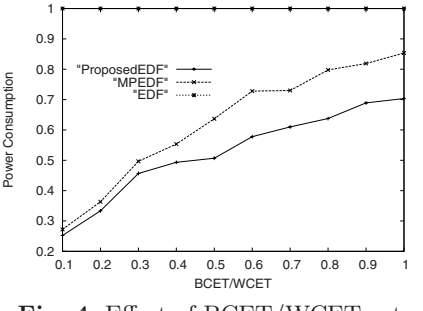


Fig. 4. Effect of BCET/WCET rate

6 Conclusion

In this paper, we have proposed a low power real-time scheduling scheme that minimizes power consumption and reassigns polling order on the wireless sensor networks. The proposed scheme aims at enhancing the power performance based on the maximum constant power method along with the additional auxiliary schedule which can switch the primary one when a specific channel falls in a bad state or time slot is unused. The simulation results show that the proposed scheme is able to enhance success ratio and it also provides minimized power consumption. Finally, we are to apply the power management scheme proposed in this paper to the RM style polling framework combined with an error control mechanism.

References

1. Shin, Y., Choi, K., Sakurai, T.: Power optimization of real-time embedded systems on variable speed processors. Proc. of the International Conference on Computer-Aided Design (2000) 365-368
2. Gruian, F.: Hard Real-Time Scheduling for Low-Energy Using Stochastic Data and DVS Processors. Proc. of the International Symposium on Low Power Electronics and Design (2001) 46-51
3. Heinzelman, W.R., Chandrakasan, A., Balakrishnan, H.: Energy-Efficient Communication Protocol for Wireless Microsensor Networks. Pro. of the 33rd Hawaii International Conference on System Sciences (2000)
4. Liu J.: Real-Time Systems. Prentice Hall (2000)
5. Adamou, M., Khanna, S., Lee, I., Shin, I., Zhou, S.: Fair real-time traffic scheduling over a wireless LAN. Proc. IEEE Real-Time Systems Symposium (2001) 279–288
6. Singh, S., Raghavendra, C.: PAMAS : Power Aware Multi-Access protocol with Signalling for Ad Hoc Networks. ACM Computer Communications (1999)
7. LAN MAN Standards Committee of the IEEE Computer Society: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications, IEEE std 802.11-1999 edition (1999) also available at *http://standards.ieee.org/getieee802/download/802.11-1999.pdf*.
8. Dam, T.V., Langendoen, K.: An Adaptive Energy-Efficient MAC Protocol for Wireless Sensor Networks. ACM Sensys '03 (2003) 171-180
9. Rajendran, V., Obraczka, K., Garcia-Luna-Aceves, J.J.: Energy-Efficient, Collision-Free Medium Access Control for Wireless Sensor Networks. ACM Sensys '03 (2003)
10. Shah, S., Chen, K., Nahrstedt, K.: Dynamic bandwidth management for single-hop ad hoc wireless networks. ACM/Kluwer Mobile Networks and Applications (MONET) Journal. 10 (2005) 199-217
11. Kim, W., Shin, D., Yun, H., Kim, J., Min, S.: Performance Evaluation of Dynamic Voltage Scaling Algorithms for Hard Real-Time Systems. J. Low Power Electronics. Vol. 1 (2005) 1-11
12. Aydin, H., Melhem, R., Mosse, D., Alvarez, P.M.: Dynamic and aggressive scheduling techniques for power-aware real-time systems. Proc. of IEEE Real-Time Systems Symposium (2001) 95-105

Analysis of an Adaptive Key Selection Scheme in Wireless Sensor Networks

Guorui Li[1], Jingsha He[2], and Yingfang Fu[1]

[1] College of Computer Science and Technology
Beijing University of Technology
Beijing 100022, China
{liguorui,fmsik}@emails.bjut.edu.cn
[2] School of Software Engineering
Beijing University of Technology
Beijing 100022, China
jhe@bjut.edu.cn

Abstract. Sensor networks are suitable for a variety of commercial and military applications due to their self-organization characteristics and distributed nature. As a fundamental requirement for security functionality in sensor networks, key management plays a central role in authentication and encryption. In this paper, we describe an Adaptive Key Selection (AKS) scheme for multi-deployment in sensor networks and analyze the scheme in the aspects of security, connectivity and overhead. Our analysis shows that the AKS scheme can greatly improve the connectivity of sensor nodes while maintaining the security of an existing multi-deployment scheme.

Keywords: sensor networks, security, key predistribution, multi-deployment.

1 Introduction

Recent advances in micro-electro-mechanical systems, electronics and wireless communications have made it practical now to develop and deploy low-cost, high-performance and low-power sensor nodes. These nodes are equipped with sensing, processing and communication capabilities. In such a network, security is important to guarantee confidentiality, integrity and availability of transported data. As the basic requirement for security functionality, key management plays a central role in data encryption and in authentication. However, due to energy and resource limitations in sensor nodes, many ordinary security mechanisms are deemed impractical, if not infeasible, in sensor networks.

There are currently three types of key management schemes that are commonly used in sensor networks: trusted server scheme, self-enforcing scheme, and key predistribution scheme. The first type of scheme, i.e., the trusted server scheme, relies on a trusted server for key distribution and management. The second type of scheme, i.e., the self-enforcing scheme, relies on asymmetric cryptography for key distribution and management using public key certificates. However, the lack of a trusted infrastructure in application environments and limited computation and energy resources in sensor nodes make these two types of schemes undesirable. The third type of scheme, i.e., the

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 409–416, 2007.
© Springer-Verlag Berlin Heidelberg 2007

key predistribution scheme, is one in which cryptographic keys are predistributed in all sensor nodes prior to deployment [1]. There are already several key predistribution schemes that have been proposed, such as the basic probabilistic key predistribution scheme [2], the q-composite key predistribution scheme [3], the random pairwise key scheme [3], the random subset assignment scheme [4], the grid-based key predistribution scheme [4], the closest pairwise key predistribution scheme [5], and the closest polynomials predistribution scheme [5]. The Adaptive Key Selection scheme is developed based on the above basic schemes.

As time goes by, some sensor nodes may be destroyed, compromised or dead. Since these nodes no longer work properly, the sensor network may become fragmented. The consequence is that not all collected data can be transmitted to the sink node that collects the data. Therefore, new sensor nodes have to be deployed in the network to replace those out-of-function nodes to reestablish a connected network. However, little work has been done so far to address this concern.

One simple method is to deploy a group of new sensor nodes that are loaded with predistributed keys selected from the same key pool as that for the previously deployed set of sensor nodes. We call this scheme the basic multi-deployment scheme. The benefit of this scheme is that the newly deployed sensor nodes share the same group of predistributed keys as those for the previously deployed sensor nodes so that they can establish connections fairly easily. However, its shortcoming is equally obvious. That is, the sensor nodes that have already been compromised will have negative impact on the newly deployed sensor nodes. As the result, the newly deployed sensor nodes may not be safe from the very beginning of their deployment when the percentage of the compromised sensor nodes exceeds a certain number.

Arjan Durresi et al. proposed the Secure Continuity for Sensor Networks (SCON) scheme [6]. In this scheme, the sensor nodes that belong to different deployment sets are loaded with predistributed keys from different key pools. And the bridge nodes with large memory and high computation power are deployed at the same time to help the sensor nodes establish secure links between the sensor nodes in the new deployment set and those in the previous deployment set. However, the probability of establishing a secure link between any two sensor nodes that belong to two different deployment sets is very low because the bridge node is only loaded with predistributed keys that are randomly selected from the key pool of the new deployment set and that of the previous deployment set. Therefore, extra help is needed from actors that possess arbitrary moving capacity so that they can be deployed in lower connectivity areas to help sensor nodes establish secure links.

In this paper, we describe an Adaptive Key Selection (AKS) scheme for multi-deployment in sensor networks. This scheme can be applied to the hierarchical wireless sensor networks in multiple deployments of sensor nodes. We use three types of network elements in this scheme: base station, cluster head nodes and ordinary sensor nodes. We assume that the base station stores all the predistributed keys for every sensor node and the cluster head nodes have strong computation, memory and communication power and can communicate with the base station using an asymmetric encryption algorithm. Every cluster head node executes the AKS algorithm to select the optimal key set and assists in establishing secure links between any two sensor nodes that belong to two different deployment sets. Our analysis shows that our scheme can greatly improve the connectivity between any two sensor nodes that belong to two

different deployment sets. The flexibility of the AKS scheme ensures that it can be combined with any key predistribution scheme described in [1-5].

The rest of the paper is organized as follows. In the next section, we describe the AKS scheme. In Section 3, we analyze the connectivity, security and overhead aspects of this scheme. In Section 4, we identify some related work in sensor network security. Finally, in Section 5, we conclude this paper and discuss some future research directions.

2 The Adaptive Key Selection Scheme

In the AKS scheme, all sensor nodes in the network are classified into three different types: the base station, the cluster head nodes and the ordinary sensor nodes. These nodes perform different functions to achieve the goal, which is described in this section.

2.1 Key Determination in the Cluster Head Nodes

The strong computation, memory and communication power of the cluster head nodes enable these nodes to use an asymmetry encryption algorithm to secure communication with the base station node. Each cluster head node operates by following the procedure below:

(1) A cluster head node broadcasts a query message to acquire the ID information of its neighboring sensor nodes.
(2) After receiving a query message, a sensor node transmits its ID information to the cluster head node.
(3) The cluster head node collects the ID information from the neighboring sensor nodes, encrypts such information in a message using its private key K_{pri} and sends the message to the base station. The base station verifies the message using the cluster head node's public key and retrieves the ID information.
(4) The base station encrypts the predistributed keys for the sensor nodes contained in the message using the cluster head node's public key K_{pub} and sends the resulting message to the cluster head node.
(5) The cluster head node decrypts the received message from the base station using its private key and gets the predistributed keys for its neighboring sensor nodes.
(6) The cluster head node runs the AKS algorithm presented below to select the optimal key set.

2.2 The Adaptive Key Selection (AKS) Algorithm

The Adaptive Key Selection algorithm works as follows:

(1) For all the sensor nodes, S_1, …, S_m, that belong to the same deployment set, we build a $n \times m$ matrix M to describe key predistribution status where n is the number of different keys and m is the number of sensor nodes. We set M_{ij}=1 if sensor node S_i is predistributed with the jth key, otherwise, M_{ij}=0.
(2) We sum the matrix M by column and the resulting $n \times 1$ vector V represents the number of each predistributed key in the deployment set.

(3) We select the maximum element in vector V and the corresponding row number i_{max} is the number of the optimal key selected in this round. If there is more than one such maximum element, we just select the first such element.
(4) We set all the elements of column j to be 0 if M_{imaxj}=1 to exclude the nodes that hold the selected optimal key.
(5) Repeat steps (2)-(4) until all the elements in M become 0.

When the algorithm completes, we have the queue for optimal key selection for the deployment set. All the keys in this queue are critical keys and they form the minimum spanning set of all the predistributed keys for the deployment set. The closer the position of a key to the front of the queue, the higher the number of sensor nodes that share this key, and this key should, therefore, be selected with a higher priority.

2.3 Key Establishment in Ordinary Sensor Nodes

In an ordinary sensor node, key establishment includes three phases: (1) key predistribution phase, (2) direct key establishment phase and (3) path key establishment phase. Two sensor nodes that belong to two different deployment sets can establish a secure path key with the help of the cluster head node. The cluster head node can certainly do so for two sensor nodes in the same deployment set because it stores all the critical keys for this deployment set. If it stores all the critical keys for different deployment sets, we can see that the length of any path key is no more than two hops.

3 Analysis

We analyze the AKS scheme in terms of security, connectivity and overhead to demonstrate its feasibility. In particular, we compare the AKS scheme with the SCON scheme in the analysis where appropriate to show that the former is more advantageous over the latter in some key aspects of multi-deployment in sensor networks.

3.1 Security Analysis

Fig. 1 shows the relationship between the percentage of compromised links between sensor nodes in a new deployment and that of compromised nodes in the previous deployment in the basic multi-deployment scheme [6]. We can thus see that the

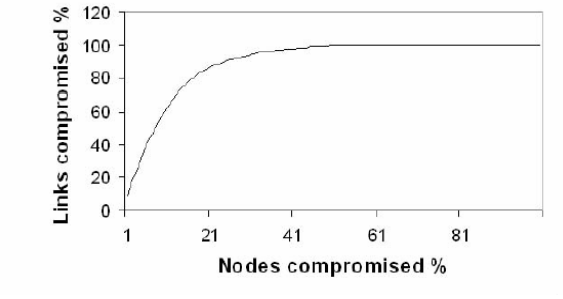


Fig. 1. Relationship between the percentage of compromised links and that of compromised nodes in the basic multi-deployment scheme

effectiveness of a new deployment of sensor nodes is heavily influenced by the number of compromised sensor nodes in the previous deployment set. In the SCON and the AKS schemes, however, the sensor nodes of a new deployment set are immune to the compromised sensor nodes of the previous deployment set, for the new nodes are loaded with keys predistributed from a different key set. That is, the percentage of compromised links in the SCON and the AKS scheme is 0.

3.2 Connectivity Analysis

In the SCON scheme, the number of keys that are randomly selected from the new deployment set and predistributed into the bridge node is the same as that from the previous deployment set. Therefore, the probability of establishing a link between any two sensor nodes that belong to two different deployment sets is:

$$\left(1 - \frac{C_{n-k}^{k}}{C_{n}^{k}}\right)^{2}$$

where n is the size of key predistribution pool and k is the number of predistributed keys in each sensor node. And the bridge node is predistributed with $2k$ number of keys.

In the AKS scheme, the probability of establishing a link between two senor nodes that belong to two different deployment sets is higher than that in the SCON scheme. This is because the keys stored in the cluster head node are selected from the optimal key selection queues of the respective deployment sets. To validate our claim, we did some simulation in which we used 200 as the number of sensor nodes for each deployment set and 1,000 as the size of the pool for predistributed keys. Each sensor node would be predistributed with k keys, where k=0,…,100, that were randomly selected from the key pool. Fig. 2 shows that connectivity between any two sensor nodes that

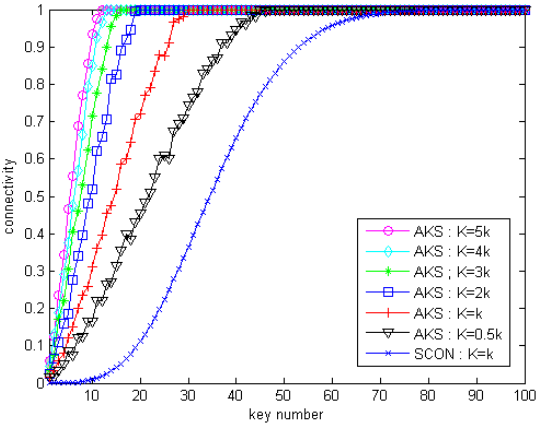


Fig. 2. The connectivity between sensor nodes that belong to two different deployment sets in the SCON and the AKS schemes, respectively

belong to two different deployment sets in the SCON scheme and that in the AKS scheme. We can thus conclude that network connectivity using the AKS scheme is much better than that using the SCON scheme.

Fig. 3 shows the relationship between connectivity among sensor nodes that belong to two different deployment sets in the AKS scheme and the number of keys k in each sensor node. It also shows the result for different K, the number of keys stored in a cluster head node. We can thus see that the higher the number of keys K is in a cluster head node, the higher the connectivity is between any two sensor nodes. We can also see that even when the number of keys in a cluster head in the AKS scheme is only half of that in the SCON scheme, the connectivity between any two sensor nodes is much higher in the AKS scheme than that in the SCON scheme.

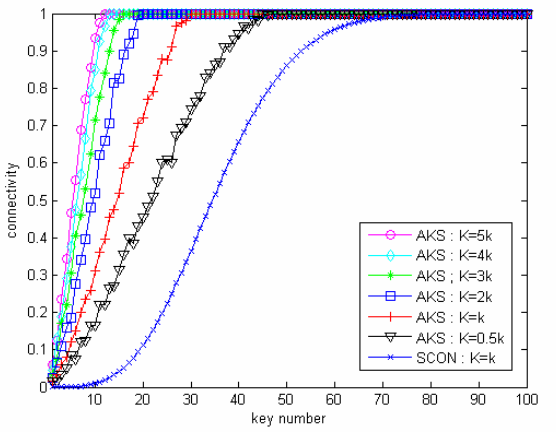


Fig. 3. The connectivity between sensor nodes that belong to two different deployment sets in the AKS scheme for different key storage sizes in the cluster head nodes

3.3 Overhead Analysis

The AKS scheme would incur communication and computational overhead to the cluster head nodes while having little impact on the ordinary sensor nodes. The communication overhead mainly results from key request and reply messages between the cluster head nodes and the base station node, and the computational overhead mainly results from encryption and decryption computation in the cluster head nodes for requesting and receiving keys from the base station node as well as from optimal key selection in the AKS algorithm. That is why we require that the cluster head nodes have the necessary computation, storage and communication power.

In order to decrease the communication overhead and the computational overhead, we can predistribute all the keys in every deployment set to the cluster head nodes before the actual deployment. After deployment, the cluster head nodes can collect the ID information of its neighboring sensor nodes and only reserve the optimal keys selected using the AKS algorithm while removing all the other keys to enhance the security of key management [7].

4 Related Work

Nowadays, there are many studies in the area of security in wireless sensor networks. These studies are mostly focused on key management, authentication, and vulnerability analysis. In addition to studies on key predistribution schemes [1-6], intrusion detection system (IDS) is also very important to detect compromised sensor nodes and ensure the security of the whole network [8, 9]. Furthermore, Wood and Stankovic identified a number of DoS attacks in sensor networks [10] and Deng et al. described a path-based DoS attack and proposed a solution using one-way hash chains to protect end-to-end communication against this type of DoS attacks [11].

5 Conclusion and Future Work

In this paper, we described an Adaptive Key Selection scheme and the corresponding algorithm for multi-deployment in sensor networks. Our analysis shows that the AKS scheme can greatly improve the connectivity of sensor nodes while maintaining the security in existing multi-deployment schemes. In the future, we will focus on developing methods to protect the cluster head nodes and to detect compromised sensor nodes in order to further improve the security of sensor networks.

References

1. Du, W.L., Deng, J., Han, Y.S., Chen, S., Varshney, P.K.: A Key Management Scheme for Wireless Sensor Networks Using Deployment Knowledge. Proc. IEEE INFOCOM 2004. (2004) 586-597
2. Eschenauer, L., Gligor, V.D.: A Key-Management Scheme for Distributed Sensor Networks. Proc. 9th ACM Conference on Computer and Communications Security. (2002) 41-47
3. Chan, H., Perrig, A., Song, D.: Random Key Predistribution Schemes for Sensor Networks. Proc. IEEE Symposium on Research in Security and Privacy. (2003) 197-213
4. Liu, D.G., Ning, P.: Establishing Pairwise Keys in Distributed Sensor Networks. Proc. 10th ACM Conference on Computer and Communications Security. (2003) 52-61
5. Liu, D.G., Ning, P.: Location-based Pairwise Key Establishments for Static Sensor Networks. Proc. 2003 ACM Workshop on Security in Ad Hoc and Sensor Networks. (2003) 72-82
6. Durresi, A., Bulusu, V., Paruchuri, V., Barolli, L.: Secure and Continuous Management of Heterogeneous Ad Hoc Networks. Proc. 20th International Conference on Advanced Information Networking and Applications. (2006) 511-516
7. Simplot-Ryl, D., Simplot-Ryl, I.: Connectivity Preservation and Key Distribution in Wireless Sensor Networks Using Multi-deployment Scheme. Proc. 3rd International Conference on Ubiquitous Intelligence and Computing. (2006) 988-997
8. Silva, A., Martins, M., Rocha, B., Loureiro, A., Ruiz, L., Wong, H.: Decentralized Intrusion Detection in Wireless Sensor Networks. Proc. 1st ACM International Workshop on Quality of Service & Security in Wireless and Mobile Networks. (2005) 16-23

9. Roman, R., Zhou, J.Y., Lopez, J.: Applying Intrusion Detection Systems to Wireless Sensor Networks. Proc. 3rd Consumer Communications and Networking Conference. (2006) 640-644
10. Wood, D., Stankovic, J.A.: Denial of Service in Sensor Networks. IEEE Computer, Vol. 35, No. 10. (2002) 54-62
11. Deng, J., Mishra, S.: Defending Against Path-based DoS Attacks in Wireless Sensor Networks. Proc. 3rd ACM workshop on Security of Ad hoc and Sensor Networks. (2005) 89-96

Unusual Event Recognition for Mobile Alarm System

Sooyeong Kwak, Guntae Bae, Kilcheon Kim, and Hyeran Byun

Dept. of Computer Science, Yonsei University, Seoul, Korea, 120-749
{ksy2177,gtbae,kimkch}@cs.yonsei.ac.kr, hrbyun@yonsei.ac.kr

Abstract. This paper proposes an unusual event recognition algorithm, which is a part of a mobile alarm system. Our systems focus on unusual event. When the system detects the unusual event, the photos of emergency situation are passed to the user's portable devices such as mobile phone or PDA along with event description to help the user's final decision. The system combines the foreground segmentation, object tracking and unusual event recognition to detect the *Drop off*, *Abandon* and *Steal bag* event. The event recognition module constructs the Bayesian network of each event and uses inference algorithm to detect the unusual event. The proposed system tested in PETS2006 and CAVIAR dataset. The proposed algorithm showed good results on the real world environment and also worked at real time speed.

Keywords: Mobile alarm system, Background subtraction, Object tracking, Event recognition.

1 Introduction

Visual surveillance is a major research area in computer vision. The recent rapid increase in the number of surveillance cameras has led to a strong demand for automatic methods of processing their outputs. Due to this fact, the necessity of automatic techniques which process and analyze human behaviors and activities is more evident each day.

There have been a number of famous visual surveillances systems. The IBM Smart Surveillance System[1] is a middleware offering for use in surveillance systems and provides video based behavioral analysis capabilities. W4[2] system employs a combination of shape analysis and tracking, and constructs model of people's appearances in order to detect and track groups of people as well as monitor their behaviors even in the presence of occlusion and in outdoor environment. The VSAM system[3] can monitor activities over a large area using multiple cameras that are connected into a network. It can detect and track multiple persons and vehicles within cluttered scenes and monitor their activities over long periods of time.

This paper proposes an unusual event recognition algorithm, which is a part of a mobile alarm system. Unlike most previous surveillance system, our systems focus on unusual event. When the system detects the unusual event, the photos of emergency situation are passed to the user's portable devices such as mobile phone or PDA along with event description to help the user's final decision. Fig. 1 shows that overall system flow chart of the proposed event recognition for mobile alarm system.

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 417–424, 2007.
© Springer-Verlag Berlin Heidelberg 2007

Our proposed system can be divided into three parts: foreground segmentation module detects the location of person and bag using background subtraction method; object tracking module tracks the detected objects which deal with occlusion of multiple objects; and event recognition module which integrates the tracking results in order to recognize three unusual events.

This rest of this paper is organized as follows. Video is segmented into background and foreground regions by a background subtraction algorithm described in section 2. Section 3 presents tracking algorithm. Section 4 describes the event recognition method. Some experimental results are given in Section 5 and then summary and conclusions are presented in Section 6.

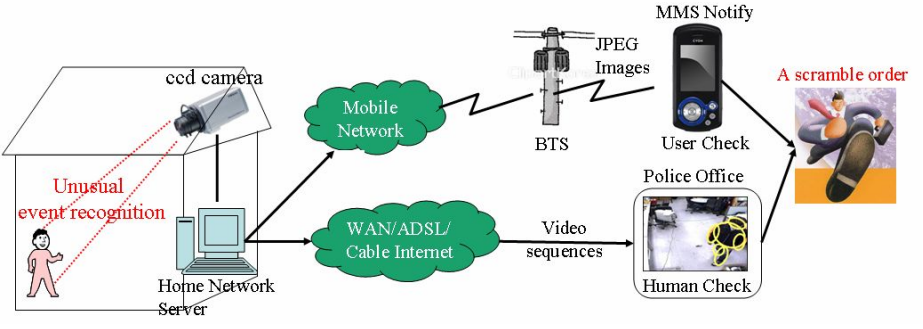


Fig. 1. Overall system flow chart of the proposed event recognition for mobile alarm system

2 Foreground Segmentation

In order to extract the foreground region, we used the background subtraction method. Background subtraction has been widely used to detect and track moving object obtained from a static camera. We proposed modified Horprasert's algorithm which call the sequential Horprasert's algorithm.

Horprasert et al.[4] proposed the statistical background model that separates the brightness from the chromaticity component in the batch mode. The batch mode requires long training time and the training images are to be stored in memory while the sequential model does not. Furthermore, the sequential algorithm is of great advantage in practice, especially when cameras can move around and stop to detect foreground objects. We modify the Horprasert's algorithm to work in the sequential mode.

In Horprasert's background model a pixel p is modeled by 4-tuple $< \mu_p, \sigma_p, a_p, b_p >$ where is μ_p the expected color value, σ_p is the standard deviation of RGB color value, a_p is the variation of the brightness distortion, and b_p is the variation of the chromaticity distortion of pixel p. We calculate sequences of $< \mu_p{}^t, \sigma_p{}^t, a_p{}^t, b_p{}^t >$ at every tth frame when the scene is stable with the assumption that the sequence approximates $< \mu_p, \sigma_p, a_p, b_p >$ as t becomes bigger. It is shown in equation (1).

$$< \mu_p, \sigma_p, a_p, b_p > \rightarrow < \mu_p{}^t, \sigma_p{}^t, a_p{}^t, b_p{}^t > \quad (1)$$

The sequences are calculated by the following. We define the expectation and standard deviation of the color vector of pixel p as $\mu_p{}^t = (\mu_r{}^t(p), \mu_g{}^t(p), \mu_b{}^t(p))$ and $\sigma_p{}^t = (\sigma_r{}^t(p), \sigma_g{}^t(p), \sigma_b{}^t(p))$, respectively, up to the tth frame. For sequential background training process we calculate $\mu_p{}^t$ and $\sigma_p{}^t$ using equation (2) and (3).

$$\mu_i^t(p) = \frac{t-1}{t}\mu_i^{t-1}(p) + \frac{1}{t}C_i^t(p) \quad (2)$$

$$\sigma_i^t(p)^2 = \frac{t-1}{t}\sigma_i^{t-1}(p) + \frac{1}{t}\left(C_i^t(p) - \mu_i^t(p)\right)^2 \quad (3)$$

Where i=r,g,b and $C_p{}^t = (C_r{}^t(p), C_g{}^t(p), C_b{}^t(p))$ is the observed color of pixel p at the tth frame. The brightness and chromaticity distortions can be obtained using the temporal mean $\mu_p{}^t$ and standard deviation $\sigma_p{}^t$. The variation of the brightness distortion $a_p{}^t$ and the chromaticity distortion $b_p{}^t$ are calculated by equation (4).

$$(a_p^t)^2 = \frac{t-1}{t}(a_p^{t-1})^2 + \frac{1}{t}(a_p^t - 1)^2, \ (b_p^t)^2 = \frac{t-1}{t}(b_p^{t-1})^2 + \frac{1}{t}(\gamma_p^t)^2 \quad (4)$$

The details of the algorithms can be found in [5]. Fig 2 shows results of proposed background subtraction.



(a) (b) (c) (d)

Fig. 2. The results of Background subtraction (a) Original image in CAVIAR dataset (b) Results of Background subtraction in CAVIAR dataset (c) Original image in PETS2006 dataset (d) Results of Background subtraction in PETS2006 dataset

3 Object Tracking

For detected object tracking, we take the results of foreground segmentation which is described by a bounding box (It can be know the location and size of the tracking object). The tracking module applies Senior's method[6] to handle the occlusion and merge/split between multiple object using appearance model. This model is used to localize object during partial occlusions, detect complete occlusions and resolve depth ordering of objects during occlusions.

At first, the bounding box distance between each of the foreground region and all the currently active tracks is computed to track correspondence. If the distance is lower than the threshold value, we consider that each foreground region associated to tracks. And then, it analyzes four possible results such as existing object, new object, merge detected and split detected. Each case has the predefined rules which are well described in Senior paper[6]. Fig. 3 shows that the processes of tracking when multiple objects are occluded in PETS2006 Dataset.



Fig. 3. The processes of tracking when multiple objects are occluded in PETS2006 Dataset, with frame numbers

4 Event Recognition

We define three events such as *Drop off*, *Abandon*, and *Steal bag*. In order to recognize these events, we construct Bayesian network in each event which shown in Fig. 4. Our event structure refers to [7]. *Drop off bag* and *Steal bag* events need three pieces of evidence and *Abandon bag* event needs two pieces of evidence. All evidences of each event are described in Table 1. Useful information of tracking object such as speed, direction and the distance of between two object can computed from the bounding box and it used as evidence. If all evidences are observed, we can know the probability of each event using Bayesian inference[8]. Given the evidence, the posterior probability is computed using the equation (5).

$$P(Event \mid e_k^1, e_k^2, ..., e_k^n) = \frac{\prod_{i=1}^{n} P(e_k^i \mid Event) P(Event)}{\prod_{i=1}^{n} P(e_k^i \mid Event) P(Event) + \prod_{i=1}^{n} P(e_k^i \mid \neg Event) P(\neg Event)} \tag{5}$$

where $\{Dropp\ off, Abandon, Steal\} \in Event, \{D,A,S\} \in k, \ i = 1,2,...,n$

To compute equation (5), several prior and conditional probabilities are needed. The prior and the conditional probability can be determined by expert. These are shown in equation (6). When the posterior probability of each event is over the 0.5, the alarm is triggered.

$$P(Event) = 0.5, P(\neg Event) = 0.5\ , P(e_k^i \mid Event) = 0.9, P(e_k^i \mid \neg Event) = 0.5 \tag{6}$$

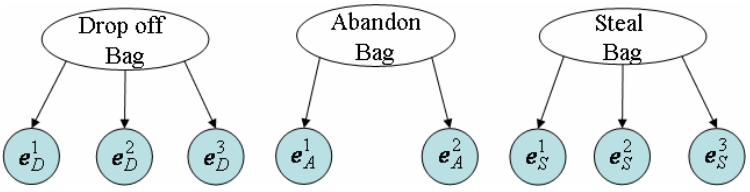


Fig. 4. The structure of Bayesian networks (a) Bayesian model of *Drop-off bag* event (b) Bayesian model of *Abandon bag* event (c) Bayesian model of *Steal bag* event

Table 1. The definitions of evidence

Event	Evidence
Drop off bag	e_D^1 : The bag did not appear 0.1 second ago.
	e_D^2 : The bag shows up now.
	e_D^3 : The distance between the person and the bag now is less than 30 pixels
Abandon bag	e_A^1 : The distance between the person and the bag 0.1 second ago is less than 40 pixels
	e_A^2 : The distance now is larger than 60 pixels
Steal bag	e_S^1 : The person approaches the bag and the distance between the person and the bag is less than 60 pixels.
	e_S^2 : The person stops near the bag.
	e_S^3 : The person takes away the bag

5 Experimental Results

The proposed system was implemented in C/C++ and was run on Pentium IV-3.0 GHz PC with 1G RAM. We used the PETS2006 Dataset S1 (Take 1-C)[9] and CAVIAR Test Dataset (Person leaving bag by wall)[10] in order to detect the three unusual event, such as the *Drop off*, *Abandon*, and *Steal bag*, in a real world environment. The description of test videos shows below. The average running time of the proposed system was about 13fps.

PETS2006 Dataset S1 (Take 1-C): The scenarios are filmed from multiple cameras and involve multiple actors. The scenario contains a person with a rucksack who loiters before leaving the item of bag unattended. The dataset consists of 1479 frames recording activities in Victoria metro station. The video sequence includes a total of 24 moving objects, people and bag appearing at close, medium and fat distances from the camera. The image size is 360x288 pixels.

CAVIAR Test Dataset (Person leaving bag by wall): These include a person leaving a package in a public place. The dataset consists of 837 frames recording activities in the entrance lobby of the INRIA Labs at Grenoble, France. The resolution is half-resolution PAL standard(384x288 pixels, 25 frames per second).

5.1 Performance of Object Tracking

Performance of the object tracking was evaluated with respect to the ground truth in each frame of test sequences. The evaluation method refers to [11]. Each frame tested to see if the number of objects as well as their sizes and locations match the corresponding ground truth data for that particular frame. To evaluate the tracking algorithm, we compute True Negative(TN), True Positive(TP), False Negative(FN), and False Positive(FP) for every frame in the sequence. These definitions describe below. Also, we compute the *Accuracy* using the equation (7).

- TN: Number of frames where both ground truth and system results agree on the absence of any object.

- TP: Number of frames where both ground truth and system results agree on the presence of one or more objects, and the bounding box of at least one or more objects coincides among ground truth and tracker results.

- FN: Number of frames where ground truth contains at least one object, while system either does not contain any object or none of the system's objects fall within the bounding box of any ground truth object.

- FP: Number of frames where system results contain at least one object, while ground truth either does not contain any object or none of the ground truth either does not contain any object or none of the ground truth's objects fall within the bounding box of any system object.

Table 2 shows the performance of the tracking algorithm. Because of the foreground segmentation part does not detect the small moving objects in PETS2006 dataset, the *Accuracy* is not good.

$$Accuracy = \frac{TP + TN}{TF} \tag{7}$$

where TF is total number of frames.

Table 2. Performance of the tracking algorithms

	PETS2006 Dataset S1 (Take 1-C)	CAVIAR Test Dataset (Person leaving bag by wall)
TP	1258	599
TN	46	195
FP	45	15
FN	130	31
Accuracy	0.881	0.945

5.2 Performance of Event Recognition

In order to evaluate the event recognition, we compare the frame number when an even is triggered with our results and the ground truth. The performance of the event recognition shows in Table 3. The key frame of the three events such as *Drop off*, *Abandon* and *Steal bag* shown in Fig. 5. We can see in Table 3, the alarm event is detected within an error of 25 frames except the abandon bag event in PETS2006 Dataset S1. In that sequence, the merge and split algorithm in tracking module has not separate quickly the human and the bag.

Table 3. Performance of event recognition

	PETS2006 Dataset S1 (Take 1-C)		CAVIAR Test Dataset (Person leaving bag by wall)	
	Ground truth	Our result	Ground truth	Our result
Drop off bag	1922	1935	949	974
Abandon bag	2086	2100	988	1003
Steal bag	none	none	1348	1359

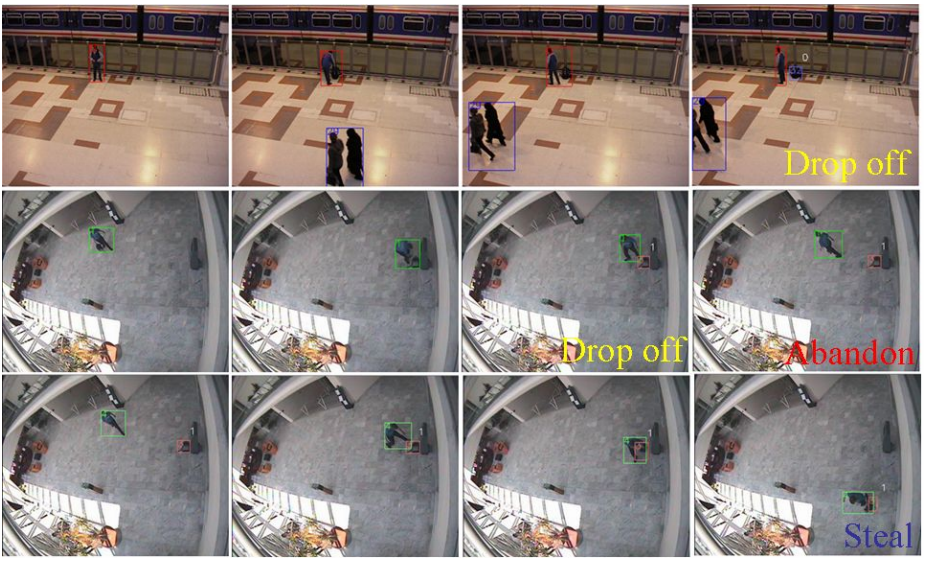


Fig. 5. It shows the key frame of the three events. First row shows the *Drop off bag* event in PETS2006 test dataset and the second row shows the *Drop off bag* and *Abandon bag* event in CAVIAR test dataset. The *Steal bag* event in CAVIAR test dataset is shown at last row.

6 Summary and Conclusions

This paper proposed unusual event recognition for mobile alarm system. Our system has three main modules such as moving object detection, tracking and event recognition. In

order to detect the foreground object, we used the background subtraction method. After candidate foreground region are detected, appearance model is used for moving object tracking. In this paper, we also used Bayesian inference algorithm in order to recognize the unusual event. The proposed algorithm showed good results on the real world environment and also worked at real time speed. The proposed framework can be easily employed or integrated into a variety of vision surveillance systems.

Acknowledgments. This research was supported by the Ministry of Information and Communication, Korea under the Information Technology Research Center support program supervised by the Institute of Information Technology Assessment, IITA-2005-(C1090-0501-0019).

References

1. Chiao-Fe Shu, Hampapur A., Lu M., Brown L., Connell J., Senior A., and Yingli Tian : IBM smart surveillance system (S3): a open and extensible framework for event based surveillance, IEEE Conference on Advanced Video and Signal Based Surveillance (2005) 318-323
2. I. Haritaoglu, D. Harwood, and L. S. Davis : W : Real-time surveillance of people and their activities, IEEE Transaction on Pattern Analysis and Machine Intelligence, (2000) Vol. 22, 809–830
3. R. T. Collins, A. J. Lipton, T. Kanade, H. Fujiyoshi, D. Duggins, Y. Tsin, D. Tolliver, N. Enomoto, O. Hasegawa, P. Burt, and L.Wixson : A system for video surveillance and monitoring, Carnegie Mellon University, Pittsburgh, PA, Technical Report, CMU-RI-TR-00-12, (2000)
4. T. Horprasert, D. Harwood, L.S. Davis: A statistical approach for real-time robust background subtraction and shadow detection. Proceeding of IEEE Frame RateWorkshop (1999) 1-19.
5. Jung-Ho Ahn and Hyeran Byun: Human silhouette extraction method using region based background subtraction, International Conference on Mirage 2007 (To be appear)
6. Andrew Senior: Tracking people with probabilistic appearance models, Proceedings 5th IEEE International Workshop on PETS, (2002)
7. Fengjun Lv, Xuefeng Song, Bo Wu, Vivek Kumar Singh, and Ramakant Nevatia.: Left-Luggage Detection using Bayesian Inference, Proceedings 9th IEEE International Workshop on PETS, (2006) 83-90
8. D'Ambrosio : Inference in Bayesian networks, AI Magazine (1999) 21-35
9. http://homepages.inf.ed.ac.uk/rbf/CAVIARDATA1/
10. http://www.cvg.rdg.ac.uk/PETS2006/index.html
11. Faisal Bashir and Fatih Porikli : Performance Evaluation of Object Detection and Tracking Systems, Proceedings 9th IEEE International Workshop on PETS, (2006) 7-13

Information Exchange for Controlling Internet Robots

Soon Hyuk Hong, Ji-Hwan Park, Key Ho Kwon, and Jae Wook Jeon

School of Information and Communication Engineering, Sungkyunkwan University
Chunchun-Dong, Jangan-Gu, Suwon City, Korea 440-746
jwjeon@yurim.skku.ac.kr

Abstract. In order to control a remote robot over the internet, the operator needs to be able to exchange information with it in real-time. However, the camera image generally used for providing feedback information is too large to send in real-time over the internet. Furthermore, it takes a long time for the operator to send complex robot commands repetitively. These time delays in exchanging feedback and command information between the remote robot and its operator make it difficult to control internet robots efficiently. This paper proposes an information exchange technique for controlling a remote robot over the internet, in which the desired information is extracted from the image in order to reduce the time required to send the feedback information, and a task-level command is used to reduce the time required to send the robot commands. As a result, the proposed technique can help an operator to control a remote robot in real-time over the internet.

Keywords: internet robot, robot, information exchange.

1 Introduction

In the early stages of robotics, remote robots were controlled through a dedicated network, however, more recently, remote robots have been controlled over the internet, because of its use of a standard communication protocol and the fact that it is available almost everywhere. Remote robots controlled over the internet are called internet-based robots or internet robots. The first internet robots to be introduced were stationary remote robots operating in a limited geographical region [1-2] and, subsequently, autonomous internet robots operating in a wider region were developed [3-9].

It takes a long time to transmit large amounts of information over the internet because of its narrow bandwidth and unpredictable transmission time. Therefore, in order to control internet robots, abstract commands that do not contain detailed information about the task at hand and reduced feedback data should be used.

A virtual teaching environment was used as a teaching interface for assembly tasks in [10]. In the virtual teaching environment, the teaching motion of the operator is formulated into symbolic operation commands, which are sent to an actual robot in the form of task commands. Instead of using detailed information to control a robot, symbolic task-level commands such as “Grasp part A” are sent. In the autonomous mobile robot called Xavier described in [4], several computer systems were used for control, perception, and planning. If the operator selects a block in the two-dimensional building map

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 425–432, 2007.
© Springer-Verlag Berlin Heidelberg 2007

through a web-base interface, Xavier can visit the corresponding desired office using its autonomous navigation system. In [5], an overhead camera and an additional camera in the mobile robot supplied images to the local operator. The local operator could then provide the mobile robot with movement commands by pointing at a particular location in these images. The coordinates of the designated location were then sent to the mobile robot.

In the case of remote robots, in addition to the robot commands, the feedback information is also very important. Consequently, various data exchange techniques have been developed for the purpose of reducing the transmission time of this feedback information. In [11], fourteen magnetic sensors and one range sensor were used to track the human motion, and their data were sent to the host computer in order to render the human motion in 3D. Instead of sending all of the data, which consisted of nine double floating point numbers per sensor, a quaternion consisting of four double floating point numbers per sensor was sent to the host computer, in order to reduce the data transfer time and to improve the graphic rendering speed. In [12], the data size can be changed according to the priority, which is determined by the data content and the task condition. If the control information is more important than the image information, then the image data is divided into several parts and each part is transmitted with the whole set of control data through the network.

This paper proposes an information exchange technique for controlling a robot over the internet, which used a task-level command to reduce the time required to send commands to the robot and extracts the desired information from an image in order to reduce the time required to send the feedback information. The proposed information exchange technique is applied to a mobile robot that can be operated by either manual control or supervisory control over the internet. The manual control mode requires fast information exchange and is therefore different from the supervisory control mode.

In section 2, an internet robot system is described. In section 3, the information exchange technique for controlling this internet robot is proposed. In section 4, the operator interface of the internet robot is explained in detail. Some experiments are described in section 5. Finally, our conclusion is given in section 6.

2 An Internet Robot System

As shown in Fig. 1, a local operator can control the mobile remote robot over the internet using either the manual or supervisory control mechanisms. The local operator can use the web-based robot interface to control the remote robot. In order to send information about the tasks that result from the local operator commands, a web server at the remote site sends either camera images of the remote robot or the locations of the remote robot calculated from these images. The remote robot receives the commands from the operator over the internet and performs the desired tasks.

If the operator gives a command to the remote robot, the remote robot performs its corresponding task and the remote web server returns the result of the task to the operator. In this way, the operator can follow the status of the robot and its tasks and formulate subsequent commands based on the information provided by the web server.

3 The Information Exchange Technique

In order to improve the real-time performance in a communication network, various data exchange techniques have been developed with the goal of reducing the communication time [1, 2, 5]. We propose an efficient technique for exchanging information between the operator and the remote robot for both the manual and supervisory control modes, as shown in Fig. 1.

The proposed technique simplifies the robot commands in order to transfer them to the remote robot quickly. Since it does not take a long time to transfer these simple commands, the local operator can easily control the remote robot over the internet.

For the manual control mode, the proposed technique uses eight keyboard inputs for the robot commands employed to make the remote robot move. These keyboard inputs can adjust the velocity and direction of the remote robot, as shown in Table 1. Using these keyboard inputs, the operator can easily drive the remote robot in the "hands-on" control fashion.

For the Supervisory control mode, the proposed technique uses the mouse to select a destination point or via-points designed to make the remote robot move and issue the corresponding command messages in the form of a sequence of short messages to the remote robot.

In order to transfer the information sent by the remote robot, the proposed technique uses two kinds of feedback data, the image sequence and the robot's location on the captured image. In order to transmit the real image sequences, RTP (Realtime Transport Protocol) provided by JMF (Java Media Framework) is used, and the robot location information is transmitted by peer-to-peer socket communication.

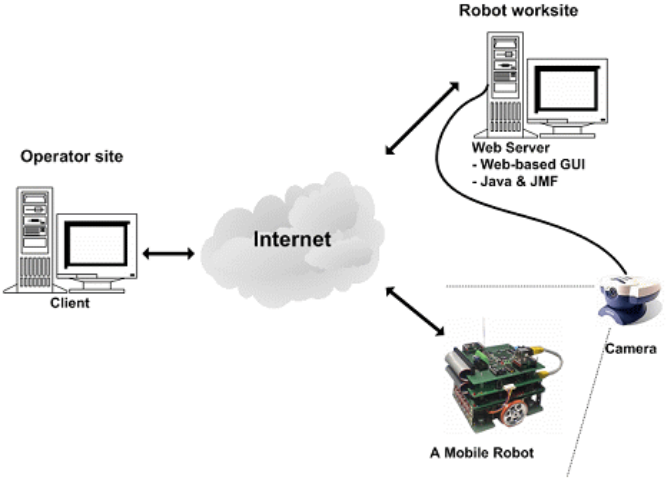


Fig. 1. An internet robot system

The robot's location can be obtained from the image sequence in real-time, while the camera is capturing the successive images. Since it may take a long time to transfer the image sequence over the internet, the robot's location can be used as complementary

information to the image sequence. As shown in Fig. 2, while the camera at the remote site is capturing the images, the robot's location can be obtained using an elliptical tracking algorithm that can track the moving object as the primitive, an ellipse, by calculating the gradient of the contour of the ellipse [13]. Then, both the robot's location and the image sequence can be transmitted over the internet. The information pertaining to the robot's location can be transmitted to the local site faster than the image sequence information, because its size is smaller than that of the image. Thus, the operator is informed of the robot's location and can control the remote robot in a timely manner, even though the image data has not yet been received.

4 Operator Interface

We developed an operator interface to allow the operator to control the remote robot easily. In order to construct the GUI, a Java applet is embedded in the web page and the JMF (Java Media Framework) is used to provide the image of the remote robot. The web server provides the GUI and the operator can access the web page used to download it in the form of a Java applet. Then, the operator can make the TCP/IP connection to the remote robot directly for the purpose of controlling the robot. The operator can monitor the robot's movement through the live images obtained from the web server.

In order to provide the operator with the necessary visual information, we use one camera and the JMF toolkit. In particular, RTP (Realtime Transport Protocol) is used to achieve better performance for the peer-to-peer connection. As shown in Fig. 3, the operator interface consists of a Virtual workspace, Vision image viewer, State message board, and Robot control panel.

The Virtual workspace displays the robot's current position. Also, the operator can select a position in the virtual workspace, in order to move the robot along the desired path and so be able to visualize the robot's movement in the simulation before sending the corresponding commands to the real robot.

The Vision image viewer shows the image of the camera connected to the web server at the remote worksite. The real image also provides the operator with important feedback information in the manual control mode, even though the images are transmitted with some delay.

The State message board shows the information of the operator interface and the communication results from the robot. The information contained in the State message board facilitates the task of the operator.

Table 1. Commands for a remote robot

Number	Acceleration	Deceleration	Description
8 (↑)	a	d	Forward movement
2 (↓)	a	d	Backward movement
4 (←)	a	d	Left turn
6 (→)	a	d	Right turn
7	a	d	Forward left turn with a radius
9	a	d	Forward right turn with a radius
1	a	d	Backward left turn with a radius
3	a	d	Backward right turn with a radius

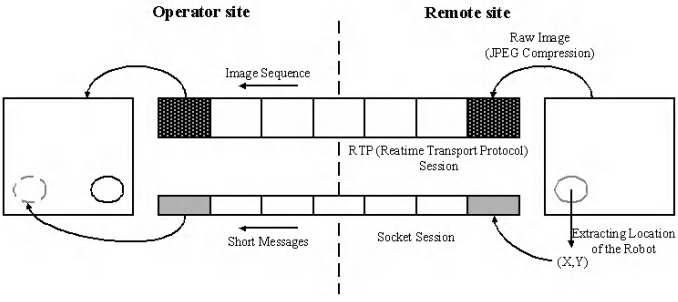


Fig. 2. Transmission of information about the remote robot

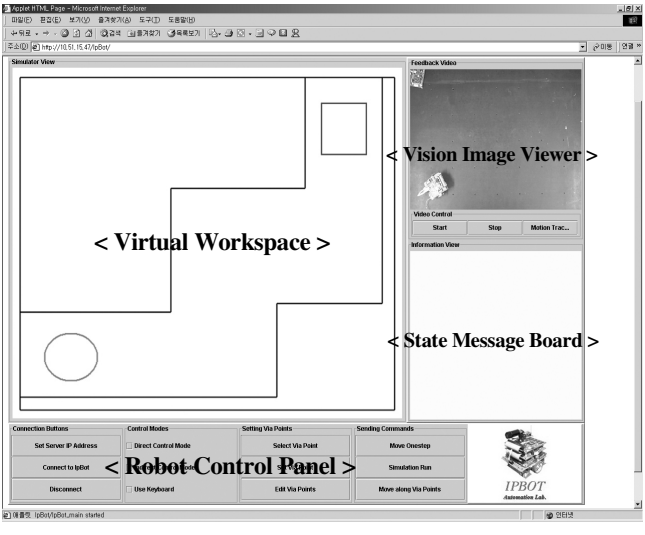


Fig. 3. Operator interface

The operator interface provides the operator with three control modes. In the Direct Control Mode, the operator can exert manual control over the robot and, thus, can control the robot's movement by using the keyboard. The operator determines the robot's position through the Vision image viewer and controls the robot by pressing the keyboard so as to adjust the robot's velocity and direction. The other two control modes are the Supervisory Control Mode and the Task Scheduling Mode. In these modes, the operator can use the mouse to indicate the subsequent via-points or a destination point. The 'Select position' button in the operator interface is used for this purpose. In the Supervisory Control Mode, the operator can make the robot move by indicating the next position and pressing the 'Move one-step' button. Then, by checking the robot's movement in the corresponding image, the operator can determine the robot's subsequent movement and continue to operate the robot. The Supervisory Control Mode is a basic control mode for Internet robot control, because it provides a very simple interface which allows the operator to move the robot in a visual fashion [1, 2, 4]. In the Task Scheduling Mode, the operator can predict the robot's movement

by indicating a set of via-points corresponding to the path to the destination and previewing the resultant robot movement by means of a simulation. There is a 'Set position' button used for storing the via-points and a 'Simulation run' button for simulating the robot movement in the virtual workspace.

Although many Internet robots have been designed which use standard communication protocols to make the connection and a simple user interface, these robots have some limitations. When a large number of users access these systems, the web server has to make a new process for each request, thereby increasing its overhead. Therefore, it is not possible to achieve an adequate level of interactivity using this method [1, 2]. In the proposed technique, the use of the Java platform allows the user interface to be more interactive and the performance to be enhanced with the standard communication protocol.

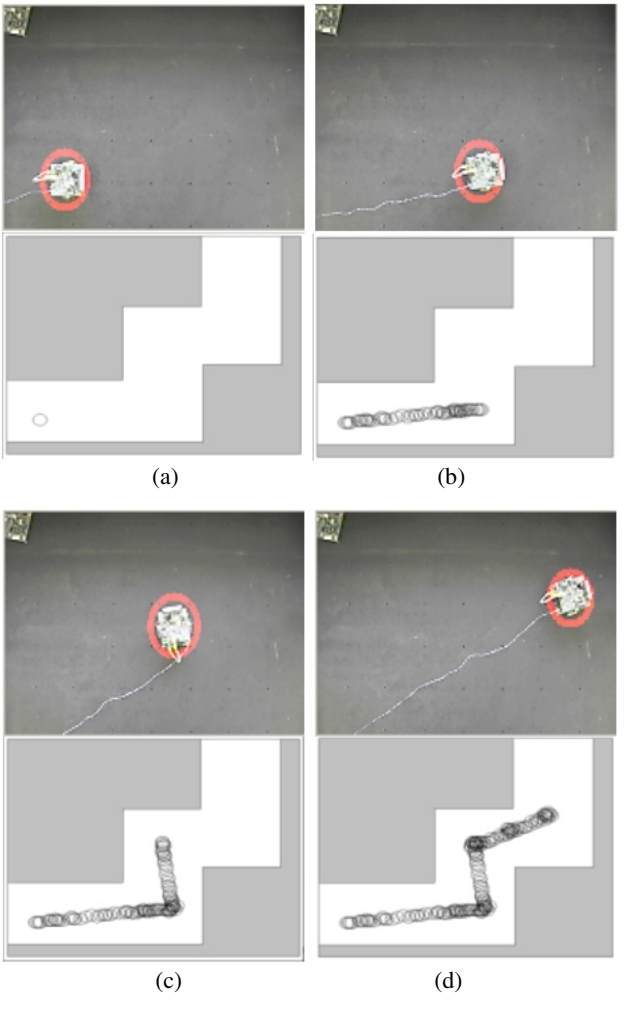


Fig. 4. Manual control of the remote robot

5 Experiments

In the system shown in Fig. 1, we performed manual control with the Internet based mobile robot by using short command messages and feedback information.

The operator can use the operator interface described in section 4 to control the robot. In the operator interface, the operator can select the 'Use Keyboard' check button to perform manual control using the eight directional keyboard inputs. In this way, the operator can easily drive the remote robot in the "hands-on" control fashion.

The image sequences that provide the feedback information are susceptible to be transmitted over the internet with some delay, depending on the environmental conditions. Even though the local and remote sites were near to each other, the image sequence sometimes froze or successive frames were not synchronized and, consequently, it was necessary to replay part of the sequence. However, the additional information sent in the form of short messages was successfully received with only a small delay and overlaid as the ellipse on the real images that represents the robot's current position. This small delay comes from the capturing of the raw image and processing of the tracking algorithm. Although the original algorithm was slightly modified before applying it to our system, it showed good performance in the tracking of the moving robot.

Fig. 4 shows the manual control of the remote robot, where the robot's location is represented by an elliptical tracker. In this mode, the additional information provided by the image sequences is useful in controlling the robot remotely over the internet.

6 Conclusion

This paper proposes an information exchange technique for controlling an internet robot, which makes use of a task-level command for the purpose of reducing the time required to send robot commands and extracts the desired information from an image in order to reduce the time required to send the feedback information. We also developed an operator interface that uses the proposed information exchange technique for controlling the internet robot easily. Using the developed operator interface, the internet robot can easily be driven by means of the manual and supervisory control modes.

Acknowledgements. This research was supported by the MIC, Korea under ITRC, IITA-2006-(C1090-0603-0046).

References

1. Ken Goldberg, Michael Mascha, Steve Gentner, Nick Rothenberg, Carl Sutter, and Jeff Wiegley. : Desktop Teleoperation via the World Wide Web. IEEE International Conference on Robotics & Automation. (1995) 654-659
2. Eric Paulos, John Canny. : Delivering Real Reality to the World Wide Web via Telerobotics. IEEE International Conference on Robotics & Automation. (1996) 1694-1699.
3. Paul G. Bakes, Kam S. Tso, and Gregory K. Tharp. : Mars Pathfinder Mission Internet-Based Operations Using WITS. IEEE International Conference on Robotics & Automation. (1998) 284-291

4. Reid Simmons, Joaquin L. Fernandez, Richard Goodwin, Sven Koenig, and Joseph O'sullivan. : Lessons Learned from Xavier. IEEE Robotics & Automation Magazine. June (2000) 33-39
5. Patrick Saucy, Francesco Mondada. : KhepOnTheWeb: Open Access to a Mobile Robot on the Internet. IEEE Robotics & Automation Magazine. March (2000) 41-47
6. Dirk Schulz, Wolfram Burgard, Dieter Fox, Sebastian Thrun, and Armin B. Cremers. : Web Interfaces for Mobile Robots in Public Places. IEEE Robotics & Automation Magazine. March (2000) 48-56
7. Sebastien Grange, Terrence Fong and Charles Baur. : Effective Vehicle Teleoperation on the World Wide Web. Proceedings of the 2000 IEEE International Conference on Robotics & Automation. April (2000) 2007-2012
8. Kuk-Hyun Han, Shin Kim, Yong-Jae Kim, Seung-Eun Lee and Jong-Hwan Kim. : Implementation of Internet-Based Personal Robot with Internet Control Architecture. IEEE International Conference on Robotics & Automation. May (2001) 217-222
9. Paul E. Rybski, Ian Burt, Tom Dahlin, Maria Gini, Dean F. Hougen, Donald G. Krantz, Florent Nageotte, Nikolas Papanikolopoulos, and Sascha A. Stoeter. : System Architecture for Versatile Autonomous and Teleoperated Control of Multiple Miniature Robots. IEEE International Conference on Robotics & Automation. May (2001) 2917-2922
10. Hiroyuki Ogata, Tomoichi Takahashi. : Robotic Assembly Operation Teaching in a Virtual Environment. IEEE Transactions of Robotics and Automation, vol 10, no. 3. (1994) 391-399
11. Tom Molet, Ronan Boulic, Serge Rezzonico, and Daniel Thalmann. : Architecture for Immersive Evaluation of Complex Human Tasks. IEEE Transactions on Robotics and Automation, vol. 15, no. 3. June (1999) 475-485
12. Takafumi Matsumaru, Syun'ichi Kawabata, Tetsuo Kotoku, Nobuto Matsuhira, Kiyoshi Komoriya, Kazuo Tanie, and Kunikatsu Takase. : Task-based Data Exchange for Remote Operation System through a Communication Network. Proceedings of the 1999 IEEE International Conference on Robotics and Automation. May (1999) 557-564
13. Stan Birchfield. : Elliptical Head Tracking Using Intensity Gradients and Color Histograms. IEEE Conference on Computer Vision and Pattern Recognition. June (1998) 232-237

A Privacy-Aware Identity Design for Exploring Ubiquitous Collaborative Wisdom

Yuan-Chu Hwang and Soe-Tsyr Yuan

Department of Management Information Systems, National Chengchi University, Taiwan
No.64, Sec.2, ZhiNan Rd., Wenshan District, Taipei City 11605, Taiwan (R.O.C)
{ychwang,yuans}@mis.nccu.edu.tw

Abstract. Privacy and security has been considered as the top criterion for the acceptance of e-service adoption. In this paper, we proposed a privacy aware identity protection design that improves the flaw in cheap pseudonym design under the limitation of ambient e-service environment. This design is being applied to a collaborative iTrust platform for ambient e-service, a new scope of mobile ad-hoc e-service. The collaborative iTrust platform highlights the collaborative power from nearby user groups to eliminate potential risk and provide appropriate estimation for trust decision in the ubiquitous environment.

Keywords: unlinkability, privacy, collaborative trust.

1 Introduction to Ambient e-Service

The notion of ambient e-services is proposed to identify a new scope of ubiquitous e-service, which addresses dynamic collective efforts between mobile users (enabled by Mobile-P2P), dynamic interactions with ambient environments (envisioned by Location-Based Services), moment of value, and low cost service provision [1]. Mobile users can exchange their information wirelessly and proceed highly extensive interactions based on the mobile P2P technology. The collective effort is based on the collaborative interactions of mobile users, which facilitate the low cost service provision. Grounded on location-based service, location information of mobile users can be acquired. Hence, ambient e-service then can provide personal, timely and relevant services to mobile users.

Comparing with the client/server design, an ambient e-service has two major distinguished features. First, it is not possible to effectively attain the collective efforts tailored to the contexts of the users with the client/server architecture. Second, the P2P design grows the number of connections by a significantly rapid pace especially in an open space.

In order to encourage users to embrace ambient e-service, privacy protection is the vital barrier. Privacy and security are very important concerns today. In ubiquitous e-services environment, who can be trusted? Are the nearby mobile users trustworthy? What do they know about me? These questions underlie the interactions and collaborations between mobile users.

Before we turn to the identity protection design, it will be useful to be aware of the natural limitations of ambient e-service environment. Ambient e-Service is presented

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 433–440, 2007.
© Springer-Verlag Berlin Heidelberg 2007

in a distributed environment, under mobile P2P ad-hoc network constraints. The mobile device is limited to lightweight computational loading with restricted storage capacity. Moreover, the identities in ambient e-service environment are localized and short-term lived.

In an ambient e-service environment, the mobile peers' composition is dynamic. Users may not stay in the environment for a long time. Mobile peers in such environment form a dynamic-organized structure. Unlike traditional e-commerce environment, users do not have permanent identities. Lack of sustained identification, the trust and reputation data for mobile agents cannot be tracked. The uncertain situation raises the risk for the disclosure of a user's detailed transaction histories to unfamiliar nearby users. How to protect user's private matters under the dynamic environment becomes a major challenge especially when users are encouraged to participate in the collaborative interactions of ambient e-service.

2 Privacy Aware Identity Protection Design

Ensuring privacy protection is one of the major incentives to encourage users to embrace the ambient e-service especially when the collective efforts of the e-services were originate from those collaborative interactions between mobile peers. Only when users are assured that their privacy or sensitive data are secured from invasion, the critical mass of the use of the ambient e-service are then unfolded.

In order to design a privacy aware identity protection system for ambient e-service, some critical issues should be considered:

- **Unlinkability:** Fulfill the untraceable demands for all participants within the ambient e-service environment. Make sure user's identities are not linkable to the real personal identities. Ensure all users within the e-service environment are not able to recognize other's true identity via their involved transaction histories.
- **Tolerance of the limitations of ambient e-service environment**: The identity protection design should create a fair reputation evaluation mechanism that values user's private matters seriously given the limitations of ambient e-service environment (i.e., capabilities of device computation and storage) while retaining users to participate.

A simple way to achieve unlinkability is pseudonymity, which replaces a user's true identities with a pseudonym used in user interactions. In this paper, a multiple layered pseudonym method is designed for interaction under ambient e-service environment. Before any interactions can be executed, peers require an identity for the service environment. We use the Interaction Pseudonym as an agent identity for the user. It should be noted that a user may possess several different agents for various e-services. An agent's identity is produced according to the service. A user can activate an agent identity or discard a specific identity based on their needs. Even if the user has many agent identities, all identities still share the same global reputation data. When an identity is created, it inherits the concurrent reputation from user's global reputation data. The reputation data for each identity does not exist separately. No matter how many identities belong to the user, he can keep only one global reputation. The diagram (Fig. 1) represents the general design concepts and the relationships of

three kinds of pseudonyms. Only the Interaction Pseudonyms appear in the interaction environment. Interaction pseudonyms are generating through the same Active Pseudonym but without any linkage relationship. Interaction pseudonyms are cost-free (i.e. cheap pseudonyms); user can generate/discard them freely. However, user can not change their active pseudonym without cost. The algorithm of generating interaction pseudonyms is then shown in Table 1.

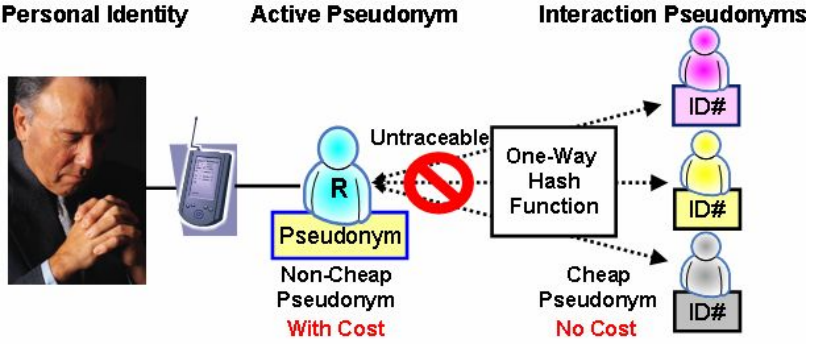


Fig. 1. General idea and relationships between the pseudonyms

The Active Pseudonym represents a pseudonym that entails that the original identity must be responsible for some of the switching cost. As mentioned in section 1, the identities' life-time is short owing to the ambient e-service's nature. However, as addressed in [3], reputation systems generally take on three properties: (1) entities are long-lived; (2) feedback about current interactions is captured and distributed; (3) past feedback guides buyer decisions. Consequently, a conceptual long-lived identity for ambient e-service is indispensable. Opposite to the cheap pseudonyms that allow users to discard cheaply pseudonyms arbitrarily, the non-cheap pseudonym accept user to resume new pseudonyms but charges replacement expense. The non-cheap pseudonym is to encourage users to address themselves on just and honorable behaviors.

The interaction pseudonyms are separated from one another without any linkage to their previous interaction pseudonyms or their personal identities. However, due to the inheritance process, their past behaviors can be ranked by a simplified reputation which prevents the defects of cheap pseudonyms such as whitewashing. (e.g., malicious user may discard their pseudonym, and register a. new pseudonym with a fresh reputation.).

Moreover, a one-way hash function is used to isolate the interaction pseudonyms and the active pseudonym. One-way hash function is an algorithm that turns messages or text into a fixed string of digits, usually for security or data management purposes. The "one way" means that it's nearly impossible to derive the original text from the string. It is also hard to find two strings that would produce the same hash value. As the diagram indicates, even if someone wants to trace a certain person, the only identity can be recognized is his Interaction Pseudonym which is an unlinkable short-term lived identity.

Whenever a user enters the ambient e-service environment, a short-lived interaction pseudonym is created by Anonymizer module. It inherits the briefed reputation

and assigned role attributes. Because of the time stamp, the interaction pseudonym is localized and temporary valid within the service environment. Once its user leaves the service environment, the interaction pseudonym expires. Even when using the same device, a different interaction pseudonym is assigned to the user when they return to the service environment.

The goals of using our pseudonym design for ambient e-service interactions are three folds: (1).Exclude a unique personal pseudonym for interactions to protect users from possible tracking and profiling. (2). Use multiple interaction pseudonyms to enhance the complexity of identity tracing. (3).Abstract the design of role/relationship pseudonyms for service version selection and delivery. (i.e. Versioning the services by specific types for performance consideration)

Table 1. Interaction Pseudonym Generation Algorithm

Algorithm: Generation of an Interaction Pseudonym based on the Active Pseudonym
Input: *AP*: active pseudonym string *T*: time stamp *H*: one-way hash function Output: *IP*: encrypted interaction pseudonym with preference attributes and inherited reputation data
begin 1. Get selected active pseudonym string *(AP)* 2. Form an Pseudonym Message*(PM)* by mixing *AP* with time stamp*(T)* 3. Hash the Pseudonym Message to an Encrypted Pseudonym*(IP)*, *H(PM)* → *IP* 4. Assign attributes and preference data to *IP* 5. return *IP* end

A relevant work of the gradational pseudonym structure was proposed by Pfitzmann and Hansen [2]. They defined the differences between five pseudonym types and classified them into four categories. Their design utilized the classes of pseudonyms for increasing available anonymity and realizing unlinkable transactions. However, their design assumptions were grounded in the concept of centralized and permanent identity management and thus cannot be applied to the ambient e-service environment.

3 Collaborative iTrust Platform: A Sample Application

This section exemplifies an application of our identity protection design for ambient e-service so as to the unlinkability of one's personal identity. This application is unfolded with a collaborative iTrust platform that utilizes the privacy aware identity protection design to assure the properties of safety and fairness required for ambient e-Service. The collaborative iTrust platform (shown as Figure 2) includes several function modules. Due to the page limitation, we will elaborate the identity related core modules as follows.

- **Profile Management:** In iTrust platform, mobile users can manage their profile settings through a friendly user interface. The profiles includes their preferences and the roles they would like to play, and various attributes such as user's willingness to participate, the will to disclose their interaction experiences, the risk level they can tolerate, and the reliability threshold for determining whether to interact with nearby peers. Once an identity has been generated, those settings will be assigned to the interaction pseudonym automatically.

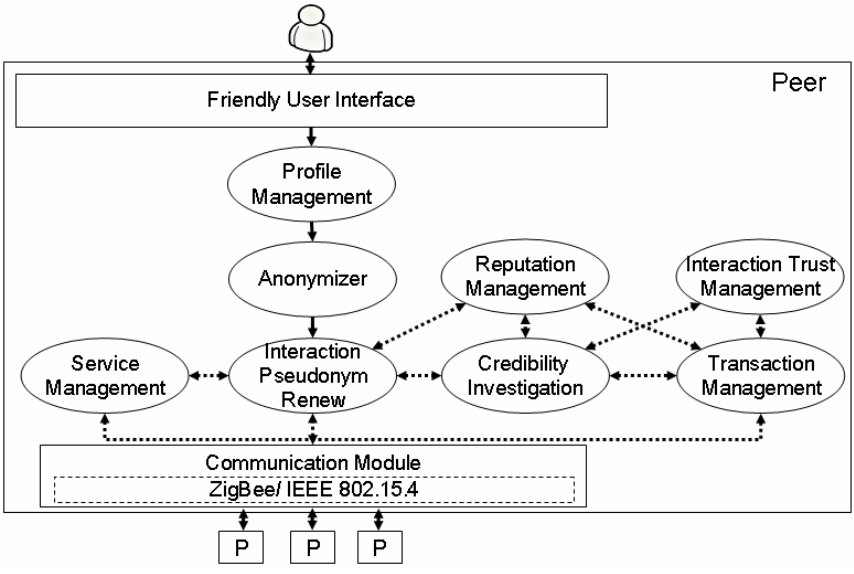


Fig. 2. Macro view of the Collaborative iTrust platform

- **Anonymizer:** All interactions within the ambient e-service environment are using the "Interaction Pseudonyms" instead of user's real identity or personal pseudonyms. The main function for the Anonymizer is to generate diverse occasional interaction pseudonyms based on their given identity for various kinds of e-services. Those interaction pseudonyms are valid for a short period, and are localized to the corresponding e-service acquired. Because the randomized interaction pseudonyms are not linked to real personal identities and are valid for a limited range, others will be unable to trace their real owner via the interaction pseudonyms. Those interaction pseudonyms are generated by the Anonymizer and will inherit the attribute parameters automatically through the Profile management module. They are able to execute the versioning process and cope with the service management module to reduce irrelevance transmission and improve the efficiency of interaction.
- **Interaction Pseudonym Renew:** The iTrust design has overcome the problem of the dynamic composition of surrounding peers that may change rapidly. The Interaction Pseudonym Renew module is used to update the list of current nearby users, which exhibits all available nearby peer interaction pseudonyms. Users can interact with peers around themselves through the Communication module. The Interaction

Pseudonym Renew module is connected with the Reputation Management module, which may immediately update the global reputation of peers so that all devices in range may access it. Each exchange and transmission within the Service Management module, as well as information inquiry when performing creditability investigation, is targeted to those identities obtained by Interaction Pseudonym Renew module.

- **Communication Module:** The ZigBee based communication module makes use of the security services that are already present in the 802.15.4 security specification. ZigBee infrastructure security includes network access control, integrity of packet routing, and prevention of unauthorized use of packet transport. ZigBee application data security includes message integrity, authentication, freshness, and privacy.[6]

In short, the aforementioned components are tightly coupled with the privacy aware identity protection design for ambient e-service security. However, the capabilities of the collaborative iTrust platform also depend on the functions of additional components (related to the confidential reputation design). The confidential reputation design is intend to modify current public reputation system design and try to separate the detailed transaction history from the reputation data. This separation aims to ward off the bias reputation estimation from vindictive act of users holding up each other or flattering each other to falsify good comments. The confidential reputation design rests on blind-signed scheme that keeping the evaluation targets from being aware of the referee's identity. With this platform design, we expect that users will respond their faithful opinions during their transaction interactions within the ambient e-service environment. Moreover, not only the transaction history can be treated as the data source for trust estimation, the user's direct interaction experience is also considered. Cooperating with the creditability investigation module, the platform provides heterogeneous information based on the user's experiences among one's social network for estimating unfamiliar users.

4 Significance and Contribution

Our privacy-aware identity design concept has several beneficial results: (1) Encourage fair transactions that prevent bad users from avoiding punishment. In other words, the design assures bad users cannot regenerate their new identity without cost. (2) For those good users who devote their energies to maintain good reputations, they can accumulate their good reputation and gain some rewards (i.e. get special discount or better chance for transaction) due to the reputation inheritance. This will encourage users to maintain their reputations. With the application of the identity design, the collaborative iTrust platform accordingly is empowered to realize the vision of ambient e-service in terms of the following perspectives:

- **Deliberation of short-term lived pseudonyms:** Revises existing long term identity design concepts and ensures the unlinkability of identities.
- **Distributed data process consideration:** Each interaction record within the ambient e-service environment relies on the computational loading and data storage in the mobile device instead of the centralized server database.

- **Lightweight consideration:** Different from the existing works of centralized gradational pseudonym design (the quantity of interaction data will be expanded in exponential growth), our design method integrates a user's conceptual role information and relationship information into an abstract public information attached to the user's identity. This attached abstract information can then provide hints to filter out inadequate service information that will reduce unnecessary data transmission. Moreover, the adoption of the time stamp design is exerted to omit those overdue historical interaction data, which will further improve the strength of interaction pseudonym's unlinkability.
- **Convenience requirement:** Under the versioning scheme, irrelevant service information is filtered out. Only highly correlated services are delivered to the requester. The versioning design reduces communication costs and system loading while improving service efficiency.

A pilot study of user perceived protection involved 38 college students, 24 students male and 14 female. All students had at least two years experience in network communications and understand the potential privacy invasion in an internet environment. The participants were divided into two groups 19 each. Each group was shown a future shopping mall scenario in an ambient e-service environment. We explained that users had to perform shopping activities in a risky environment that contained 50% honest and 50% malicious users. The operational interface we showed was the same for both groups, except for the availability of the identity protection design mechanism. Group 1 had a traditional permanent identity, under which users themselves can determine whether to reveal personal information to others. Group 2 was given the multi-layered identity protection design that allows users to use multiple identities for various interactions. Based on the measures for perceived control in [4] and [5], we asked participants about the perceived power of information control and helplessness during privacy protection tasks. 68.4% of the participants in Group 1 reported that they would prefer to have some advanced protection mechanism for their identities. 84.2% of Group 2 participants reported that they feel their identities are effectively protected by the identity protection design. For power of information control, no significant difference between two groups (80.2% and 78.9%) was observed. Participants felt that they had the power of information control.

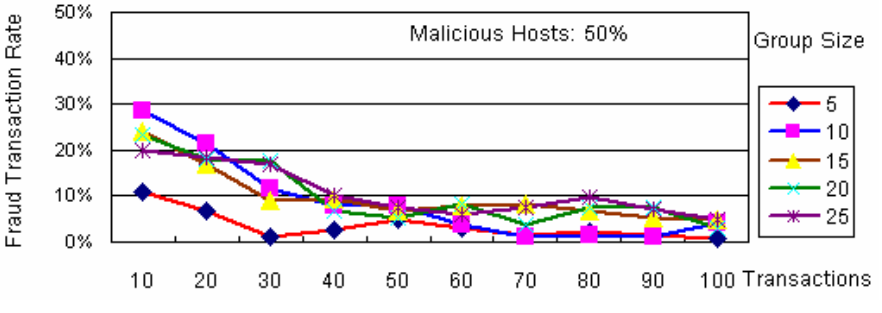


Fig. 3. Evaluation results on decreasing fraud transition rate

In the ambient e-service environment it is lack of information available for users to estimate which user is trustworthy. This problem is more serious when a new market is opened since the global reputation of each identity is zero and may not satisfy the user's trustworthiness threshold. This will lead to a desolate e-service environment since users may be afraid to transact with unfamiliar users. As the number of transactions increases, more interaction experience is stored in the environment. For the decision quality of the overall collaborative iTrust platform, we develop a prototype system for simulation. Up to 100 transactions were executed by users with different group sizes. In testing the trustworthiness performance, we have considered practical factors such as the population of malicious hosts and good hosts and the available peer numbers for simulating realistic conditions of applications. Figure 3 then shows the evaluation results for collaborative trust measured as successfully decreased fraud transaction rate using our proposed iTrust platform. Even for the risky environment of 50% malicious hosts, the collaborative trust quickly decreases the presence of malicious transactions. For heterogeneity information sources, the risk level for interact with unfamiliar peer is still significantly decreased when applying the collaborative iTrust platform.

5 Conclusion and on Going Works

This paper proposes a privacy aware identity protection design that specifies how the identity protection structure works to resolve the defect of privacy invasion problems for existing e-commerce designs. Furthermore, we also brief a collaborative iTrust platform that exerts the identity design to deliver the visions of ambient e-service with an integrated consideration of trust, reputation and privacy requirements. The identity design aims to encourage the users' participation with the concept of unlinkability (guaranteeing the protection of a user's private matters in the ambient e-service environment). We have implemented the identity design and the device platform, and our on-going works include the evaluation of the design and the platform from different perspectives (trust, reputation, privacy, efficiency, usability, etc.). Preliminary evaluation results indicate the design suits the privacy protection requirement in an ambient e-service environment.

References

1. Y. Hwang and S. Yuan. A Roadmap for Ambient E-Service: Applications and Embracing Model. *International Journal of E-Business Research*. 3(1): 51-73. 2007.
2. A. Pfitzmann and M.Hansen. Anonymity, unlinkability, unobservability, pseudonymity, and identity management--- a consolidated proposal for terminology. http://dud.inf.tu-dresden.de/anon-terminology. August, 2005.
3. P. Resnick, R.Zeckhauser, E. Friedman and K. Kuwabara. Reputation systems: Facilitating trust in Internet interactions. *Communications of the ACM*. 43(12): 45-48. 2000.
4. B. Rohrmann. Empirische Studien zur Entwicklung von Antwortskalen für die sozialwissenschaftliche Forschung. *Zeitschrift für Sozialpsychologie*. 9: 222-245. 1978.
5. S. Spiekermann. Perceived Control: Scales for Privacy in Ubiquitous Computing Environments. *Online Proceedings of the 10th International Conference on User Modeling*. Edinburgh, Scotland. July, 2005.
6. ZigBee Org. ZigBee Specification Ver. 1.0, http://www.zigbee.org. 2005.

Performance Comparison of Sleep Mode Operations in IEEE 802.16e Terminals*

Youn-Hee Han[1], Sung-Gi Min[2], and Dongwon Jeong[3]

[1] School of Internet-Media, Korea University of Technology and Education, Cheonan, Korea
yhhan@kut.ac.kr
[2] Department of Computer Science and Engineering at Korea University, Seoul, Korea
sgmin@korea.ac.kr
[3] Department of Informatics and Statics, Kunsan National University, Gunsan, Korea
djeong@kunsan.ac.kr

Abstract. IEEE 802.16e Task Group has developed an IEEE 802.16 amendment supporting mobile terminals such as PDAs, phones, and laptops. The energy efficiency design is an important research issue for mobile terminals due to limited battery power of them. In this paper, we study the energy efficiency design of IEEE 802.16e. For the purpose of power efficient operation, there are two sleep mode operations related to normal data traffic in the current IEEE 802.16e specification. We propose analytical models of the two sleep mode operations and conduct their performance comparison in terms of the power consumption and the frame response time.

Keywords: IEEE 802.16e, Sleep Mode.

1 Introduction

Broadband wireless access systems use base stations (BSs) to provide broadband multimedia services to business or homes as an alternative to wired ′last-mile′ access links that use fiber, cable, or digital subscriber lines. A promising advance of such systems is the IEEE 802.16 standard, which provides fixed-wireless access between the subscriber station and the Internet service provider (ISP) [1,2]. The IEEE 802.16 standardization is also expected to support quality of service (QoS) for real time application such as voice over IP. For such QoS support, it defines four different scheduling services: Unsolicited Grant Service (UGS), Real-Time Variable Rate (RT-VR), Non-Real Time Variable Rate (NRT-VR), and Best Effort (BE) services [3].

Whereas the existing IEEE 802.16 standards address fixed wireless applications only, IEEE 802.16e standard [4] aims to serve the needs of fixed, nomadic,

* This work was supported by the Korea Research Foundation Grant funded by the Korean Government (MOEHRD) [KRF-2006-331-D00539].

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 441–448, 2007.
© Springer-Verlag Berlin Heidelberg 2007

and fully mobile networks. It enhances the original standard with mobility so that Mobile Subscribe Stations (MSSs) can move during services. It adds mobility support to the original standard so that MSSs can move while receiving services. Mobility of MSSs leads the power consumption aware design, which is one of the primary objectives due to limited computation power of MSSs [5]. In order to get the solution with the lowest power consumption, optimization over all aspects of system implementation must be employed, including the algorithms, architectures, circuit design, and manufacturing technology. In this paper, we are interested in reducing power consumption for IEEE 802.16e MSSs in terms of MAC algorithms.

In the current IEEE 802.16e specification, three sleep mode operations, named Power Saving Class (PSC) 1, 2, and 3, are defined for the purpose of power efficient operations. Among the three PSCs, PSC 1 and 2 are related to data traffic, while PSC 3 serves for management procedure such as periodic raging. While PSC 1 serves data traffic carried in NRT-VR or BE connection, PSC 2 serves UGS or RT-VR data flow. In [6] and [7], the authors proposed the novel models to investigate the energy consumption in IEEE 802.16e. However, they evaluated only PSC 1 and did not perform any analysis about PSC 2 and PSC 3. In this paper, we propose analytical models of PSC 2 as well as PSC 1 and conduct their performance comparison in terms of their power consumption and frame response time. From this analytical study, we can get the features of the sleep mode operations in detail and obtain some hints about the way to determine right values of the system parameters used in sleep mode operation.

This paper is organized as follows. In Section 2, we represent the current operational rules of the two sleep mode operations, PSC 1 and 2, based on the IEEE 802.16e specification. In Section 3, we give an analysis about PSC 1 and 2. Section 4 conducts the performance evaluation and reveals the features of the operations, and concluding remarks are given in Section 5.

2 Sleep Mode Operations in IEEE 802.16

An IEEE 802.16 MSS has two modes: wakeup mode and sleep mode. The sleep mode is defined for the purpose of power efficient operations. Before entering the sleep mode, an MSS in the wake mode sends the sleep request message to a serving BS. After receiving the sleep response message which notifies whether the sleep request of the MSS is approved or not, the MSS can enter the sleep mode. In the sleep response message, the following relevant parameters are included: the start frame number for first sleep window, the minimum sleep interval, the maximum sleep interval, and the listening interval. The intervals are presented as units of MAC frames. After getting approval, the MSS goes into the sleep mode, gets sleep for an interval, and then wakes up to check whether there are frames for it. If there are frames for the MSS, it goes to the wakeup mode. Otherwise, the MSS is still in the sleep mode and gets sleep for another interval. The MSS keeps performing the above procedure until it goes to the wakeup mode.

In the first power saving mode, PSC 1, sleep intervals are defined as follows. At the first sleep interval, the minimum sleep interval is used. Then each sleep interval is doubled until the maximum sleep interval is reached, and then the sleep interval keeps the maximum sleep interval. After each sleep interval, the MSS temporarily wakes up a short interval, called listening interval, to listen the traffic indication message broadcasted from the serving BS, and the message includes information about MSSs to whom the BS has frames waited. Furthermore, the MSS can terminate the sleep mode if there is an out-going frame, mostly because of the user's manual interaction.

IEEE 802.16 supports four QoS supports: UGS, RT-VR, NRT-VR, and BE. Each scheduling service is characterized by a mandatory parameter set of QoS parameters, which is adjusted to best describe the guarantees required by the applications that the scheduling service is designed for. UGS is designed to support real-tme applications with strict delay requirements. UGS is free from any contention of frame transmission. The BS provides fixed size data grants at periodic intervals to the UGS flows. UGS can be used for constant bit-rate (CBR) service. RT-VR and NRT-VR flows are polled through the bandwidth request polling. RT-VR is designed to support real-time applications with less stringent delay requirements. The supported applications may generate variable-size data packets at periodic intervals, such as MPEG video and VoIP with silence suppression. RT-VR flows is prevented from using any contention requests. While RT-VR can be used for real-time service, NRT-VR can be used for non-real time service such as bandwidth-intensive file transfer. The main difference between RT-VR and NRT-VR is that NRT-VR connections are reserved a minimum amount of bandwidth. Finally, BE is for applications with no rate or delay requirements.

PSC 1's behavior well fits the behavior of demand created by random (or bursty) IP traffic, like WEB browsing. Such traffic typically is carried in NRT-VR or BE connection. But demand created by UGS and in some cases of RT-VR connections may have different pattern. For UGS (and RT-VR) flow, PSC 2 is recommended to be used. PSC 2 defines the different sleep interval from PSC 1's one. At the first sleep interval, the minimum sleep interval is used. However, the size of all subsequent sleep intervals is the same as the initial one. As a result, the definition of sleep mode for an MSS has been extended.

3 Analytical Models

In this paper, we are interested in the sleep mode operation affected by frames delivered from network to MSSs. For the model of incoming traffic flow, we use a simple *train model*. A frame train is defined in [8] as a burst of frames arriving from the same source and heading to the same destination. If the spacing between two frames exceeds some inter-train time, they are said to belong to different trains. For both PSC 1 and 2, we assume that the train arrival processes follow a Poisson distribution with the same rate λ. That is, UGS (or RT-VR) traffic and NRT-VR (or BE) traffic arrive at the same rate on the average. Accordingly, the inter-train times follow an exponential distribution with the mean $1/\lambda$ (unit time).

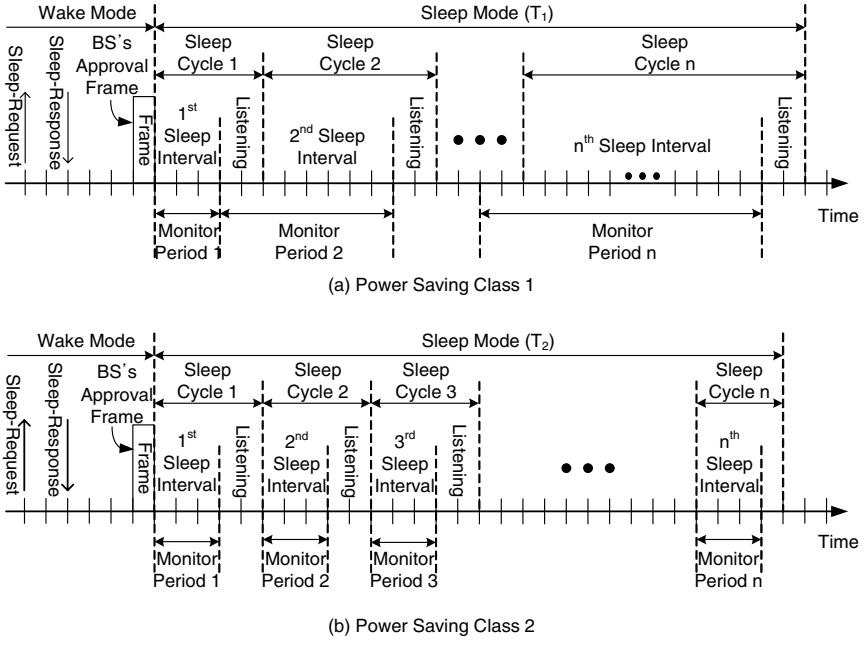


Fig. 1. Sleep Mode Operations of IEEE 802.16e

We will use T_1 and T_2 to respectively denote the sleep mode interval of PSC 1 and 2. A sleep mode interval includes one or more sleep cycles, which are illustrated in Fig. 1 (a) and (b). A sleep cycle includes a sleep interval and a listening interval. It is assumed that the same listening interval L is used for both PSC 1 and 2. Let t_1^1 and t_2^1 denote the initial sleep intervals of PSC 1 and PSC 2, respectively.

Let t_1^i denote the length of MSS's i-th sleep interval in PSC 1. The maximum sleep interval is $t_{max}(= 2^M t_1^1)$ where M is the number of increments when the sleep window size reaches t_{max}. Then, t_1^i is defined as follows:

$$t_1^i = \begin{cases} 2^{i-1} t_1^1 & if\ 1 \le i < M \\ t_{max} & if\ i \ge M. \end{cases} \tag{1}$$

Similarly, let t_2^i denote the length of MSS's i-th sleep interval in PSC 2. For every i, according to the specification of PSC 2, it is simply defined as follows:

$$t_2^i = t_2^1. \tag{2}$$

In PSC 1, L is only used for sending the traffic indication message to MSSs and enabling MSS synchronization with BS. So, the transmission of traffic arrived during a sleep interval begins at the end of the following listening interval L. When there is no traffic addressed to an MSS during a sleep interval and frames arrive in the listening interval immediately following the sleep interval, these

packets are buffered and the traffic indication message for these packets will be notified at the next listening interval. In PSC 2, as opposite to PSC 1, the transmission of traffic arrived during a sleep interval begins at the end of the sleep interval. That is, during L, an MSS can receive (or send) any frames from (or to) the serving BS without any exchange of notification. Based on these assumptions, we also define the monitor period to be the time interval when the frames arrived at BS should be buffered and released after i-th sleep cycle. For PSC 1 and PSC 2, the monitor periods are denoted by s_1^i and s_2^i, respectively. They are given by

$$s_1^i = \begin{cases} t_1^1 & if\ i = 1 \\ L + t_1^i & if\ i \geq 2. \end{cases} \tag{3}$$

$$s_2^i = t_2^i\ =\ t_2^1. \tag{4}$$

Let n_1 denote the number of sleep cycles before an MSS goes to the wake mode in PSC 1. In PSC 1, the current sleep mode turns to the wake mode when a frame train arrives during a monitor period. Let e_1^i denote the event that there is the train arrival for an MSS at a BS during the monitor period i of PSC 1. Then, we have

$$Pr(e_1^i = true) = 1 - e^{-\lambda s_1^i}. \tag{5}$$

$$Pr(n_1 = 1) = Pr(e_1^1 = true) = 1 - e^{-\lambda s_1^1} = 1 - e^{-\lambda t_1^1}. \tag{6}$$

and for $k \geq 2$

$$\begin{aligned} Pr(n_1 = k) &= Pr(e_1^1 = false \wedge e_1^2 = false \wedge \cdots \wedge e_1^{k-1} = false \wedge e_1^k = true) \\ &= \prod_{i=1}^{k-1} Pr(e_1^i = false) \cdot Pr(e_1^k = true) \\ &= e^{-\lambda \sum_{i=1}^{k-1} s_1^i} \cdot \left(1 - e^{-\lambda s_1^k}\right) = e^{-\lambda \left((k-1)L + \sum_{i=1}^{k-1} t_1^i\right)} \cdot \left(1 - e^{-\lambda (L + t_1^k)}\right). \end{aligned} \tag{7}$$

Let n_2 denote the number of sleep cycles before MSS goes to the wake mode of PSC 2. Unlike PSC 1, PSC 2's sleep mode can turn to the wake mode when a frame train arrives during a listening interval or when it arrives during a monitor period. Similarly with Equations (5)-(7), we have

$$\begin{aligned} Pr(n_2 = k) &= e^{-\lambda (k-1)(t_2^{k-1} + L)} \cdot \left(1 - e^{-\lambda (t_2^k + L)}\right) \\ &= e^{-\lambda (k-1)(t_2^1 + L)} \cdot \left(1 - e^{-\lambda (t_2^1 + L)}\right). \end{aligned} \tag{8}$$

Let P_S and P_L denote the power consumption units per a unit of time in the sleep interval and the listening interval, respectively. Assuming PC_1 denote the power consumption during the PSC 1's sleep mode interval. Then, assuming that $E[\cdot]$ stands the average function, we can get the average power consumption $E[PC_1]$ as follows:

$$E[PC_1] = \sum_{k=1}^{\infty} Pr(n_1 = k) \cdot \sum_{i=1}^{k} (t_1^i P_S + LP_L)$$
$$= (1 - e^{-\lambda t_1^1})(t_1^1 P_S + LP_L)$$
$$+ \sum_{k=2}^{\infty} e^{-\lambda\left((k-1)L + \sum_{i=1}^{k-1} t_1^i\right)} \cdot \left(1 - e^{-\lambda(L + t_1^k)}\right) \cdot \sum_{i=1}^{k} (t_1^i P_S + LP_L). \quad (9)$$

Let R_1 denote the response time of a train's first frame in PSC 1. In our paper, like [6], the response time represents the amount of time required for data packets to be delivered to an MSS after they are buffered at a BS. Since train arrivals follow Poisson distribution, the arrival events are random observers [9] to the sleep intervals. Therefore, we have

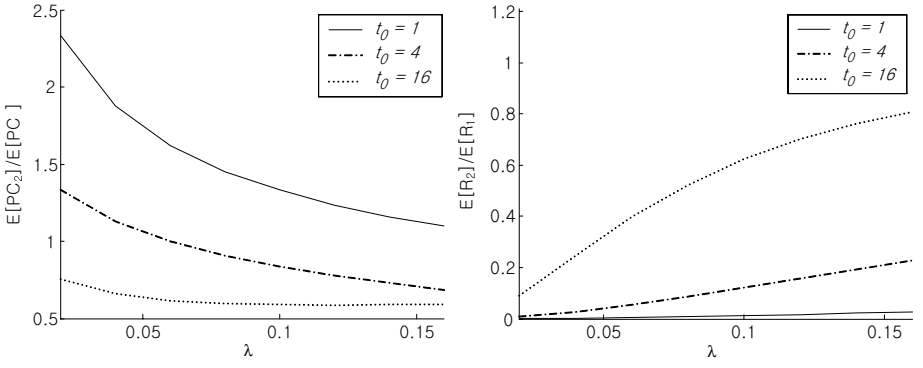
$$E[R_1] = \sum_{k=1}^{\infty} Pr(n_1 = k) \cdot \frac{s_1^k}{2}$$
$$= (1 - e^{-\lambda t_1^1}) \frac{t_1^1}{2} + \sum_{k=2}^{\infty} e^{-\lambda\left((k-1)L + \sum_{i=1}^{k-1} t_1^i\right)} \cdot \left(1 - e^{-\lambda s_1^k}\right) \frac{s_1^k}{2}. \quad (10)$$

The PSC 2's last sleep cycle can finish without the full listening interval since PSC 2 allows MSS and BS to exchange frames even during the listening interval. For getting $E[T_2]$ exactly, therefore, it is required to model the last sleep cycles minutely. Again, the last sleep cycle can end up with a sleep interval plus the one among the following two cases: 1) no listening interval if a frame train arrives during the last monitor period, and 2) an interval time ($< L$) when a frame train arrives after the last monitor interval. In second case, for simplicity, we just assume that the arrival time of a frame train is distributed uniformly on the listening interval L. So, we can get the average power consumption $E[PC_2]$ during the PSC 2's sleep mode interval as follows:

$$E[PC_2] = \sum_{k=1}^{\infty} Pr(n_2 = k) \cdot \left(\sum_{i=1}^{k-1} (t_2^i P_S + LP_L) + (1 - e^{-\lambda s_2^k}) \cdot t_2^k P_S + e^{-\lambda s_2^k} \cdot \left\lceil \frac{L}{2} \right\rceil P_L \right)$$
$$= \left(1 - e^{-\lambda(t_2^1 + L)}\right) \cdot \sum_{k=1}^{\infty} e^{-\lambda(k-1)(t_2^1 + L)}$$
$$\cdot \left((k-1)(t_2^1 P_S + LP_L) + (1 - e^{-\lambda t_2^1}) \cdot t_2^1 P_S + e^{-\lambda t_2^1} \cdot \left\lceil \frac{L}{2} \right\rceil T_L \right). \quad (11)$$

Let R_2 denote the response time of a train's first frame in case of PSC 2. If the first frame arrives when MSS is in a sleep interval, it should be buffered at the serving BS until the sleep interval comes to end. Such buffering makes the response time long. Otherwise (the first frame arrives when MSS is in a listening interval), it does not cause the response time to be extended. Therefore, we have

$$E[R_2] = (1 - e^{-\lambda t_2^1}) \cdot \frac{t_2^1}{2} + e^{-\lambda t_2^1} \cdot 0 = (1 - e^{-\lambda t_2^1}) \cdot \frac{t_2^1}{2}. \quad (12)$$



(a) $E[PC_2]/E[PC_1]$ with regard to λ (b) $E[R_2]/E[R_1]$ with regard to λ

Fig. 2. Results of Performance Evaluation

4 Performance Evaluation

In the previous work [7], the author also conducted simulation for PSC 1's sleep mode operation and validated his analytic results with the simulation results. In this paper, we did not conduct a simulation for verification of our analysis. But, we assert that our analytical results are also reliable since our analytical method is based on [7]'s one and we extend its research results. For our analysis, the following default parameters are used: $L = 1$, $t_{max} = 1024$ (thus, $M = 10$), $P_S = 1$, and $P_L = 10$. We also assume that two initial sleep intervals of PSC 1 and PSC 2 are set to the same value t_0 (that is, $t_1^1 = t_2^1 = t_0$).

An interesting result is shown from Fig. 2 (a). Since PSC 1 has mostly longer sleep intervals than PSC 2, at a first glance, it might also seem that PSC 1's energy efficiency is generally better than PSC 2's one. However, PSC 2's energy efficiency is always better than PSC 1's one if t_0 is high (e.g., $t_0 = 16$). Such energy gain is furthermore almost insensitive to λ (the values of PC_2/PC_1 are distributed around 0.6 in case of $t_0 = 16$). These are justified by the fact that PSC 1's listening interval is only used for sending the traffic indication message to MSSs, while an MSS can exchange a frame with its serving BS during PSC 2's listening interval. So, the last sleep cycle of PSC 2 can finish without the full listening interval. If initial sleep interval is so long, a frame train is very likely to arrive in the first or second sleep cycle.

Of a practical interest is the case of $t_0 = 4$. We can see PSC 2's energy efficiency is just similar to PSC 1's one when $t_0 = 4$. If $\lambda > 0.06$, it is even better than PSC 1's one. It is because that a high λ is also likely to make a frame train arrive in the first or second sleep cycle. If λ is very high, therefore, a very small value of t_0 (e.g., $t_0 = 1$) is a not bad choice, which also guarantees a very fast frame response time.

Fig. 2 (b) also provides the comparison about the response time of a train's first frame. It shows that the dependence of the response time from the values

of λ is marginal if the values of t_0 are small (e.g., $t_0 = 1$ or 4). We also see that PSC 2's reduction of the response time is remarkably enormous with such small values of t_0. When $t_0 = 4$, PSC 2 achieves more than 80% of response time reduction. When $t_0 = 1$, the reduction is distributed even between 95% and 99%. This result is of fundamental reason for PSC 2 to be developed for UGS (or RT-VR) traffic.

5 Conclusions

In this paper, we studied the details of two sleep mode operations, PSC 1 and PSC 2, presented by the recent specification of IEEE 802.16e. We also presented analytical models of PSC 1 and PSC 2 to compare them in terms of the two performance metrics:The power consumption ($E[PC_1]$ and $E[PC_2]$) and the response time of a train's first frame ($E[R_1]$ and $E[R_2]$).

This paper has focused how much PSC 2 reduces the response time of a train's first frame. We also studied how much it sacrifices the energy efficiency for reducing the response time. From our analysis results, we can conclude that the small values of t_0 (e.g., $t_0 = 1$ or 4) can make PSC 2 reduce the response time greatly (up to 80% $\sim$ 99%) at not much expense of the energy efficiency. Moreover, if λ is very high, very small values of t_0 does not much sacrifice the energy efficiency while preserving a very low frame response time.

References

1. Std. 802.16-2004, I.: IEEE 802.16 Local and Metropolitan Area Networks - Part 16: Air Interface for Fixed Broadband Wireless Access Systems (2004)
2. Eklund, C., Marks, R.B., Stanwood, K.L., Wang, S.: IEEE Standard 802.16: A Technical Overview of The WirelessMAN Air Interface for Broadband Wireless Access. IEEE Communications Magazine **40**(6) (June 2002) 98–107
3. Chu, G., Wang, D., Mei, S.: A QoS Architecture for The MAC Protocol of IEEE 802.16 BWA System. In: IEEE International Communications, Circuits and Systems and West Sino Expositions. Volume 1. (July 2002) 435–439
4. Std 802.16e 2005, I., Std 802.16-2004/Cor 1-2005, I.: IEEE Standard for Local and Metropolitan Area Networks Part 16: Air Interface for Fixed and Mobile Broadband Wireless Access Systems Amendment 2 (February 2006)
5. Xiao, Y., Chen, C.L.P., Kinateder, K.J.: An Optimal Power Saving Scheme for Mobile Handsets. In: Sixth IEEE Symposium on Computers and Communications (ISCC'01). (July 2001) 192–197
6. Seo, J.B., Lee, S.Q., Park, N.H., Lee, H.W., Cho, C.H.: Performance Analysis of Sleep Mode Operation in IEEE 802.16e. In: IEEE 60th Vehicular Technology Conference (VTC2004-Fall). Volume 2. (Sept. 2004) 1169–1173
7. Xiao, Y.: Energy Saving Mechanism in The IEEE 802.16e Wireless MAN. IEEE Communications Letters **9**(7) (2005) 595–597
8. Jain, R., Routhier, S.A.: Packet Trains - Measurements and A New Model for Computer Network Traffic. IEEE Journal on Selected Areas in Communications **4**(6) (1986) 986–995
9. Ross, S.M.: Stochastic Processes. Wiley (1995)

Performance Evaluation of the Optimal Hierarchy for Cellular Networks

So-Jeong Park[1,*] ,Gyung-Leen Park[1,**], In-Hye Shin[1], Junghoon Lee[1], Ho Young Kwak[2], Do-Hyeon Kim[2], Sang Joon Lee[2], and Min-Soo Kang[3]

[1] Department of Computer Science and Statistics, Cheju National Univ., Korea
{joypark,glpark,ihshin76,jhlee}@cheju.ac.kr
[2] Faculty of Telecommunication and Computer Engineering, Cheju National Univ., Korea
{kwak,kimdh,sjlee}@cheju.ac.kr
[3] Department of Information and Communication Engineering, Hanyangcyber Univ., Korea
mskang@hycu.ac.kr

Abstract. Reducing the location update cost has been a critical research issue since the location update process requires heavy signaling traffics. One of the solutions is employing the hierarchical structure in cellular networks to reduce the location update cost. This paper is not only to measure the average location update cost in cellular networks using the hierarchical structure but also to determine the degree of the hierarchy which minimizes the location update cost. The paper proposes an advanced analytical model to obtain the optimal hierarchy in cellular networks, considering the update rates of the two registers, HLR and VLR, as well as the update delays for them. Also, the paper provides the threshold values as a guideline for the network administrator to design the optimal hierarchical structure in cellular networks.

1 Introduction

Many researchers have studied on diverse design problems of the location management schemes in cellular networks. The location management means the task finding the location of the mobile station (MS) in order to receive the seamless services properly while the MS moves [1-3]. The cellular networks has two types of databases, the home location register (HLR) and the visitor location register (VLR). The mobile switching center (MSC) associated with a specific VLR is in charge of several base station controllers (BSCs), lower control entities which in turn control several base stations (BSs). The MSCs are connected to the backbone wired network such as public switching telephone network (PSTN). The network coverage area is divided into smaller cell clusters called location areas (LAs). The VLR stores temporarily the service profile of the MS roaming in the corresponding LA while the HLR stores permanently the user profile and points to the VLR associated with the LA where the user is currently located.

* This research was supported by the MIC (Ministry of Information and Communication), Korea, under the ITRC support program supervised by the IITA (IITA-2006-C1090-0603-0040).
** The corresponding author.

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 449–456, 2007.
© Springer-Verlag Berlin Heidelberg 2007

The total location update cost consists of the update cost of HLR and that of VLR. Since the HLR is connected to many VLRs, it has heavy signaling traffics [4-6]. The super location area (SLA) [7], a group of LAs, has been proposed for hierarchical networks to reduce the location update cost of the HLR. The paper develops an analytical model to obtain the optimal size of SLA which minimizes the total location update cost which includes both of the update rates of the HLR and VLR as well as the update delays for them. Also, the paper provides threshold values as a guideline that allows network administrators to design an optimal hierarchical structure for cellular networks depending on the given environments.

2 The Location Management Using the Hierarchical Structure

An example of the architecture of the scheme is depicted in Figure 1 [7].

Each SLA consists of 7 LAs and again each LA has 7 cells as shown in Figure 1. Each MSC/VLR pair covers an SLA. When an MS enters a new LA, the LA could belong to either a different SLA or the same SLA. If the LA belongs to a different SLA, both of the VLR and the HLR are updated to record the LA where the MS is located. On the other hand, if the new LA belongs to the same SLA, only the VLR is updated to record the new LA. The conventional scheme, which doesn't employ any hierarchy, should update not only the VLR but also the HLR whenever an MS enters a new LA regardless of any SLA. The previous research [7] could not answer the questions like "What is the optimal number of levels in the hierarchy when designing cellular networks?". The next section proposes an analytic model to answer the questions.

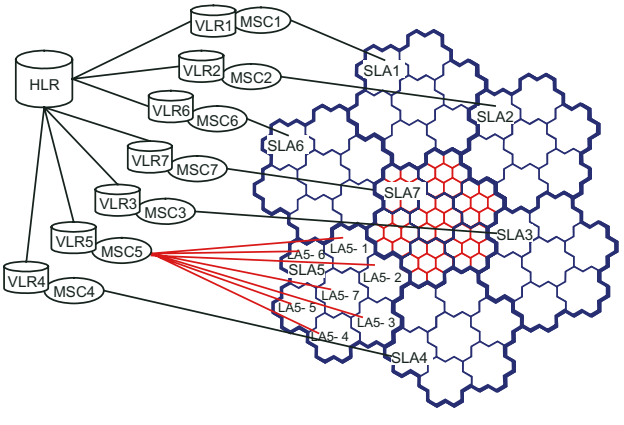


Fig. 1. The cellular architecture using hierarchical structure

3 The Proposed Analytical Model

This section evaluates the total update rate of the VLR and that of the HLR as well as the total update delay of the VLR and that of the HLR.

The followings assumptions are made in the model.

1. The service area is divided into hexagonal cells of the equal size.
2. A mobile user moves independently to one of neighboring cells with the uniform distribution.
3. The dwell time in any cell for a mobile user is an exponentially distributed random variable with the average value, $\overline{T_d}$.
4. The update cost between a cell and a neighboring cell is 1.

The notations used in the proposed model are depicted in Table 1.

Table 1. The Notations Used in the Analytic Model

$\overline{K}$	The average number of mobile users in a cell		
K_{MS}	The total number of mobile users in the whole network		
d	The size of an LA	d_s	The size of an SLA
$\overline{T_d}$	The average dwell time of a mobile user		
N_c	The number of cells in an LA : $3d^2-3d+1$		
N_{bc}	The number of boundary cells in an LA : $6(d-1)$		
N_{Sla}	The number of LAs in an SLA : $3{d_s}^2-3d_s+1$		
N_{Sc}	The number of cells in an SLA : $(3d^2-3d+1)(3{d_s}^2-3d_s+1)$		
N_{SbLA}	The number of boundary LAs in an SLA : $6(d_s-1)$		
$\overline{R_{LA}}$	The average location update rate per mobile user in an LA		
R_{VLR}	The total location update rate of VLR		
$\overline{R_{SLA}}$	The average location update rate per mobile user in an SLA		
R_{HLR}	The total location update rate of the HLR		
d_T	The size of the whole cellular network		
$[d_T]$	The greatest positive number not larger than d_T		
K_{SLA}	The number of SLAs in the whole network		
K_{bSLA}	The number of SLAs which does not form complete rings in the whole network		
l_{LA}	The distance between the centers of two neighboring LAs		
r	The average distance from the center of an SLA to the corresponding ring		
l_{SLA}	The distance between the centers of two neighboring SLAs		
l_{SLAl}	The longest distance among the center of an SLA and the boundary cells in the SLA		
$\overline{D_{VLR}}$	The average location update delay of the VLR		
$\overline{D_{HLR}}$	The average location update delay of the HLR		
C	The total location update cost in whole network		

In Figure 2, the LA consists of 19 cells. The boundary LAs of the SLA are classified into 6 vertex LAs and 12 side LAs. A mobile user may cross the border (a boundary cell) with probability of 1/6, 2/6, or 3/6 according to the location of the cell.

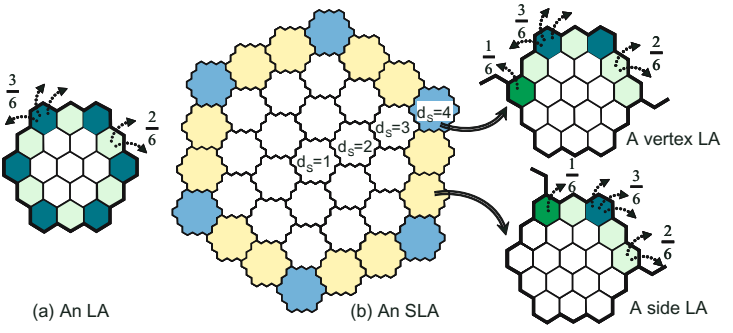


Fig. 2. (a) The Structure of an LA (b) The Structure of an SLA

The location update of the VLR occurs when a mobile user crosses the boundary cells of LA. Thus the total update rate of the VLR is obtained by Equation (2) which is derived from Equation (1) in [7], as shown in Figure 2-(a).

$$\overline{R_{LA}} = 6 \times \left\{ \frac{2}{6} \times \{6(d-2)\} + \frac{3}{6} \times 1 \right\} \times \overline{K} \times \frac{1}{\overline{T_d}} = (2d-1) \cdot \overline{K} \cdot \frac{1}{\overline{T_d}} \tag{1}$$

$$R_{VLR} = K_{SLA} \times \left(N_{Sla} \times \overline{R_{LA}} \right) = \frac{K_{MS}}{N_{Sc} \times \overline{K}} \times \left(N_{Sla} \times (2d-1) \times \overline{K} \times \frac{1}{\overline{T_d}} \right) = \frac{(2d-1)}{\left(3d^2 - 3d + 1\right)} \cdot K_{MS} \cdot \frac{1}{\overline{T_d}} \tag{2}$$

When a mobile user moves into another SLA, the location update of the HLR happens in the boundary cells in the current SLA. Consequently, the total update rate of the HLR is obtained by Equation (4) using Equation (3) considering the vertex LAs and the side LAs, as shown in Figure 2-(b).

$$\overline{R_{SLA}} = 6 \left[(d_s - 2) \left\{ \frac{2}{6} \times 6(d-1) + 1 \right\} + \left\{ \frac{3}{6} 6(d-1) + 1 \right\} \right] \times \overline{K} \times \cdot \frac{1}{\overline{T_d}} = (2d_s - 1)(2d-2) \cdot \overline{K} \cdot \frac{1}{\overline{T_d}} \tag{3}$$

$$R_{HLR} = K_{SLA} \times \overline{R_{SLA}} = \frac{(2d_s - 1)(2d-1)}{\left(3{d_s}^2 - 3d_s + 1\right)\left(3d^2 - 3d + 1\right)} \cdot K_{MS} \cdot \frac{1}{\overline{T_d}} \tag{4}$$

If a VLR is located at the center of an SLA, the update delay of the VLR can be regarded as the distance between the center of an SLA and the boundary cell of the SLA where the update of the VLR happens. In Figure 3, the shadowed parts, where the update of VLR occurs, are the boundary cells of LAs. The SLA forms three rings which connect the center of each LA. The rings are depicted as Ring1, Ring2, and Ring3 in the figure. Then, the average update delay of the VLR is the average distance between the center of the SLA and the corresponding ring.

The average distance is obtained by calculating the radius of a circle. The distance, l_{LA} is obtained by Equation (5) because $\triangle ABC$ is a regular triangle, where $ACB = 60^{\circ}$, $\overline{AC} = \overline{AB} = \overline{BC} = (2d-1)$, and $\overline{CD} = (d-1)$.

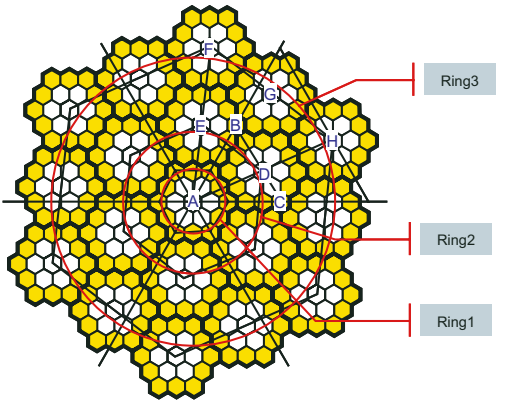


Fig. 3. The Architecture of an SLA with d_s=3

$$l_{LA} = \overline{AD} = \sqrt{\overline{AC}^2 + \overline{CD}^2 - \overline{AC} \times \overline{CD}} = \sqrt{3d^2 - 3d + 1} \tag{5}$$

Therefore the average distance from the center of the SLA to the corresponding ring is obtained by Equation (6). The average update delay of the VLR, consequently, is obtained by Equation (7).

$$r = \sqrt{\frac{3\sqrt{3}}{2\pi}} \times \overline{FH} = \sqrt{\frac{3\sqrt{3}}{2\pi}} \times (d_s - 1) \times l_{LA} \tag{6}$$

$$\begin{aligned} \overline{D_{VLR}} &= \frac{1 \times \sqrt{\frac{3\sqrt{3}}{2\pi}} \cdot (d-1) + 6\sqrt{\frac{3\sqrt{3}}{2\pi}} \cdot l_{LA} + 12\sqrt{\frac{3\sqrt{3}}{2\pi}} \cdot 2 \cdot l_{LA}}{N_{Sla}} + \cdots + \frac{6(d_s - 1)\sqrt{\frac{3\sqrt{3}}{2\pi}} \cdot (d_s - 1) \cdot l_{LA}}{N_{Sla}} \\ &= \frac{\sqrt{\frac{3\sqrt{3}}{2\pi}} \cdot (d-1) + \sqrt{\frac{3\sqrt{3}}{2\pi}} \cdot \sqrt{3d^2 - 3d + 1} \cdot \left(2d_s^{\,3} - 3d_s^{\,2} + d_s\right)}{3d_s^{\,2} - 3d_s + 1} \end{aligned} \tag{7}$$

Assume that HLR is centrally located in the network because the HLR is logically only one in the whole network.

In Figure 4, $\triangle ABC$ is a regular triangle and ACB is 60°. $\overline{AC}$ and $\overline{CD}$ are obtained by Equation (8) and (9), respectively. Consequently, the distance between SLA1 and SLA2, l_{SLA}, is obtained by Equation (10).

$$\begin{aligned} \overline{AC} &= \overline{AE} + \left(\overline{FG} + \overline{HI}\right) - \left(\overline{JK} + \overline{LM}\right) + \left(\overline{NO} + \overline{PQ}\right) + \overline{RD} \\ &= (d-1) + (d_s - 1)(2d-1) - (d_s - 1)(d-1) + (d_s - 1)(2d-1) + d = 3dd_s - d - d_s \end{aligned} \tag{8}$$

$$\begin{aligned} \overline{CD} &= \left(\overline{EF} + \overline{GH} + \overline{IJ}\right) + \left(\overline{KL} + \overline{MN}\right) + \left(\overline{OP} + \overline{QR}\right) \\ &= d_s(d-1) + (d_s - 1)(d-1) = 3dd_s - 2d - 2d_s + 1 \end{aligned} \tag{9}$$

$$l_{SLA} = \overline{AD} = \sqrt{\overline{AC}^2 + \overline{CD}^2 - \overline{AC} \times \overline{CD}} = \sqrt{9d^2 d_s{}^2 - 9d^2 d_s - 9dd_s{}^2 + 9dd_s + 3d^2 + 3d_s{}^2 - 3d - 3d_s + 1} \quad (10)$$

The size of the whole cellular networks is obtained by Equation (11) because the SLAs are arranged in rings.

$$d_H = \frac{1}{2} + \sqrt{\frac{1}{4} - \frac{1}{3}(1 - K_{SLA})} \quad (11)$$

Outer cells in the network cannot complete the ring for the SLA configuration. The largest size of a complete ring is the greatest positive number not larger than d_H. The number of SLAs which do not complete their outer rings is in Equation (12).

$$K_{bSLA} = K_{SLA} - \left(3[d_H]^2 - 3[d_H] + 1\right) \quad (12)$$

As shown in Figure 4, if $d_H = 1$ then the average updating delay of HLR is the average distance between the center and the boundary cell in the central SLA. $\triangle AST$ is a regular triangle and ATI is 60°. $\overline{AT}$ and $\overline{IT}$ are obtained in Equation (13) and (14), respectively. Therefore the longest distance, l_{SLA1}, is obtained by Equation (15).

$$\overline{AT} = \overline{AE} + \left(\overline{FG} + \overline{HI}\right) = (d-1) + (d_s - 1)(2d - 1) = 2dd_s - d - d_s \quad (13)$$

$$\overline{IT} = \left(\overline{EF} + \overline{GH}\right) = (d_s - 1)(d - 1) = dd_s - d - d_s + 1 \quad (14)$$

$$l_{SLA1} = \overline{AI} = \sqrt{\overline{AT}^2 + \overline{IT}^2 - \overline{AT} \times \overline{IT}} = \sqrt{3d^2 d_s{}^2 - 3d^2 d_s - 3dd_s{}^2 + 2dd_s + d^2 + d_s{}^2 - d - d_s + 1} \quad (15)$$

Consequently, the average update delay of HLR is obtained by Equation (16).

$$\begin{aligned} \overline{D_{HLR}} &= \frac{\sqrt{\frac{3\sqrt{3}}{2\pi}} \times l_{SLA1} + 6\sqrt{\frac{3\sqrt{3}}{2\pi}} \times l_{SLA} + 12\sqrt{\frac{3\sqrt{3}}{2\pi}} \times 2 \times l_{SLA}}{K_{SLA}} + \cdots \\ &+ \frac{6([d_H]-1)\sqrt{\frac{3\sqrt{3}}{2\pi}} \times ([d_H]-1) \times l_{SLA} + K_{bSLA} \times \sqrt{\frac{3\sqrt{3}}{2\pi}} \times [d_H] \times l_{SLA}}{K_{SLA}} \\ &= \frac{\sqrt{\frac{3\sqrt{3}}{2\pi}} \times l_{SLA1} + \sqrt{\frac{3\sqrt{3}}{2\pi}} \times l_{SLA} \times \left(2[d_H]^3 - 3[d_H]^2 + [d_H]\right)}{K_{SLA}} + \frac{K_{bSLA} \times \sqrt{\frac{3\sqrt{3}}{2\pi}} \times [d_H] \times l_{SLA}}{K_{SLA}} \end{aligned} \quad (16)$$

Finally, the total location update cost is obtained by Equation (17).

$$C = R_{VLR} \times \overline{D_{VLR}} + R_{HLR} \times \overline{D_{HLR}} \quad (17)$$

The more details of the derivations are omitted due to the space limit. Interested readers may refer to [8].

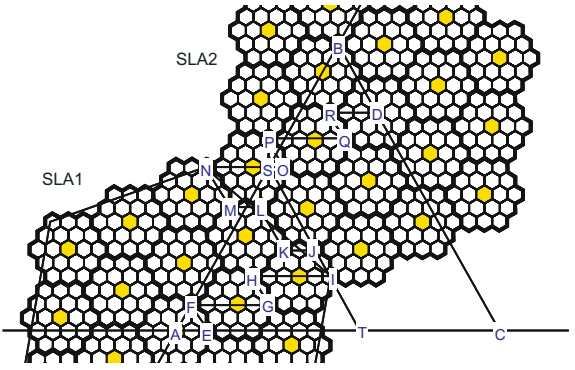


Fig. 4. The central SLA and a neighboring SLA

4 The Results of Performance Evaluation

This section shows the results of performance evaluation in terms of the total location update cost calculated using the model developed in the previous section.

Figure 5-(a) shows the total update cost according to the dwelling time while Figure 5-(b) does according to the size of an LA. The optimal size of an SLA is 7 irrespective of dwelling time as shown in Figure 5-(a). Figure 5-(b) shows that the larger the LA size is, the smaller the optimal size of an SLA become. The optimal size of an SLA is 14, 7, and 5 when the LA size=1, 3, and, 5, respectively.

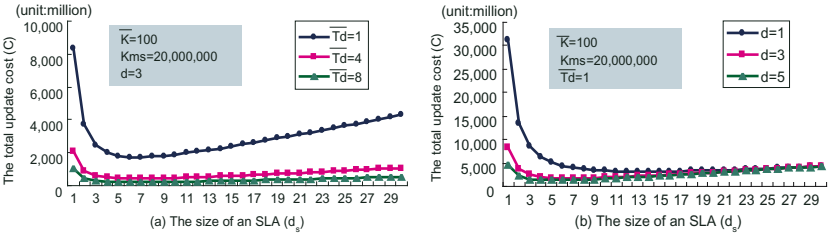


Fig. 5. The total update cost according to the dwelling time (a) and the LA size (b)

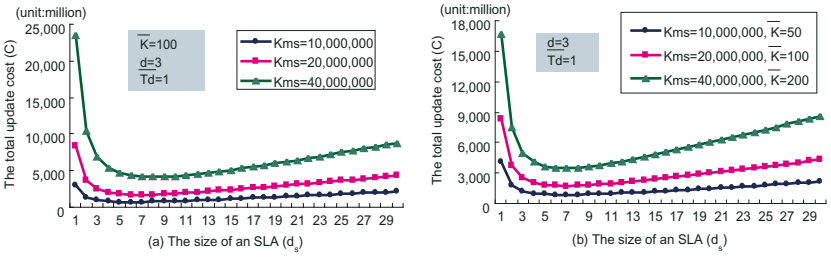


Fig. 6. The total update cost according to K_{MS} (a) and K_{MS} with same number of cells (b)

Figure 6-(a) shows that the larger the number of the total mobile users is, the larger the optimal size of an SLA is. The optimal size of an SLA is 6, 7, and 8 when K_{MS} = 10, 20, and 40 billion, respectively. Figure 6-(b) shows a change of the total update cost according to the total mobile users in the same number of cells (K_{MS} / $\overline{K}$ =200,000) in the network. The optimal size is constantly 7 in all cases.

5 Conclusion

The paper proposed an analytical model to obtain the optimal size of the SLA in cellular networks, considering the update rates of the HLR and that of the VLR as well as the update delays for them. The result obtained from the model shows that the number of cells in the whole cellular network affects sensitively the optimal size of SLA. Also, the model provides the threshold values as a guideline when network administrators aim to design the optimal hierarchical structure for cellular networks.

References

1. Kyamakya, K., Jobmann, K.: Location Management in Cellular Networks: Classification of the Most Important Paradigms, Realistic Simulation Framework, and Relative Performance Analysis. The IEEE Vehicular Technology Conference, Vol. 54. (2005) 687-708
2. Ali, S.Z.: Location Management in Cellular Mobile Radio Networks. IEEE International Symposium on PIMRC, Vol. 2. (2002) 745-749
3. Assouma, A.D., Beaubrun, R., Pierre, S.: A Location Management Scheme for Heterogeneous Wireless Networks. IEEE International Conference on Wireless and Mobile Computing, Networking And Communication, Vol. 2. (2005) 51-56
4. Kim, S., Smari, W.W.: Reducing Location Management Costs in Hierarchical-based Model of Mobile and Wireless Computing Systems. IEEE International Conference on Information Reuse and Integration, (2003) 428-435
5. Morris, D., Aghvami, A.H.: A Novel Location Management Scheme for Cellular Overlay Networks. IEEE Transaction on Broadcasting, Vol. 52. (2006) 108-115
6. Fan, G., Zhang, J.: A Multi-layer Location Management Scheme that Bridges the Best Static Scheme and the Best Dynamic Scheme. IEEE International Conference on Mobile Data Management, (2004) 757-760
7. In-Hye Shin, Gyung-Leen Park.: On Employing Hierarchical Structure in PCS Networks. ICCSA (2) (2003) 155-162
8. So-Jeong Park, Gyung-Leen Park.: An Analytical Model for Performance Evaluation of Hierarchical Structure in Cellular Networks. Tech. Report #06-122, Cheju National University, Korea.

Channel Time Allocation and Routing Algorithm for Multi-hop Communications in IEEE 802.15.3 High-Rate WPAN Mesh Networks*

Ssang-Bong Jung, Hyun-Ki Kim, Soon-Bin Yim, and Tae-Jin Lee

School of Information and Communication Engineering,
Sungkyunkwan University, Suwon 440-746, Korea
{jssbong,hyunki,sbyim,tjlee}@ece.skku.ac.kr

Abstract. IEEE 802.15.3 High-rate Wireless Personal Area Networks (WPANs) have been developed to communicate with devices within 10m at high speed. A mesh network made up of a parent piconet and several child piconets can support multi-hop communications. In this paper, we propose an efficient Channel Time Allocation (CTA) method and routing algorithm for multi-hop communications in IEEE 802.15.3 high-rate WPAN mesh networks. The proposed CTA allocation method allows sufficient CTA allocation for the relay PNCs. The proposed routing algorithm is tree-based, and the routing tables of parent PNCs and child PNCs are created by efficient device discovery process to find the shortest paths via parent PNCs or subordinate child PNCs. We evaluate the performance of the proposed algorithms via simulations. The proposed algorithms are shown to extend the communication range by multi-hop exchange of packets and to provide sufficient Channel Time Allocations (CTAs) for parent and child PNCs.

Keywords: IEEE 802.15.3 WPAN, multi-hop, mesh network, routing.

1 Introduction

Recently, a lot of personal devices (DEVs) have emerged, and they are often required to interact wirelessly to make information accessible and to exchange data without physical cables. Wireless Personal Area Networks (WPANs) can interconnect DEVs located around person's comfortable workspace wirelessly. WPAN is a developing communication technology to allow short-range wireless ad-hoc connectivity among portable consumer electronics and communication DEVs. Although Bluetooth has introduced as the first WPAN technology [1], it provides the data rate of about 1Mbps, which is limited in the sense it is hard to support high-rate or real-time multimedia traffic [2].

A new high-rate WPAN standard has been developed by IEEE 802.15.3 task group (TG) [3]. The target applications of the high-rate WPAN can be divided into two categories. The first one is multi-megabyte file transfer such as images and mp3 files. The second one is distribution of real-time video and high-quality audio. IEEE 802.15.3

* This research was supported by the Ministry of Information and Communication, Korea, under the ITRC IITA-2006-(C1090-0603-0046) and by grant No.R01-2006-000-10402-0 from the Basic Research Program Korea Science and Engineering Foundation of Ministry of Science & Technology. Dr. Tae-Jin Lee is the corresponding author.

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 457–465, 2007.
© Springer-Verlag Berlin Heidelberg 2007

high-rate WPAN [3] operates in the unlicensed 2.4GHz ISM band, and supports the data rate of up to 55Mbps and ad-hoc mode allowing a number of DEVs to communicate with one another in a piconet. A piconet in IEEE 802.15.3 high-rate WPAN consists of several DEVs and one Piconet Coordinator (PNC). If a DEV needs channel time on a regular basis, it asks the PNC for isochronous channel time. A mesh network can be formed by connecting piconets to extend coverage and communication area. As a result, it enhances network reliability and improves network throughput [4]. A mesh network in IEEE 802.15.3 high-rate WPAN consists of a parent piconet and its child piconets. The child PNC of a child piconet is a member of the parent piconet and it also operates as the PNC in the child piconet. A new child piconet is constructed under an established piconet, and the established piconet then becomes the parent piconet. So a child piconet is dependent on the parent piconet [5] and requires channel time allocation to the parent piconet and is synchronized with the parent piconet's timing.

In the current IEEE 802.15.3 high-rate WPAN specification, efficient routing algorithms for multi-hop mesh networks are not defined and only basic frame structure for the formation of a single channel mesh network is defined. Therefore, in this paper, we propose an efficient WPAN mesh network routing algorithm, Parent PNCs and Child PNCs, for multi-hop communications and evaluate its performance. Our proposed mesh network routing algorithm is a tree-based routing algorithm. The routing tables of parent PNCs and child PNCs are created by efficient device discovery process to find the shortest paths via parent PNCs or subordinate child PNCs. And, we also propose CTA allocation mechanism for time division of superframes. We analyze device discovery time and conduct simulations to evaluate its performance. Simulation results validate the desired features of the proposed routing algorithms for multi-hop communications.

The rest of the paper is organized as follows. Section 2 gives an overview of the IEEE 802.15.3 high-rate WPAN Media Access Control (MAC) and mesh networks. In Section 3, we describe our proposed channel time allocation and routing for WPAN mesh networks. Section 4 provides the performance analysis of the proposed algorithm for multi-hop communications. Finally, Section 5 concludes the paper.

2 IEEE 802.15.3 High-Rate WPAN MAC and Mesh Network

An IEEE 802.15.3 piconet consists of a PNC and more than one DEV. A piconet is managed by the PNC, and supports ad-hoc peer-to-peer connection. The PNC provides the basic timing for the piconet with beacons and it allows independent DEVs to communicate one another. The channel time allocation for data transmission in a piconet is given in the superframes. Each superframe begins with a beacon, and Contention Access Preiod (CAP), and Channel Time Allocation Period (CTAP) follow. The beacon is used to set the timing synchronization and to communicate management information for the piconet, and it is generated and broadcast by the PNC. The CAP is used to exchange association request and response command, channel time request and response command, and asynchronous data. During the CAP, DEVs try to access the channel in a distributed fashion using Carrier Sense Multiple Access/Collision Avoidance (CSMA/CA). The CTAP is composed of channel time allocations (CTAs), and management CTAs (MCTAs), which allows TDMA-like medium access.

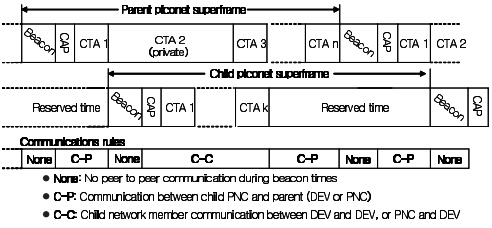


Fig. 1. Parent piconet and child piconet superframe relationship

IEEE 802.15.3 high-rate WPAN allows a DEV of a piconet to request the formation of a subsidiary piconet. The original piconet is referred to as the parent piconet and the subsidiary piconet is referred to as a child or a neighbor piconet [6]. We call a mesh network as a set of piconets consisting of a parent piconet and its child piconets. Fig. 1 illustrates the relationship between the superframe of the parent piconet and that of the child piconet. The None period denotes that there is no peer-to-peer communications during beacon times. The C-P period denotes communications between a child PNC and a DEV of the parent PNC or the PNC of the parent, and the C-C period denotes communications between child piconet DEV and DEV, or child piconet PNC and DEV. A child PNC is a member of the parent piconet and thus it can exchange data with any DEV in the parent piconet. A parent piconet and its child piconets operate under the same frequency channel. A mesh network can be formed for multi-hop communications with the aid of child PNCs and the parent PNC. How to form a WPAN mesh networks [7] and to route packets in the network are some of the most important issues. We focus on efficient routing of WPAN mesh networks.

3 Proposed Channel Time Allocation and Routing for WPAN Mesh Networks

A mesh network made up of a parent piconet and its child piconets incurs multi-hop communications. Therefore, we need efficient CTA allocation method and routing algorithm for multi-hop communications among DEVs. In this section, we present a mechanism for DEV discovery to create routing tables. Then we present our proposed a CTA allocation method and a routing algorithm. We assume that the PNC of a parent piconet and those of child piconets know the information on the DEVs of the parent piconet and the child piconets, respectively.

3.1 Device Discovery to Form Routing Tables

In this section, we present a method to create routing table. Fig. 2 shows a mesh network example with two levels (L) and routing tables of a parent piconet and child piconets. The parent piconet and child piconets use distinct PNIDs. The parent piconet uses PNID 0 when $L = 0$. The number of child piconets when $L = 1$ is two and the child piconets use PNID 1 and 2. The number of child piconets when $L = 2$ are three and the child

piconets use PNID 3, 4, and 5. The parent piconet has to know all the PNIDs of the child piconets and the topology under the PNC to make a routing table. When a child piconet is formed, the PNC of the child piconet informs its member DEV Identifiers (DEVIDs) and PNID to its parent. The PNC that receives the descendents information pass it all the way up to the root parent PNC (PNC with PNID 0), and the PNC then creates the routing table. In this way, a tree-shape PNC hierarchy is constructed (see Fig. 3). Therefore, the root parent piconet can have all the PNIDs. Child piconets only have to know PNIDs of its subordinate child piconets.

Bandwidth allocation for the parent piconet and its child piconets has been present in [7]. To explain the DEV discovery for routing table formation and the CTA allocation for a DEV or PNC, we define the following notations:

- S : Maximum superframe size
- T_B : Beacon time
- T_{CAP} : Channel access period
- T_{CTA} : Channel time allocation
- N_p : Number of DEVs in the parent piconet
- L : Number of levels
- N_{PNID} : Number of PNIDs
- n_i : Number of DEVs in level i, $\quad i = 1, 2, \cdots, L$
- m_i : Number of child piconets in level i, $\quad i = 1, 2, \cdots, L$
- R^i_{CTA1} : CTA allocation remainder to one piconet in level i, $\quad i = 1, 2, \cdots, L$
- $T^i_{S_size}$: Available superframe size for level i in mesh network, $\quad i = 1, 2, \cdots, L$
- $T^i_{Discovery}$: Device discovery time in level i, $\quad i = 1, 2, \cdots, L$
- $T^i_{C_size}$: Available CTA size of child piconets in level i, $\quad i = 1, 2, \cdots, L$
- $C^0_{P_CTA}$: Available CTA of parent PNC with PNID *0*
- $C^j_{D_CTA1}$: Available CTA of a DEV with PNID j, $\quad j = 1, 2, \cdots, N_{PNID}$
- $C^j_{C_CTA}$: Available CTA of child PNCs with PNID j, $\quad j = 1, 2, \cdots, N_{PNID}$
- Sn_j : Number of source DEVs with PNID j, $\quad j = 0, 1, \cdots, N_{PNID}$
- Rn_j : Number of relay PNCs with PNID j, $\quad j = 0, 1, \cdots, N_{PNID}$
- B^j_{CTA} : Allocated bandwidth in a piconet with PNID j, $\quad j = 0, 1, \cdots, N_{PNID}$
- r : CTA ratio between the parent PNC and its child PNCs
- R_{C_PNC} : Basic CTA ratio between the parent PNC and its child PNCs

We can calculate the available CTA size of child piconets in level i, $T^i_{C_size}$ by Eq. (1). The $T^i_{C_size}$ consists of a beacon time and a CAP, and the basic CTAs for the DEVs. Since m_i child piconets request beacon time, CAPs, and CTA allocation remainder (R^{i+1}_{CTA1}) for child piconets for the next level $i+1$,and, n_i DEVs in level i are assigned T_{CTA},

$$T^i_{C_size} = (T_B + T_{CAP} + R^{i+1}_{CTA1}) \times m_i + (T_{CTA} \times n_i), \quad i = 1, 2, \cdots, L. \tag{1}$$

The $T^0_{S_size}$ consists of a beacon time and a CAP, and the basic CTAs for the parent piconet DEVs and is computed as Eq. (2). We can calculate the available superframe size for level i, $T^i_{S_size}$ by Eq. (3). The R^1_{CTA1} is the CTA allocation remainder to one child piconet for level 1.

$$T^0_{S_size} = T_B + T_{CAP} + (T_{CTA} \times N_p) \tag{2}$$

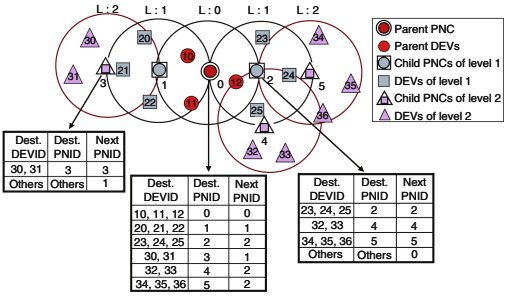


Fig. 2. Routing table example of a mesh network

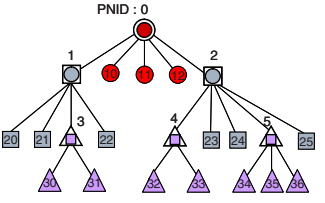


Fig. 3. The tree hierarchy of Fig. 2 example

$$T^{i}_{S_size} = T^{0}_{S_size} + R^{l}_{CTA1} + \left(\sum_{n=1}^{i} T^{n}_{C_size} \right), \quad i = 1, 2, \cdots, L. \tag{3}$$

The DEV discovery ($T^{0}_{Discovery}$) time in level 0 consists of a $T^{0}_{S_size}$ and a beacon time and is computed by Eq. (4). We can calculate the device discovery time for level i, $T^{i}_{Discovery}$ as Eq. (5).

$$T^{0}_{Discovery} = T^{0}_{S_size} + T_B \tag{4}$$

$$T^{i}_{Discovery} = (T^{i}_{S_size} \times (i+1)), \quad i = 1, 2, \cdots, L. \tag{5}$$

3.2 Proposed Channel Time Allocation Method and Routing Algorithm

The CTA allocation and routing algorithm have a great effect on efficient data transmission. In this section, we propose an efficient CTA allocation method and a routing algorithm. We assume that parent PNCs and child PNCs can support packet relay. Source DEVs will generate a new packet for data transmission. If the parent PNC or child PNCs receive the packet, they relay the packets. So CTAs of a superframe are allocated to source DEVs and the relaying PNCs. The parent PNC is located at the center of data transmission. Consequently, the traffic of parent PNC often experience heavy traffic. The proposed CTA allocation method allocates sufficient CTAs for the relay of the parent PNC. We can calculate the available CTA of a DEV in piconets with PNID *0*, $C^{0}_{D_CTA1}$ by Eq. (6). The $C^{0}_{P_CTA}$ denotes available CTAs of its parent PNC in a parent

```
PROCEDURE Routing Algorithm in Parent PNC and Child PNCs
1: while (a packet is received)
2:     if (there exists the destination PNID in the routing table)
3:         if (the destination PNID of the packet is the same as its own PNID)
4:             Send the packet to the destination DEVs
5:         else
6:             Send the packet to the next hop according to the routing table
7:     else
8:         Forward the packet to the upper PNC
```

Fig. 4. Pseudocode of the proposed routing algorithm

piconet with PNID *0* and is computed as Eq. (7). The $C^0_{C_CTA}$ is available CTAs for child PNCs with PNID 0,

$$C^0_{D_CTA1} = \frac{B^0_{CTA}}{Sn_0 + (Rn_0 + 1) \times R_{C_PNC} + (2 \times r)}, \qquad r = 0, 1, \cdots, R_{max} \tag{6}$$

$$C^0_{P_CTA} = C^0_{D_CTA1} \times (R_{C_PNC} + (2 \times r)), \qquad r = 0, 1, \cdots, R_{max} \tag{7}$$

$$C^0_{C_CTA} = C^0_{D_CTA1} \times R_{C_PNC} \tag{8}$$

where R_{max} is the maximum CTA ratio between a parent PNC and its child PNCs, and B^0_{CTA} is the allocated bandwidth in the parent piconet with PNID 0 and Sn_0 is the number of source DEVs for PNID *0*. And, Rn_0 is the number of child piconets for PNID *0*. The child PNCs are the member of their parent piconet and R_{C_PNC} is the basic CTA ratio the parent PNC and child PNCs. The R_{C_PNC} value is set to two. The 1 in Rn_0+1 denotes the number of CTAs for the parent piconet and the 2 in $2 \times r$ denotes the number of CTAs for uplink and downlink relay.

We can calculate the available CTA of a DEV in child piconets with PNID *j*, $C^j_{D_CTA1}$ by Eq. (9). The $C^j_{C_CTA}$ presents the available CTAs of child PNCs in child piconets with PNID *j* and is computed as Eq. (10).

$$C^j_{D_CTA1} = \frac{B^j_{CTA}}{Sn_j + (Rn_j + 1) \times R_{C_PNC}}, \qquad j = 1, 2, \cdots, N_{PNID} \tag{9}$$

$$C^j_{C_CTA} = C^j_{D_CTA1} \times R_{C_PNC}, \qquad j = 1, 2, \cdots, N_{PNID} \tag{10}$$

We then describe the proposed routing algorithm in the parent PNCs and the child PNCs. The proposed routing algorithm is a tree-based routing algorithm as in Fig. 3, which shows the tree hierarchy structure associated with Fig. 2. The routing tables of parent PNCs and child PNCs are created by an efficient device discovery process to find the shortest paths via parent PNCs or subordinate child PNCs as explained in Section 3.1. The pseudocode of the proposed routing algorithm is shown in Fig. 4. While parent PNCs or child PNCs receive a packet, if there exists the PNID of a destination DEV in the routing table, either parent PNC or child PNCs compare the destination PNID with its own PNID. If the destination PNID is the same as its own PNID, the packet is sent to the destination DEV. Otherwise, either parent PNCs or child PNCs send to the next

Table 1. Simulation parameters

Parameter	Value
Number of DEVs	100
Channel bandwidth	55 Mbps
Number of S-D pairs	$5 \sim 30$
Maximum buffer size	307,200 byte
Packet size	$64 \sim 2{,}032$ bytes
Beacon time (T_B)	0.6 ms
CAP time (T_{CAP})	1.00 ms
Maximum Superframe size (S)	65.535 ms
Guard time (T_{Gt})	0.0328 ms
SIFS	0.01 ms
MPEG-4 traffic rate	4 Mbps

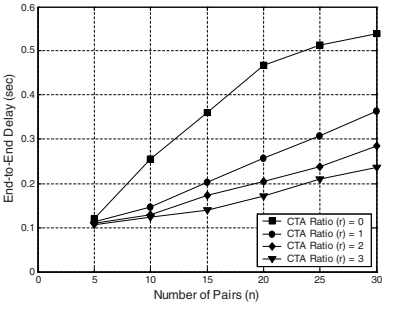


Fig. 5. End-to-end delay vs. the number of S-D pairs

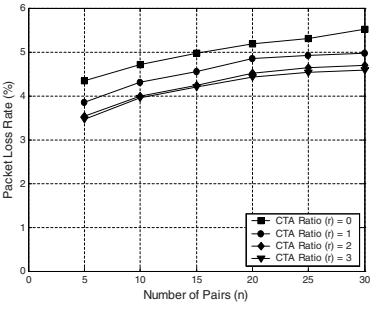


Fig. 6. Packet loss rate vs. the number of S-D pairs

hop according to the routing table. If there does not exist the destination PNID in the routing table, the packet is forwarded to the upper PNC. Because of the tree hierarchy, a packet is destined to the proper destination DEV by the proposed routing mechanism.

4 Performance Evaluation

We evaluate the performance of the proposed CTA allocation method and the routing algorithm for multi-hop communications via simulations. The simulation parameters are summarized in Table 1. The DEVs are assumed to be located uniformly in a square area of 50m $\times$ 50m. The mesh network formation algorithm and the bandwidth allocation for the parent piconet and child piconets are adopted as in [7]. We use MPEG-4 traffic model for isochronous traffic and its data rate of is 4 Mbps. The length of a MAC Service Data Unit (MSDU) is generated uniformly in $64 \sim 2{,}032$ bytes. The 802.15.3 protocol adds Short Interframe Space (SIFS) and guard time between individual CTAs to keep transmissions in adjacent CTAs from colliding.

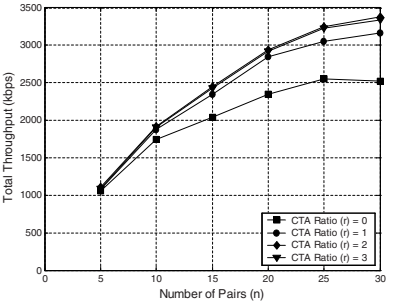


Fig. 7. Throughput vs. the number of S-D pairs

We analyze the performance of the CTA allocation method and the routing algorithm with respect to end-to-end delay, packet loss, and throughput. Fig. 5 presents the end-to-end delay by the proposed CTA allocation method and the routing algorithm. We compared end-to-end delays under various CTA ratio (r) and the number of source-destination (S-D) pairs. The end-to-end delay is about 0.54 second when r = 0 and the number of S-D pairs = 30. The end-to-end delay is improved up to approximately 0.24 second when r = 3 and number of S-D pairs = 30. As r increases, the end-to-end delay is shown to be decreased since PNCs are allocated more CTAs for relaying traffic. Fig. 6 presents the packet loss rate of the proposed algorithm. When the number of S-D pairs increases, the packet loss rate increases up to about $3 \sim 5\%$. We also analyze the total throughput. Fig. 7 shows the throughput performance of the proposed CTA allocation method and the routing algorithm. Total throughput increases as r and the number of S-D pairs increase. When r is 3 and the number of S-D pairs is 30, the total throughput becomes the maximum at about 3.4 Mbps.

5 Conclusion

In this paper, we have proposed the CTA allocation method and routing algorithm for multi-hop communications in IEEE 802.15.3 high-rate WPAN mesh networks. The proposed CTA allocation method allocates sufficient CTAs for relay PNCs. The proposed routing algorithm is simple and tree-based, and the routing tables of parent PNCs and child PNCs are created by an efficient device discovery process to find the shortest paths via parent PNCs or subordinate child PNCs. We evaluate the performance of the proposed algorithms via simulations. The end-to-end delay, the packet loss rate, total throughput indicates desired characteristics. The proposed algorithms are able to provide efficient multi-hop communications in IEEE 802.15.3 WPAN mesh networks.

In the future we intend to include the effects of various CTA allocation schemes for the relay PNCs and DEVs. Further, we are implementing and studying other features of the routing algorithm and bandwidth allocation.

References

1. IEEE, "Standards for Part 15.1: Wireless Medium Access Control (MAC) and Physical Layer (PHY) Specifications for Wireless Personal Area Networks (WPANs)," Jun. 2002.
2. J. Karaoguz, "High-Rate Wireless Personal Area Networks," IEEE Communications Magazine, vol. 39, no. 12, pp. 96-102, Dec. 2001.
3. IEEE, "Standard for Part 15.3: Wireless Medium Access Control (MAC) and Physical Layer (PHY) Specifications for High Rate Wireless Personal Area Networks (WPANs)," Sep. 2003.
4. Artimi Compony Ltd. UWB and Mesh networks white paper. Avail at http://www.artimi.com, Aug. 2003.
5. D. Trezentos, G. Froc, I. Moreau and X. Lagrange, "Algorithms for Ad-hoc Piconet Topology Initialization Evaluation for the IEEE 802.15.3 High Rate WPAN System," in Proc. of IEEE VTC, vol. 5, pp. 3448-3452, Oct. 2003.
6. X. Chen, J. Lu, Z. Zhou, "An enhanced high-rate WPAN MAC for mesh networks with dynamic bandwidth management," in Proc. of IEEE GLOBECOM, vol. 6, pp. 3408-3412, Nov.-Dec. 2005.
7. S.-B. Jung, S.-B. Yim, T.-J. Lee, S.-D. June, H.-S. Lee, T.-G. Kwon, and J.-W. Cho, "Multi-piconet Formation to Increase Channel Utilization in IEEE 802.15.3 High-rate WPAN," Springer-Verlag Lecture Notes in Computer Science, vol. 3392, pp. 1041-1049, May 2006.

Nonlinear Optimization of IEEE 802.11 Mesh Networks

Enrique Costa-Montenegro, Francisco J. González-Castaño, Pedro S. Rodríguez-Hernández, and Juan C. Burguillo-Rial

Departamento de Ingeniería Telemática, Universidad de Vigo, Spain
{kike,javier,pedro,jrial}@det.uvigo.es
http://www-gti.det.uvigo.es

Abstract. In this paper, we propose a novel optimization model to plan IEEE 802.11 broadband access networks. From a formal point of view, it is a mixed integer non-linear optimization model that considers both co-channel and inter-channel interference in the same compact formulation. It may serve as a planning tool by itself or to provide a performance bound to validate simpler planning models such as those in [3].

Keywords: IEEE 802.11, mesh networks, rooftop, planning.

1 Introduction

In this paper, we propose an optimization model to generate IEEE 802.11 resource-sharing broadband access meshes, which users themselves often manage [1]. Resource-sharing wireless networks based on IEEE 802.11 are not new [2]. In our model, a *basic node* is composed by a cable/xDSL router, an 802.11 access point and two 802.11 cards for interworking purposes. Basic nodes may serve a LAN (covering a building, for example). This model may represent user-managed rooftop networks linking building LANs, to share a pool of cable/xDSL accesses.

Our proposal relies on a set of *rules* to generate topologies with low co-channel and inter-channel interference. From them, in a previous paper [3] we derived two mesh deployment algorithms: a distributed one, to be executed by infrastructure nodes themselves, and a centralized one via a mixed integer linear optimization model. In this paper we enrich the centralized version, by adding co-channel and inter-channel interference estimates that yield a *mixed integer non-linear optimization model*. The new model may serve as a planning tool or to provide a performance bound to validate previous planning models.

Our study is based on IEEE 802.11b because it has been the most extended *legal* 802.11 substandard in the EU for a long time. It is straightforward to extend the results of this work to other substandards like IEEE 802.11a or 802.11g.

This paper is organized as follows: section 2 reviews the work in [3]. Section 3 describes the new proposal, a mixed integer non-linear optimization model satisfying the deployment rules in [3]. Section 4 presents numerical tests on a realistic scenario –a sector in Vigo, Spain–. Section 5 concludes.

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 466–473, 2007.
© Springer-Verlag Berlin Heidelberg 2007

2 Distributed IEEE 802.11 Deployment Algorithm

A single access point per basic node is a natural choice, since it can manage connections from several wireless cards. Multiple access points per basic node would compromise cell planning, due to the few channels available. According to our previous work, two wireless cards per basic node yield satisfactory performance and ensure network diversity and survivability.

2.1 IEEE 802.11b Channel Assignment

IEEE 802.11b has 13 DSSS (Direct Sequence Spread Spectrum) overlapping channels. We wish to minimize co-channel and inter-channel interference in wireless infrastructure deployment. There is co-channel interference when two access points (AP) employ the same channel, and inter-channel interference when APs or wireless cards (WLCs) with overlapping channels transmit simultaneously.

We adopt the classic cellular planning algorithm in [4] to generate a channel grid (Figure 1). We assign AP channels according to the cells they belong to. The maximum *legal* range (without boosting equipment) of card-to-access point connections is 170 meters (using a D-Link DWL-1000 AP and D-Link DWL-650 WLCs). IEEE 802.11b co-channel interference is negligible at distances over 50 m [5]. To mitigate co-channel and inter-channel interference, we allow a single *fully active* basic node per cell (the rest become *partially active*, by disabling the AP and one of the WLCs) and set cell edge length to 50 meters.

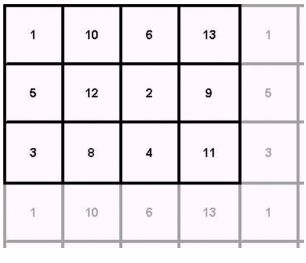


Fig. 1. Frequency pattern and cell grid

If there are several basic nodes in a cell, we need to decide which one is active. We achieved the lowest interference level when the active node is close to the cell center. Note that all basic nodes in a cell but the active one could be WLCs to reduce costs. However, this solution restricts network evolution (basic nodes can appear and disappear). Also, note that the cost *per user* is very low if the basic node serves a LAN.

2.2 Setting Wireless Links

As soon as a basic node is active, its WLCs look for the closest AP (in signal strength), i.e. the local one. The basic nodes filter the MAC addresses of their

own WLCs. However, once a WLC in basic node A is connected to the AP in basic node B, the latter must block the second card in A, since (i) overall capacity is the same due to AP sharing and (ii) there would be less diversity otherwise. It is possible to detect and avoid this *one-way dual connection establishment problem* because both WLCs in A belong to the same addressing range.

When a WLC wants to join the AP in a remote basic node, the latter must check if any of its WLCs have previously set a link in the opposite direction. The requesting WLC must notify its IP range, and the remote basic node can check if any of its own WLCs already belongs to that range (*two-way dual connection establishment problem*). If so, it will deny the connection.

In zones with many basic nodes, another connection establishment problem arises when some of them handle many connections. However, it is possible to limit the number of connections per AP at IEEE 802.11b MAC level by blocking association request frames from WLCs once the connection counter reaches the limit, keeping a reasonable throughput per connection. This has a second beneficial effect, because it favors network expansion at its edges.

We now consider co-channel and inter-channel interference. The former is presumably low due to cellular planning and cell size, specially if inter-cell links are short. Regarding inter-channel interference, if two physically adjacent IEEE 802.11b sources transmit with a two-channel separation, throughput drops to 50%, whereas a three-channel separation is practically enough to avoid inter-channel interference. It may be quite common in our case. Thus, we consider an *inter-channel interference mitigation rule*. If implemented, all elements in a basic node (WLCs and AP) only set connections with mutual frequency separation of at least three channels. This drastically reduces inter-channel interference.

2.3 Performance of the Distributed Deployment Algorithm

In [3], we simulated the distributed deployment algorithm on a realistic scenario with a significant number of basic nodes (≈ 50), corresponding to Vigo, Spain. We observed that most WLCs got connected and the resulting mesh had link diversity. Figure 2(a) shows the resulting network: black icons represent fully active basic nodes, and "$\times$" icons denote partially active ones (6 out of 46).

Access point degree is low: 2.11. At the solution, there are some unconnected APs and WLCs. This is not evident in figure 2(a), because the corresponding basic nodes are not isolated. This does not imply a waste of network resources, since those cases are mainly located at mesh edges, and thus they allow future growth. Also, over 50% APs have at most two connections, which implies a high throughput per connection. Practically all WLCs set connections (97.8%). The percentage of highest-rate links (11 Mbps) is close to 90%.

3 Improved IEEE 802.11 Deployment Algorithm

In [3], we proposed a centralized deployment algorithm based on a mixed integer linear program. Now we present a new mesh planning model that adds *explicit* counter-interference constraints, which is a mixed integer non-linear program.

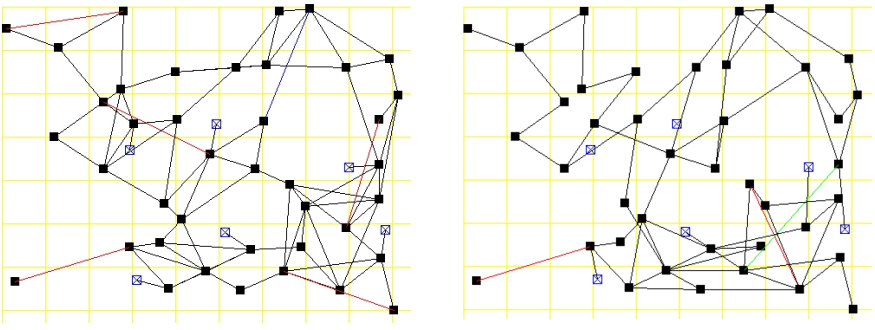


Fig. 2. Vigo: (a) Network example, (b) New algorithm with 25% frozen connections

Due to the complexity of this nonlinear model, our solver could not handle it on a Pentium IV desktop. So we decided to break it to solve it iteratively, as explained in section 4. In the next subsections we describe our model.

3.1 Sets and Constants

The main set BN contains N basic nodes b_i, $i = 0, \ldots, N-1$. This set is divided into two disjoint subsets, BN_f (fully active basic nodes) and BN_p (partially active basic nodes). Thus, $BN_p \cap BN_f = \emptyset$ and $BN = BN_p \cup BN_f$. Let d_{ij} indicate the distance between b_i and b_j, k_{ij} the capacity of the corresponding link, and ch_ap_i the channel of the AP in node b_i. If it is partially active, ch_ap_i=0. If it is fully active, ch_ap_i is the channel index of b_i plus two. Consequently, ch_ap_i is 0 or an integer in [3, 15].

3.2 Variables

Let $i, j = 0, \ldots, N-1$. The variables in the model are:

- $c1_{ij}$, $c2_{ij}$: Boolean variables. They equal 1 if WLCs #1 or #2 in b_i are connected to the AP in b_j, respectively. They equal 0 otherwise.
- ch_w1_i, ch_w2_i: real variables indicating the channels that WLCs #1 and #2 in b_i acquire once connected. The optimization model ensures that they will take integer values (see condition C6 and remark 1).
- δ_i: Boolean variable. If WLC #1 in b_i is not connected, it equals 1 to set ch_w1_i to dummy channel 18. Dummy channel 18 allows to set constraints (7)-(9) representing the inter-channel interference rule (remark 2).
- e_i, f_i: Boolean variables, to define linear constraints (8) and (9) that enforce the inter-channel interference mitigation rule in a fully active node b_i (remark 2).
- $conex_ap_i$: real variable, number of connections received by b_i. Values in [0,4].
- cap_i: real variable, aggregated capacity of the connections received by basic node i.
- cap_pon_i: real variable, average capacity of the connections bond to the AP in b_i.
- $degrk_i$: real variable, degradation in b_i due to links transmitting in channels with mutual distance k.
- $perc_dk_i$: real variable, indicating the percentage of capacity waste due to interference in basic node i by channels with mutual distance k.

3.3 Conditions

From mesh design specifications, we impose a series of conditions on model variables. The optimization tools take advantage of these conditions to reduce model size and execution time drastically. Let b_i, b_j be basic nodes in BN. Let b_p be a basic node in BN_p. Then:

C1 $c1_{ip}, c2_{ip} = 0$, since partially active nodes do not have APs.
C2 $c2_{pi} = 0$, since WLC #2 is disabled in partially active nodes.
C3 $k_{ij} = 0 \Rightarrow c1_{ij}, c2_{ij} = 0$: no connections between nodes that are far apart.
C4 $c1_{ii}, c2_{ii} = 0$, connections are forbidden within the same basic node.
C5 $\mid ch_ap_i - ch_ap_j \mid < 3 \Rightarrow c1_{ij}, c2_{ij} = 0$, due to the inter-channel interference mitigation rule. To understand this, suppose that $\mid ch_ap_i - ch_ap_j \mid < 3$ and $c1_{ij} = 1$ or $c2_{ij} = 1$. If so, one of the WLCs in b_i is connected to the AP in b_j. Consequently, at least one WLC transmitting in channel ch_ap_j is physically adjacent to b_i, whose AP transmits in ch_ap_i. Thus, there are overlapping transmissions.
C6 $ch_w2_p = conex_ap_p = cap_p = degrk_p = cap_pon_p = 0$ & $perc_dk_p = 1$: partially active nodes have no AP (so no degradation) and no WLC #2.

3.4 Constraints

1. $c1_{ij} + c2_{ij} + c1_{ji} + c2_{ji} \leq 1,\ i, j = 0, \ldots, N-1$. One-way and two-way dual connection avoidance rules.
2. $\sum_j c1_{ij} + \delta_i = 1,\ i = 0, \ldots, N-1$. WLC #1 in node b_i can set one connection at most. If the WLC is disconnected $\delta_i = 1$, and $\delta_i = 0$ otherwise.
3. $\sum_j c2_{ij} \leq 1,\ i = 0, \ldots, N-1$. WLC #2 in node b_i can set one connection at most.
4. $\sum_i (c1_{ij} + c2_{ij}) \leq 4,\ j = 0, \ldots, N-1$. Basic node b_j can accept four WLC connections at most.
5. $ch_w1_i = \sum_j (c1_{ij} \times ch_ap_j) + 18\delta_i,\ i = 0, \ldots, N-1$. WLC #1 in node b_i acquires the channel of the AP it joins, or dummy channel 18 if not connected.
6. $ch_w2_i = \sum_j (c2_{ij} \times ch_ap_j),\ b_i \in BN_f$. WLC #2 in node b_i acquires the channel of the AP it joins, or dummy channel 0 in partially active nodes.
7. $e_i + f_i = 1,\ b_i \in BN_f$. Variables e_i and f_i take complementary values. This helps us define constraints (8) and (9) below.
8. $(ch_w2_i - ch_w1_i) \geq 3e_i - 18f_i,\ b_i \in BN_f$. This constraint enforces the inter-channel interference mitigation rule when (*i*) WLCs #1 and #2 in $b_i \in BN_f$ are connected and (*ii*) $ch_w2_i > ch_w1_i$.
9. $(ch_w2_i - ch_w1_i) \leq -3f_i + 12e_i,\ b_i \in BN_f$. Same case, when $ch_w2_i < ch_w1_i$.
10. $conex_ap_i = \sum_j c1_{ji} + \sum_j c2_{ji}, b_i \in BN_f$. Connections received by a fully active node.
11. $cap_i = \sum_j (k_{ji} \times (c1_{ji} + c2_{ji})), b_i \in BN_f$. Aggregated capacity of fully active nodes.

12.1 $degr0_i_1 = \sum_j [\frac{1 - c1_{ji}}{1 + 1000(ch_ap_i - ch_w1_j)^{10}} + \frac{1 - c2_{ji}}{1 + 1000(ch_ap_i - ch_w2_j)^{10}}$
$+ (1\ if\ (ch_ap_i - ch_ap_j) = 0,\ 0 \text{ otherwise})]$, where $b_j \in BN, i \neq j$ and $d_{ij} \leq 50$.

12.2 $degr0_i_2 = \sum_j [\frac{1 - c1_{ij}}{1 + 1000(ch_ap_j - ch_w1_i)^{10}} + \sum_k \frac{(c1_{ik} - c1_{jk})^2}{2(1 + 1000(ch_w1_j - ch_w1_i)^{10})} +$
$\sum_k \frac{(c1_{ik} - c2_{jk})^2}{2(1 + 1000(ch_w2_j - ch_w1_i)^{10})}]$, where $b_j \in BN, i \neq j$ and $d_{ij} \leq 50$.

12.3 $degr0_i_3 = \sum_j [\frac{1-c2_{ij}}{1+1000(ch_ap_j - ch_w2_i)^{10}} + \sum_k \frac{(c2_{ik}-c1_{jk})^2}{2(1+1000(ch_w1_j - ch_w2_i)^{10})} + \sum_k \frac{(c2_{ik}-c2_{jk})^2}{2(1+1000(ch_w2_j - ch_w2_i)^{10})}]$, where $b_j \in BN, i \neq j$ and $d_{ij} \leq 50$.

12. $degr0_i = degr0_i_1 + degr0_i_2 + degr0_i_3$. Variables $degr1_i$ and $degr2_i$ are similarly defined for interference distances 1 and 2, respectively, with a slightly higher complexity. Let $perc_d0_i = \frac{1}{1+degr0i}$, $perc_d1_i = \frac{1}{1+0.75\times degr1_i}$, $perc_d2_i = \frac{1}{1+0.5\times degr2_i}$, $b_i \in BN_f$. Variable $perc_d0_i$ represents wasted capacity due to co-channel interference. For no co-channel interference ($degr0_i = 0$), $perc_d0_i = 1$, i.e. there is no loss. For a single interfering element, $perc_d0_i = 0.5$, and so on. Weights 0.75 and 0.5 in $perc_d1_i$ and $perc_d2_i$ represent a lower capacity loss as a result of distances 1 and 2.

13. $cap_pon_i = \frac{cap_i}{conex_ap_i}$, $b_i \in BN_f$. The average capacity of the connections bond to the AP in basic node i.

Remark 1. Although ch_w1_i and ch_w2_i are declared as continuous real variables, their feasible values are integer due to constraints (5) and (6).

Remark 2: Constraints (7)-(9) are extremely important because they are equivalent to the reverse convex constraint $\mid ch_w2_i - ch_w1_i \mid \geq 3$, which induces a disjoint feasible region. Note that $e_i = 1$ implies $ch_w2_i - ch_w1_i \geq 3$ and inequality (9) holds trivially. On the other hand, $e_i = 0$ implies $ch_w1_i - ch_w2_i > 3$ and inequality (8) holds trivially. Note the importance of dummy channel 18 for WLC #1: if we represented the disconnected state of both WLCs by dummy channel 0, constraints (8) and (9) could not be jointly feasible. The interested reader can obtain more information on modeling disjoint regions in [6] (chapters 9 and 10).

Remark 3: Due the complexity of variable $degr0_i, b_i \in BN_f$, we decided to split it in three parts (12.1 to 12.3).

Part 12.1 considers the elements causing co-channel interference at the AP in basic node i (less than 50 m away). The first term considers interfering WLCs #1. If WLC #1 in j joins the AP in i, factor $(1-c1_{ji})$ will be zero (the constraints avoid interference). Note that $1 + 1000(ch_ap_i - ch_w1_j)^{10}$ will be 1 if $ch_ap_i = ch_w1_j$ (co-channel interference) and it grows exponentially with channel distance. As a denominator, this expression penalizes the first term, which is only significant in case of co-channel interference. Alternative (clearer) formulations were possible, using the absolute value, sign or scalar functions, but the solver considers them non-smooth or discontinuous functions. The second term counts interfering WLCs #2. Finally, the third term simply counts interfering APs.

Part 12.2 considers the co-channel interference events affecting WLC #1 in basic node i (less than 50 m away). The first term represents interfering APs, and it is similar to the second term in (12.1). The second term considers interfering WLCs #1. Note that, if both WLCs #1 in i and j join the same access point, factor $\sum_k (c1_{ik} - c1_{jk})^2$ will be zero (the constraints avoid interference). However, if the WLCs join different APs, the sum of their contributions multiplied by the common factor $\frac{1}{2(1+1000(ch_w1_j - ch_w1_i)^{10})}$ will be two (which explains the 2 in

the denominator of the second term). Finally, the third term counts interfering WLCs #2. If a single WLC is connected, the denominators in the second and third terms are so large that they do not contribute to interference.

Part 12.3 considers the co-channel interference events affecting WLC #2 in basic node i (less than 50 m away). The first term counts interfering APs, like the third term in (12.1). The second term counts interfering WLCs #1, like the third term in (12.2). Finally, the third term counts interfering WLCs #2.

3.5 Objective Function

The model seeks to maximize infrastructure capacity as follows:

14. Maximize $\sum_i [cap_pon_i \times perc_d0_i \times perc_d1_i \times perc_d2_i], \quad b_i \in BN_f.$

4 Numerical Tests

We tested the new model in the Vigo scenario in [3] (Figure 2(a)). The complexity of the full MINLP (Mixed Integer Non-Linear Programming model) problem is enormous. We tried to solve its GAMS 21.4 model. The solver did not succeed on a Pentium IV at 2.4 GHz with 512 MB RAM. Even after GAMS compilation, the size of the full MINLP is 2804 rows, 4840 columns, and 32874 non-zeroes. Thus, we developed an iterative approach that considered the three interference distances (0,1,2). First we solve the linear model in [3] to get an initial value for the second step. In it, the model only considers co-channel interference ($perc_d0_i$ contributions). Then, we freeze a subset of connections without inter-channel interference, to define a new starting point for the third step, which considers interference between adjacent channels (distance one). The fourth step is defined accordingly, by considering distance-two interference. From the resulting point we start again by only taking co-channel interference into account. The algorithm should stop when most connections are fixed, yielding as a final result the intermediate solution with maximum objective function value (comprising co-channel and inter-channel interferences at distances one and two). However, we obtained results of practical interest with a single run of the first two steps. The size of the resulting compiled MINLP is 2524 rows, 4554 columns, and 25326 non-zeroes. Apparently the size is the same, but we mainly eliminate non-linear constraints.

Table 1 shows objective function (14) values at algorithm termination. We observe an improvement over [3] in all cases studied. The results are very similar when we freeze connections. This is possibly because we consider co-channel interference in first place and, since it is the most troublesome, the best connections are frozen early at the beginning. However, as we could expect beforehand, elapsed time drops drastically with the number of frozen connections. Table 1 also shows interference events associated to the objective function values ($x-y-z$: x distance 0 interference events, y distance 1 ones, z distance 2 ones). We observe an improvement in all instances of the new mathematical model. In some cases, we completely eliminate co-channel interference. In Figure 2(b) we plot the resulting network for the instance with 25% frozen connections. It is still fully connected. The average node degree is 2.74.

Table 1. Objective function (14) & improvement in interference

Test/ frozen connections	Distributed algorithm	Mathematical model in [3]	New math. model	New math. model elapsed time
Test 1 (0%)	183.3792 6 - 34 - 56	248.8336 6 - 38 - 34	259.6676 0 - 34 - 44	3600 (time limit)
Test 2 (10%)	198.2935 8 - 28 - 50	241.9191 8 - 38 - 40	261.6800 6 - 20 - 42	3600 (time limit)
Test 3 (25%)	183.7588 6 - 32 - 42	250.528 10 - 34 - 32	274.1521 2 - 24 - 36	239
Test 4 (50%)	160.3011 6 - 38 - 44	252.2052 6 - 38 - 34	265.3527 0 - 34 - 46	183.74

5 Conclusions

We have presented a new wireless mesh planning algorithm (a mixed integer nonlinear programming optimization model comprising interference constraints), which we compare with the simpler deployment algorithms in [3]. Although the new approach clearly produces better results in terms of interference minimization, it also allows us to validate the faster methods in [3]. Our algorithms do not *completely* eliminate interference (there is a trade-off between interference and connectivity). However, according to our results, both co-channel and inter-channel interference are extremely low at the solution.

References

1. Hubaux J.P., Gross T., Boudec J.Y.L., Vetterli M.: Towards self-organized mobile ad-hoc networks: the terminodes project, IEEE Commun. Mag. 1, pp. 118-124, 2001.
2. MIT Roofnet, http://www.pdos.lcs.mit.edu/roofnet, 2004.
3. Costa-Montenegro E., González-Castaño F.J., García-Palomares U., Vilas-Paz M., Rodríguez-Hernández P.S.: Distributed and Centralized Algorithms for Large-Scale IEEE 802.11b Infrastructure Planning, Proc. IEEE ISCC, Alexandria, 2004.
4. Box F.: A heuristic technique for assigning frequencies to mobile radio nets, IEEE Trans. Veh. Technol., vol. VT-27, pp. 57-74, 1978.
5. Chen J.C.: Measured Performance of 5-GHz 802.11a Wireless LAN Systems, Atheros Communications white paper, 2001.
6. Williams H.P.: Model building in mathematical programming, Wiley & sons, NY, 1999.

Securely Deliver Data by Multi-path Routing Scheme in Wireless Mesh Networks*

Cao Trong Hieu and Choong Seon Hong**

Department of Computer Engineering, Kyung Hee University
Giheung, Yongin, Gyeonggi, 449-701 Korea
hieuct@networking.khu.ac.kr, cshong@khu.ac.kr

Abstract. Wireless Mesh Networks with static Transit Access Points (TAPs) have many advantages to connect different kinds of networks. While Mobile Ad hoc Networks still have many challenges to improve because of dynamic topology and security vulnerabilities, WMNs are the best solution for wireless communication recently. To utilize the characteristics of WMN's topology, in this paper, we propose an algorithm to preserve privacy for routing. This idea comes from the fact that if we can separate data traffic into more than one path, the probability to capture all traffic from intermediate node is very small. It means it is very difficult to launch traffic analysis attacks because of traffic confidentiality. In addition, to hide securely the real source and destination addresses a new technique is proposed along with an Adaptive Key Agreement Scheme. We apply Information Entropy to model our routing traffic and highlight the robustness of the algorithm. We also present a detail traffic evaluation observed from neighboring nodes to show the availability of our proposal in term of robustness, loop free and computational overhead.

Keywords: Security, Routing, Privacy Preservation, Information Entropy, Wireless Mesh Network.

1 Introduction

Along with Mobile Ad-hoc Network, Wireless Mesh Network recently has attracted increasing attention thank for the low-cost deployment and topology flexibility [2]. WMN represent a good solution to providing wireless Internet connectivity in a large scale. This new and promising paradigm allows for deploying network at much lower cost than with classic WiFi network. However, the routing mechanism must be secure.

We consider a Mesh Topology shown in Fig. 1. In this network, multiple mesh routers communicate with each other to form a multi-hop wireless backbone that forwards user traffic to the gateways which are connected to the Internet. Client devices access a stationary wireless mesh router at its residence.

Confidentiality (privacy) is one of the most important criteria regarding security aspect. In this paper, we focus on traffic confidentiality which prevents the traffic analysis attack from the mesh router.

* This work was supported by the MIC under ITRC Project (IITA-2006-C1090-0602-0002).
** Corresponding author.

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 474–481, 2007.
© Springer-Verlag Berlin Heidelberg 2007

The key idea is if the traffic between source S and destination D goes through only one route, any intermediate node can easily observe the entire traffic between S and D.

This route is vulnerable to traffic privacy attacks. To tackle this weakness, we propose a Multi-path routing mechanism which utilizes multiple paths for data delivery and can protect attacks based on data analysis.

The rest of the paper is organized as follows: section 2 briefly discusses some related works. In section 3, we propose an algorithm to find the multi-path between two mesh routers (nodes) when end-users want to communicate with each other or access to Internet. In addition, we propose an *Adaptive Key Agreement Scheme* to encrypt the data packets and transmit through multiple disjoint paths found in the previous step. In our scheme, we introduce a new technique that can hide real source and destination addresses. To make our proposal more reliable, we apply Information Entropy to model our routing traffic and prove the robustness of the algorithm in section 4. Finally, section 5 exposes some perspectives for further work.

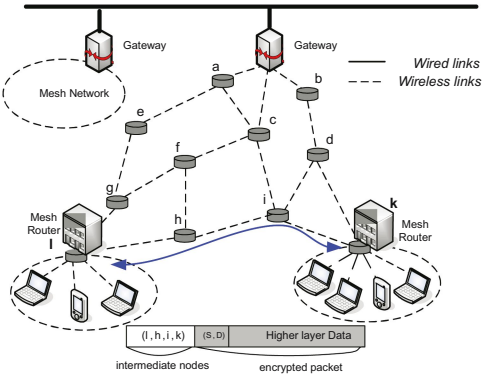


Fig. 1. General Mesh Topology

2 Related Work

WMN is a hybrid network which has both mobile parts and stationary parts. However, due to limited capacity, delay constrains [3] and the lack of security guarantees [4, 10, 11], WMNs are not yet ready for wide-scale deployment. The first problem can be solved by using multi-radio and multi-channel Transit Access Points (TAPs) [5]. The other most important challenge concerned here is security especially in routing protocol.

In the existing literature, the onion routing [12] developed by David Goldschlag et al. can secure communication through an unpredictable path but it is necessary to encrypt message between routers. This means all intermediate nodes have to involve in encryption/decryption process which cause more overhead. In wireless ad-hoc networks, authors proposed schemes for location and identity privacy in [8, 9]. However, none of them can be applied to WMN directly. Traffic forwarding relationship among nodes is strongly dependent on their locations and the network topology, which is static in WMN. In addition, WMNs have some specifications that require adaptive changes in routing protocol. Our proposed routing protocol will take into account the inherent structure and solve those existing constrains.

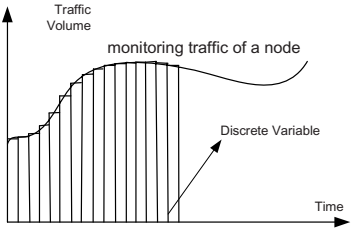


Fig. 2. Sampling continuous traffic

In reality, the traffic of a node is a continuous function of time, as shown in Fig. 2.

However, in our proposal, to apply Information Entropy for privacy preservation, we consider the traffic as discrete random variable. Therefore, as the first step, we discrete the continuous traffic into piece-wise approximation of discrete values. Then we measure the amount of traffic in each period, usually in terms of number of packets, with assumption that the packet sizes are all equal.

3 Proposed Algorithms

In this part, we propose an algorithm to preserve privacy for routing along with *Adaptive Key Management Scheme* to transfer data securely between two nodes.

3.1 Multi-path Finding Algorithm

To apply our algorithm to routing protocol, it is needed a little bit change in routing table. We define *Found Route* to count and keep the number of paths found after the algorithm is executed. *Node Occupied Status* is 0 at initial stage and is set to 1 if a node is not available or it is already in a found route. *Number_RREQ* is the number of requests sent from source to destination. Each time a route is found or *Request_Time* is over, the source will send another request and *Number_RREQ* will be counted down. In our algorithm, *Number_RREQ* is equal to the number of neighbors of source node's AP. *Request_Time* can be assigned flexibly so that it is not too long to avoid overhead and not too short to guarantee path-finding process.

Hop count (HC) is used to determine the shortest path and it is increased by 1 if *RREQ* or *RREP* is forwarded each hop. In this algorithm, *HC* is also used to avoid *RREQ*'s loop back which also causes time and energy consumption.

In *Step 1*, all nodes' states are unoccupied. The *RREQ* is sent to all neighbors of source node. Node's availability [1] will be checked in this step. As mentioned above, *Hop Count (HC)* is stored in routing table of each node and compared with new *HC* index when a *RREQ* arrives. If new *RREQ* has *HC* smaller than current one, the node will update new *HC* and go to *Step 2*.

In *Step 2*, Node's Address is compared with Destination Address in *RREQ*. If it has the same address, *Found Route* is increased, *Node Occupied Status* is set to 1 and the number of *RREQ* is decreased by 1. At this time, *Number_RREQ* and *Request_Time* are checked in Step 3 and if one of them equals 0, the algorithm is finished. Those conditions guarantee overhead avoidance.

Note that when a node does not satisfy the condition in Step 2, it will uni-cast back to notify the source and from this time it will not participate in the routing process. Moreover, the repetition of step 1 in step 2 is different from step 3 because the *Number_RREQ* is not counted down. *Number_RREQ* is only counted down when a new route is found. That is the reason why we need *Request_Time* to avoid overhead.

After the finding algorithm finished, in the routing table of involved nodes, the information about the number of routes and list of nodes in each route are stored. From that information, source node starts to send data through separate paths. As we discussed in [1], the path between source and destination in this case also need not be shortest path regarding hop count.

```
   Initial
Node's Occupied Status =
0;Found Route = 0;
   Number_RREQ = n  /* n
is the number of
neighbors of source
node*/
   Request_time = k;

   Step 1: flood RREQS to
unoccupied neighbor
nodes;
   check node's availabil-
ity & arrived_HCi;

   Step 2:
   if arrived_HCi < Cur-
rent_HCi
     { if Node_ Add ==
Destination_Add
       {Found Route ++;
Set Occupied_status = 1;
Number_RREQ --};
   else return Step 1};
       else { discast
RREQi;Set Occupied_status
= 1;
   finish};

   Step 3: repeat step 1;
           finish while {
Number_RREQ = =0 or Re-
quest_time == 0}
```

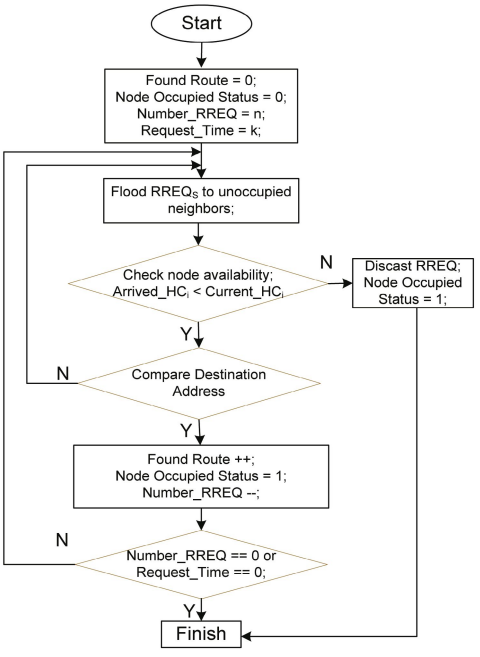


Fig. 3. Multi-Path Finding Algorithm

3.2 Adaptive Key Management Scheme

As briefly mentioned in section 1, in this part, we introduce a new technique that can hide real source and destination addresses.

After process *Multi-Path Finding Algorithm*, the source and current source AP run 2-party Diffie-Hellman in parallel with the destination and current destination AP do[13].

The key exchange includes 2 steps. At the first step (represented by solid arrows in Fig. 4), the source node and its access point (AP) choose a secret number *(x, y)* respectively, a large co-prime *(g, p)*, and exchange to make a common key $K_{S/S'AP}$:

$$K_{S/S'AP} = g^{xy} \bmod p$$

At this time, the destination node and its AP also choose a secret number *(u, v)* respectively and exchange to make a common key $K_{D/D'AP}$:

$$K_{D/D'AP} = g^{uv} \bmod p$$

In the second step (represented by dashed arrows in Fig. 4), the source AP and the destination AP run 2-party D-H in parallel with source and destination do and compute a common shared key $K_{S/D}$:

$$K_{S/D} = g^{xyuv} \bmod p$$

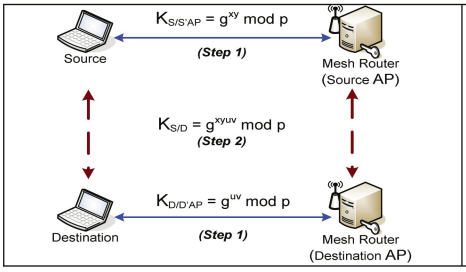


Fig. 4. Key Exchange Scenario

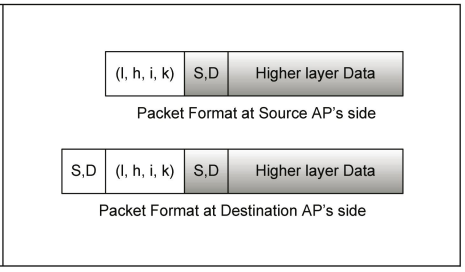


Fig. 5. Additional Field in Packet's Format

After this process, four nodes will have the same key $K_{S/D}$ and they can communicate securely.

To hide the real S/D addresses, we proposed a new technique that intermediate nodes can not extract to know address of S/D. This technique can prevent almost kinds of attacks based on data privacy.

At the source side, before transmitting, the data is split and encrypted with S/D addresses. After that, the addresses of intermediate nodes found in previous step (section 3) are attached without encryption. By this way, the intermediate nodes can only extract the source AP's address and destination AP's address. Each time an intermediate node receives a packet, it simply forwards this packet to the next hop in the address sequence. Without the need of knowing S/D address, all the packets will arrive to destination AP.

One challenge for proposed scheme is how to avoid computation overhead at receiver side because normally the destination AP will broadcast packets to all wireless clients in its range in MAC protocol. To solve this problem, the destination AP will use the common key $K_{S/D}$ to extract the S/D address in each packets and puts it in the unencrypted part before sending to its neighbor as showed in the figure 5. When all clients receive the packets, they simply compare the destination address. If a packet is for a node, it can decrypt the packet thanks to $K_{S/D}$. If the packet is not for this node, it will drop and also can not try to decrypt the packet. In briefly, this technique can make the second protection layer for privacy of data not only at intermediate compromised nodes but also at receiver side. It also puts a little more computation overhead only at S/D access points.

To illustrate the privacy preserving and evaluate the rare probability that an attacker can capture and reassemble the data from source to destination in our algorithm, in the next section, we apply *Information Entropy* (also called Shannon Entropy) into our proposal.

4 Traffic Evaluations

In the information theory, the concept of *Information Entropy* (Shannon Entropy) describes how much information there is in a signal or event. In our proposal, it is used for evaluating the traffic volume that goes through separate routing paths described above.

We discrete continuous traffic into equal-size sampling period as discussed in the section 2, and use A as the random variable of this discrete value. The probability that

the random variable A is equal to i (a node receives i packets in a sampling period) is $P(A = i)$. Likewise, $P(B^A = j)$ is the probability that B^A is equal to j. $(i, j \in N)$.

The Information Entropy of the discrete random variable A is

$$H(A) = \sum_{i=1}^{n} P(A=i) \log_2 (\frac{1}{P(A=i)}) = -\sum_{i=1}^{n} P(A=i) \log_2 P(A=i) \quad \text{(I)}$$

$H(A)$ is a measurement of the uncertainty about the outcome of A. It means if the value of A is distributed and no value predominates, $H(A)$ takes its maximum value. On the other hand, if the traffic pattern is *Constant Bit Rate (CBR)*, then $H(A) = 0$.

Similarly, we have the entropy for B^A as follows.

$$H(B^A) = -\sum_{i=1}^{n} P(B^A = i) \log_2 P(B^A = i) \quad \text{(II)}$$

B^A is a random variable representing the number of packets destined to node a observed at node b in a sampling period.

Then we define the conditional entropy of random variable B^A with respect to A as

$$H(A / B^A) = -\sum_{j=1}^{m} P(B^A = j) \sum_{i=1}^{n} p_{ij} \log_2 p_{ij} \quad \text{(III)}$$

in which, $p_{ij} = P(A = i | B^A = j)$ is the probability that $A = i$ given that $B^A = j$. $H(A|B^A)$ can be thought of as the uncertainty remained about A after B^A is known. The joint entropy of A and B^A can be shown as

$$H(A, B^A) = H(B^A) + H(A / B^A) \quad \text{(IV)}$$

The mutual information of A and B^A which represents the information we can gain about A from B^A is defined as

$$I(B^A, A) = H(A) + H(B^A) - H(A, B^A) = H(A) - H(A / B^A) \quad \text{(V)}$$

Suppose the traffic observed at b is proportional to a at any sampling period. If $B^A = j$, we can conclude that A equals to a fixed value i. In this case, we have $P(A = i | B^A = j) = 1$. This, according to *Eq. (III)*, makes the conditional entropy $H(A|B^A) = 0$. It means the uncertainty about the outcome of A when we know B^A is 0. From *Eq. (V)*, we have $I(B^A,A) = H(A)$, implying that we gain the complete information about A, given B^A. Otherwise, if B^A is independent of A, the conditional entropy $H(A|B^A)$ is maximized to $H(A)$. According to *Eq. (V)*, we have $I(B^A,A) = 0$, *i.e.*, we gain no information A from B^A. From *Eq. (V)*, we also figure out that we have to minimize the maximum mutual information $I(B^A,A)$ that any node can obtain about A to preserve privacy. In fact, since B^A records the number of packets destined to node a, it can not be totally independent of random variable A. Therefore, the mutual information should be valued between the two extremes discussed above, i.e., $0 < I(B^A,A) < H(A)$. This means that node b can still obtain partial information of A's traffic pattern.

Finally, we denote the average traffic through a node in a disjoint path as

$$T_{Avr} = \frac{1}{m} \sum_{i=1}^{m} T_i \quad \text{(VI)}$$

in which, m is the number of path found, T_i is the traffic of a node at a specific time.

We set up a simulation environment using *NS-2* and analyze the traffic of an intermediate node by the data observed from its neighbor. We analyze traffic in three cases regarding the number of found route (m = 1, 2, 3). The traffic is randomly distributed through found route in previous step, and at the same time, the total traffic imultaneously runs through those paths is 100 percent. It means the more number of found routes, the less data is transferred through a node, and the probability to capture the whole traffic is very small.

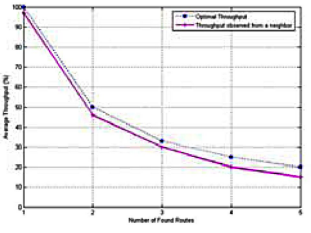


Fig. 6. Average Throughput Corresponding With Format Route

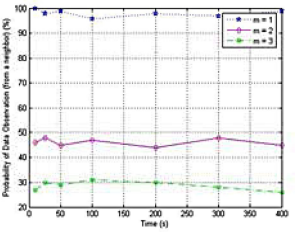


Fig. 7. Traffic observation Corresponding to Number of Found Routes

As shown in fig. 6, the obtained Average Throughput of a node in a route is always larger than the throughput observed by a neighbor of it.

In the figure 7, we monitor Traffic Throughput of a node by its neighbor in a period of time. The figure has shown that the probability of successfully capturing data will be reduced in direct proportion to the number of found routes (m). It means traffic privacy will be preserved in direct proportion to m.

5 Discussion and Conclusion

Our proposed approach in this paper is applied to WMNs which have static Mesh Router. In case of Wireless Mobile Ad-hoc Networks, it is much more difficult to maintain found routes according to the node's mobility. In fact, the routers which placed in a building are supposed to be physically protected. Therefore, they are harder to attack than the Transit Access Points (TAPs) which are placed outside. Along with current key managements and authorization schemes, the APs are almost fully protected. If some attacks occur at intermediate nodes, as shown in previous sections, the probability that attackers can capture and restore data that is sent from source to destination through several disjoint paths is very small. Note that even if attackers can capture 99%, they still can not merge the data and this stolen data is meaningless.

After a route was found, the data is split and marked before it is sent to the destination. When other routes are found, the remaining packets will be continuously sent through those paths randomly. This mechanism will reduce time consumption and also preserve data confidentiality.

In our algorithm, we especially concern about reducing overhead, so that we propose two parameters as *Request_Time* and *Number_RREQ* (discussed in section 3) to avoid time consumption. Also, the algorithm is loop free thanks to the discarded *RREQ* and the finish of participating progress of unavailable nodes in *Step 2*.

The algorithm needs a small change in routing table and can be easily applied to the current routing platforms as discussed in section 2. Also, in our environment, there are enough number of nodes to find multiple disjoint path. Of course, in the worst case, there is only one communication path (for example with only 3 mesh routers) and this scenario becomes conventional communication (one route between source and destination).

In the future work, we will discuss attack scenarios and countermeasures regarding to security analysis and continue implementing our proposal in Testbed cooperating with existing routing protocol for WMNs. In addition, we will provide specific analysis how our scheme is implemented with well-known encryption algorithms to make the communication route more secure. Also, we are working on an algorithm for privacy preservation in Mobile Wireless PAN in which the network topology always changes due to node's mobility.

References

1. Cao Trong Hieu, Tran Thanh Dai, Choong Seon Hong.: Adaptive Algorithms to Enhance Routing and Security for Wireless PAN Mesh Networks, OTM Workshops 2006, LNCS 4277, pp. 585 – 594, 2006.
2. R. Karrer, A. Sabharwal, and E. Knightly.: Enabling large-scale wireless broadband: The case for taps, In HotNets, 2003.
3. V. Gambiroza, B. Sadeghi, and E. Knightly.: End-to-End Performance and Fairness in Multihop Wireless Backhaul Networks, Proc. MobiCom, 2004.
4. Ben Salem, N.; Hubaux, J.-P.: Securing wireless mesh networks, Wireless Communications, IEEE, April 2006 Page(s):50 - 55
5. M. Kodialam and T. Nandagopal.: Characterizing the Capacity Region in Multi-Radio Multi- Channel Wireless Mesh Networks, Proc. MobiCom, 2005.
6. M. G. Reed, P. F. Syverson, and D. Goldschlag.: Anonymous connections and onion routing, IEEE Journal on Selected Areas in Communications, 16(4):482–494, 1998.
7. Shu Jiang; Vaidya, N.H.; Wei Zhao.: Preventing traffic analysis in packet radio networks, DARPA Information Survivability Conference & Exposition II, 2001. DISCEX '01, Proceedings Volume 2,12-14 June 2001 Page(s):153 – 158.
8. X. Wu and B. Bhargava.: Ao2p: Ad hoc on-demand position-based private routing protocol, IEEE Transactions on Mobile Computing, 4(4):335–348, 2005.
9. S. Capkun, J. Hubaux, and M. Jakobsson. : Secure and privacy preserving communication in hybrid ad hoc networks", Technical Report IC/2004/104, EPFL-DI-ICA, 2004.
10. Y.-C. Hu, A. Perrig, and D. B. Johnson.: Ariadne: A Secure On-Demand Routing Protocol for Ad Hoc Networks, In Proceedings of MobiCom, September 2002.
11. P. Papadimitratos and Z.J. Haas.: Secure Routing for Mobile Ad Hoc Networks, In Proceedings of CNDS, January 2002.
12. David Goldschlag, Michael Reed, Paul Syverson.: Onion Routing for Anonymous and Private Internet Connections, Communications of the ACM, Volume 42 , Pages: 39 – 41, February 1999
13. Asokan, N., and Ginzboorg, P.: Key agreement in ad-hoc networks, in Computer Communications, vol. 23, p. 1627 – 1637, 2000
14. Taojun Wu, Yuan Xue and Yi Cui.: Preserving Traffic Privacy in Wireless Mesh Networks, the International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM'06), 2006, pp. 459 – 461.

Cross-Layer Enhancement of IEEE 802.11 MAC for Mobile Ad Hoc Networks*

Taekon Kim[1], Hyungkeun Lee[2], Jang-Yeon Lee[3], and Jin-Woong Cho[3]

[1] Department of Electronics & Information Engineering, Korea University
[2] Department of Computer Engineering, Kwangwoon University
[3] Korea Electronics Technology Institute, Korea
taekonkim@korea.ac.kr, hklee@kw.ac.kr
{jylee136,chojw}@keti.re.kr

Abstract. In mobile ad hoc networks, the large amount of control overheads associated with discovering and maintaining end-to-end routing path information may not be tolerable. In this paper, we present the design and simulation of a new approach for IEEE 802.11 MAC based on the multipath routing information for ad hoc networks. The routing information about multiple paths discovered in the network layer is exploited by the MAC layer in order to forward a frame over the best hop out of multiple hop choices. The performance of our approach is compared with the one of the IEEE 802.11 MAC protocol via simulation. The results show that our proposed scheme exhibits a remarkable performance improvement over the IEEE 802.11 MAC protocol in terms of packet overhead and end-to-end throughput.

Keywords: IEEE 802.11, MANET, Cross-layer enhancement.

1 Introduction

A mobile ad hoc network (MANET) is a collection of mobile nodes dynamically organizing themselves for communication without requiring existing infrastructure. Each node in such a network operates not only as an endpoint but also as a router that has the functionality to forward data over the next hop while maintaining the route information. In MANETs, multiple network hops may be needed for communication between two distant nodes, due to the limited range of radio transmission. Therefore, the delivery of data between two nodes is much more complex and challenging in MANETs. For successful communications in such a network, a routing protocol should deal with the typical characteristics of these networks, such as limited bandwidth, high error rate, limited power capacity and node mobility. In this paper, we address a cross-layer technique between the MAC layer and routing layer and develop a new approach of IEEE 802.11 MAC that exploits such a cross-layer interaction.

* This work is supported by the ubiquitous Autonomic Computing and Network Project, the Ministry of Information and Communication (MIC) 21st Century Frontier R&D Program in Korea, and the Research Grant of Kwangwoon University in 2006.

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 482–489, 2007.
© Springer-Verlag Berlin Heidelberg 2007

In MANETs, the state of a link between two nodes is governed by the channel impairments such as the interference and fading at the receiver as well as the noise. The channel impairments could be time-varying, and significant changes in fading and interference levels may lead to a transient link failure. This link failure is often sufficient for routing and transport protocols to react, which causes operational inefficiencies. Therefore, there is a need to devise a data-forwarding mechanism that can tolerate this type of link failure at short time-scales. Furthermore, intermediate nodes shared by others may cause data transmission to defer or even fail, called link blocking. The effect of link blocking as well as link failure can be alleviated by forwarding frames via an alternative path reaching the destination. An example is shown in Figure 1. The routing protocol decides a transmission path between the nodes A and E, while the node B is being accessed by another node F and the link, A-D, is temporally broken due to the high level of fading. The transmission via the B or D node leads to transmission retries, deferring transmission, increasing delay, and wasting the bandwidth as a result. An improved approach would be choosing the next hop, A-C, on an alternate possible path, A-C-E, by the cross-layer coordination between the routing and MAC layers.

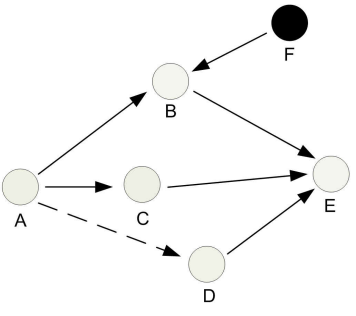


Fig. 1. Adaptive MAC protocol based on path-diversity routing

We present a new approach called anycasting [1], where multiple route information is provided to the MAC layer which is in charge of the decision on which link to forward the frame and the MAC layer must take advantage of a multiple path routing protocol. Typically, the routing protocol in the network layer decides one route out of the several paths for data forwarding, and then the MAC layer is responsible to deliver frames to the next hop along the decided route. However, let the network protocol compute multiple routing paths from the source and also from the intermediate nodes to the destination. A better approach in the MAC layer is to decide the next hop among the multiple next hops by the link status. In [2], we proposed a new data-forwarding protocol in MAC layer for vehicular ad hoc networks, where frames are forwarded to the next intermediate node without intervention of the network layer. To improve the data-forwarding performance exploiting the link status in MANETs, however, the MAC protocol requires some operational coordination between the routing and MAC layers [3]. The goal of this paper is to develop a cross-layer technique of the MAC layer, where multipath routes are discovered in the routing protocol and the virtual carrier sense mechanism is improved in the MAC protocol.

Our routing protocol is based on the Signal Power Adaptive Fast Rerouting (SPAFAR) protocol [4] which consists of two phases namely the route discovery phase and the route maintenance phase. We modify a route discovery algorithm to find multiple paths at the source and the intermediate nodes. While such a MAC layer protocol can be designed in many ways, a proper way to design is an extension of the widely-used IEEE 802.11 [5] MAC protocol.

The remainder of this paper is organized as follows. Section 2 describes the background information such as the multi-path routing protocol based on SPAFAR and the overview of the IEEE 802.11 MAC protocol. The scheme of the proposed cross-layer enhancement of the IEEE 802.11 MAC protocol in MANETs is described in Section 3. In Section 4, the simulation results for performance comparison are shown. Finally, concluding remarks are given in Section 5.

2 Preliminaries

In this section, we start by briefly reviewing the multipath routing protocol to design a cross-layer enhanced protocol of IEEE 802.11 MAC based on multipath routing information. Then, the distributed coordination function (DCF) of IEEE 802.11, the MAC layer functionality, is briefly described.

2.1 Multiple-Path Routing Protocol Based on SPAFAR

Each node of ad hoc networks keeps a *Neighbor_Table* (NT) which has an updated list of its neighbors. The NT can be easily obtained by periodic broadcasts of the beacon. Each node also keeps a *Routing_Table* (RT) which has an updated list of all the possible routes to all the potential destinations. The RT is constructed by an on-demand routing algorithm. Each element in the RT is a five-tuple of the form <*src, dst, nxt1, cnt1, nxt2, cnt2, …*>. The *src* and *dst* fields contain the unique addresses of the source and the destination node, respectively. The *nxt* field contains the address of the neighbor node to which data packets need to be forwarded. The *cnt* field contains the number of intermediate nodes from the source to the destination node on this route.

The SPAFAR protocol consists of two distinct phases, the route discovery and maintenance phases. We modify the route discovery mechanism to find multiple paths from a source and intermediate nodes to a desired destination node. When a source wants to send data to a destination and its RT does not have route information to the destination, the source initiates the route discovery mechanism to find all possible paths to the destination. The route discovery mechanism is based on request-reply operations.

An *R_Request* packet is used for the request operation from the source node and carries <*src, dst, rq_id, int_node, hop_cnt*> information. The *src* and *dst* fields contain the addresses of the source and destination respectively. The *rq_id* field contains a unique identifying number generated locally to distinguish it from the other Route Request packets. The *int_node* field keeps a sequence of all the intermediate nodes from the source to destination, while the packet traverses to the destination. The *hop_cnt* field contains the number of intermediate nodes between the source and destination. In response to a *R_Request* packet, a *R_Reply* packet is sent from the destination node and carries <*src, dst, rq_id, int_node, hop_cnt*> information. The source

field contains the address of the node that sends the *R_Reply* packet. The destination field contains the address of the node which sent the *R_Request* packet. The fields, *rq_id*, *int_node* and *hop_cnt*, contain the packet identifier, sequence of nodes from the destination to the source and the number of hops, respectively. The *int_node* field in the *R_Reply* packet is the reverse of that received in the *R_Request* packet. The detailed operation of the routing protocol is described in [4] and [6].

2.2 IEEE 802.11 Distributed Coordination Function

The IEEE 802.11 MAC protocol defines two modes of operation: Distributed Coordination Function (DCF) which allows contention access for wireless media and Point Coordination function (PCF) which requires centralized access points. DCF uses a channel access mechanism known as Carrier Sense Multiple Access with Collision Avoidance (CSMA/CA). Carrier sense is performed by a combination of physical and virtual carrier sense mechanisms. A node with packets to transmit first senses the medium. If the medium is idle for at least a certain period, DIFS, it will immediately request the channel by sending a control frame, Request to Send (RTS), to the receiver node. If the receiver correctly receives RTS, it will reply with a short control frame Clear to Send (CTS). Once the transmitter receives CTS, it will start to transfer a data frame. After the successful reception of the frame, the receiver sends an ACK to the transmitter. The exchange of RTS/CTS prior to the actual data transmission reduces the collision probability in a distributed manner and copes with the hidden terminal problem [5].

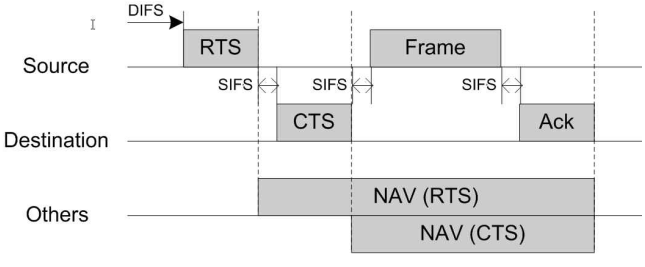


Fig. 2. Operation of IEEE 802.11 DCF

3 Next-Hop Selection Utilizing Multi-path Information

The MAC layer can acquire the information about possible next-hop options from the upper layer, and its responsibility is to transmit frames to any one of these receivers successfully. The modification of 802.11 DCF still uses the CSMA/CA algorithm, but takes advantage of multiple receivers to transmit the frame to any one of them. The routing protocol computes multiple routes between the source and destination. At each hop, the routing layer passes on the multiple next hop information to the MAC layer. The transmitter multicasts the RTS (MRTS) to these multiple receivers, and it

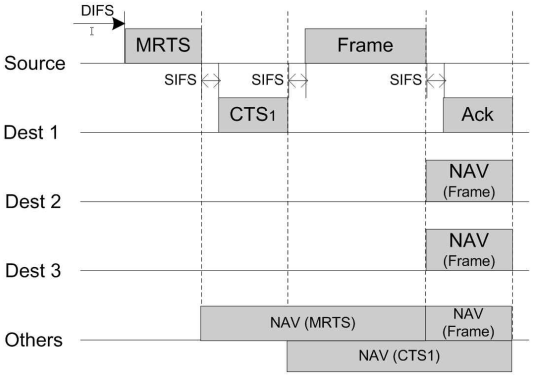


Fig. 3. First receiver's response in the cross-layer enhanced protocol of IEEE 802.11

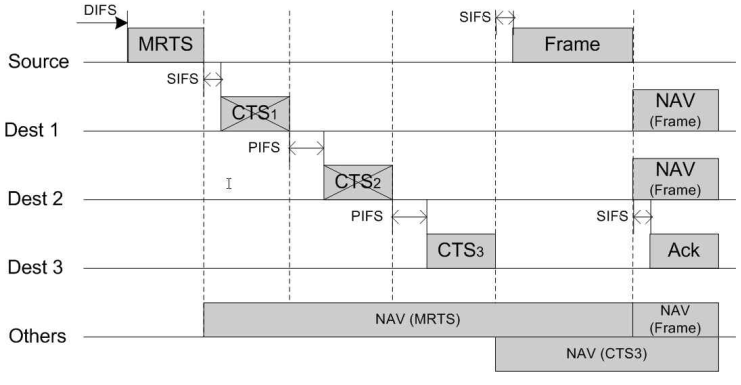


Fig. 4. Third receiver's response in the cross-layer enhanced protocol of IEEE 802.11

contains all the MAC addresses of next-hop multiple receivers. Because of practical implementation considerations, we may limit the number of next receivers to a maximum of three as shown in Figure 3 and Figure 4.

By positioning the addresses of three next receivers onto the MRTS frame, we can assign a priority order to each next hop. The priority can come from the routing or any lower layer. In the case that a shorter path to the destination gets higher priority, the routing decision in the network layer is the crucial metric for the priority. On the other hand, the information from the physical layer can be utilized to decide the priority based on the next hop that has less number of packets waiting in the queue or that has better signal strength. A combination of the above can also be used.

When the MRTS frame is broadcast to all the neighbors and all intended receivers receive the MRTS packet, the receivers respond by CTS. These CTS transmissions are intentionally delayed in time according to their priorities. The first receiver in the priority order tries to transmit the CTS after an SIFS if possible as shown in Figure 3. The second transmits the CTS after the period equal to the time to transmit a CTS, an SIFS and a PIFS if there is no transmission on the channel from the transmitter. The

third receiver transmits the CTS after the period equal to the time to transmit a CTS, two SIFSs and two PIFSs as shown in Figure 4. When the transmitter receives the CTS from the first receiver in the priority order, the transmitter transmits the DATA frame after an SIFS interval to the sender of this CTS, as shown in Figure 3. This ensures that other, lower priority receivers hear the DATA before they send CTS and suppress any further CTS transmissions. If the first and second receivers fail to transmit the CTS and the third receiver transmits the CTS, the transmitter finally forwards the DATA frame to the third receiver as shown in Figure 4.

All the receivers hearing a CTS from any intended receiver then set their NAV until the end the ACK instance. These receivers successfully sense the carrier with the exact value of NAV. Any receiver hearing only the MRTS set the maximum NAV value with the MRTS, because the total time to deliver a DATA frame cannot be guaranteed. This duration depends upon the number of receivers (a maximum of three in this paper) to which MRTS is being sent. The value of NAV set by MRTS is updated by the DATA frame which has the exact value of NAV. Furthermore, the usage of both MRTS and CTS help any other receivers identify themselves the exposed nodes or hidden nodes, and the any other receivers hearing only the MRTS set the NAV value by the DATA frame.

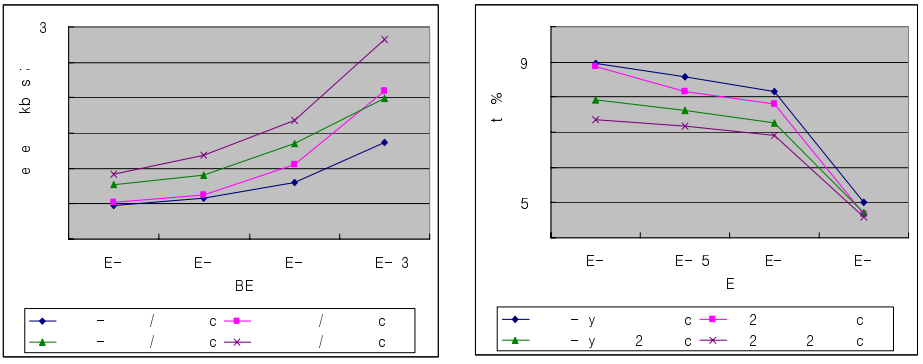
If none of the CTSs are received successfully, the transmitter goes into a random backoff and then retries again with the same receivers as in IEEE 802.11. Note that the protocol reduces to IEEE 802.11 when there is only one next hop receiver, and that when multiple next hops are indeed available and the CTS from the highest priority receiver is received successfully, this would be exactly the same as IEEE 802.11.

4 Performance Evaluation

For the simulation, the network consisting of 25, 50, 75 and 100 mobile nodes over a fixed size 100m × 100m terrain was considered. The maximum transmission power range is assumed to be 20m between nodes. A two-state Markov model [7] is used to represent the error behavior of slowly fading channels in wireless networks. The channel of bit-error rates from 10^{-6} to 10^{-3} is applied to the above simulation environment. The traffic model uses constant bit rate (CBR) traffic along with randomly chosen source-destination pairs. A traffic rate of 1 packet/sec (512 byte packet) per flow was used in the simulation. Load is varied by varying the number of traffic.

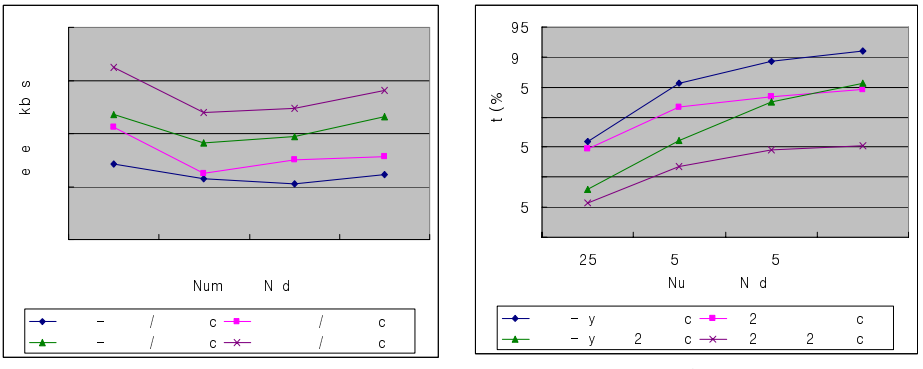
Nodes were initially placed randomly within the fixed-size physical terrain. They move to a randomly selected destination within the terrain with a constant speed (1 meter/sec). The source and destination are selected randomly, and they also move and stop continuously in random directions during the whole simulation.

The priority is decided according to the number of hops from the source and the destination. The next hop link on the shortest path has highest priority. In order to consider the temporal changes of link status, a link is marked down and the next shortest alternative is used when the number of transmissions on the link exceeds the maximum retry count. A route error is generated only when all alternatives are exhausted.



(a) (b)

Fig. 5. Simulation Results – (a) Overhead vs. BER (b) Goodput vs. BER



(a) (b)

Fig. 6. Simulation Results – (a) Overhead vs. Node density (b) Goodput vs. Node density

Figure 5 and 6 represents the simulation results in terms of the overhead and goodput according to the bit-error rate and the number of nodes, respectively. In Figure 5-a, it is shown that the proposed MAC protocol based on path-diversity remarkably reduce the overhead generated by RTSs and CTSs. In particular, the improvement increases under worse channel conditions with higher bit-error-rates. Figure 5-b also represents that the proposed MAC protocol outperforms the IEEE 802.11 MAC protocol in terms of the goodput. The metric, goodput, is defined as follows; $Goodput = \frac{\text{Number of received frames in sequence}}{\text{Number of transmitted including retransmissions}}$.

Furthermore, the performance improvement becomes larger with more traffic sources, that is, when the network is highly loaded. We can conjecture that path-diversity alleviates the congested situation in the network.

Through the simulation, we can see the IEEE 802.11 MAC protocol exploiting the multipath routing information improves the network performance especially when the qualities of channels are quite low and the network is highly loaded. The performance

improvement stems from the reduction of interactions between the MAC and network layers, and simultaneous routing deployments of multiple traffic sources.

5 Conclusion

We have proposed an enhanced IEEE 802.11 MAC protocol based on the cross-layer coordination to improve the performance of MANETs. The proposed protocol requires the cross-layer interaction between the MAC and network layers, and a routing protocol which discovers multiple paths from a source to the destination such as SPAFAR. The routing protocol in the network layer provides multipath information to leave the forwarding decision to the MAC layer. This cross-layer coordination increases the network performance with a reduced overhead cost.

The cross-layer enhancement of IEE 802.11 protocol we have proposed here can be applied to wireless sensor networks where the node density is quite high and node failures occur very often. Furthermore, the idea may be applied to wireless mesh networks where multipath routing information can be combined with an opportunistic routing protocol.

References

1. Choudhury, R., Vaidya, N., MAC-layer anycasting in ad hoc networks. ACM SIGCOMM Computer Communication Review, Vol. 34, 1(2004)75-80
2. Lee, W., Lee, H., Kim, K.: Packet forwarding based on reachability information for VANETs. ICOIN, 1(2007)
3. Jain, S., Das, S.: Exploiting Path Diversity in the Link Layer in Wireless Ad Hoc Networks. IEEE WoWMoM, 6(2005)22-30
4. Hwang, Y., Lee, H., Varshney, P.: An adaptive routing protocol for ad-hoc networks using multiple disjoint paths. IEEE VTC, 5(2001)2249-2253
5. IEEE Wireless LAN medium access control (MAC) and physical layer (PHY) specifications, IEEE standard 802.11–1997,(1997)
6. Lee, H., Lee, J., Cho, J.: An adaptive MAC protocol based on path-diversity routing in ad hoc networks. IEEE ICACT,2(2007)
7. Gilbert, E.N.: The Capacity of a Burst-Noise Channel. Bell System Technical Journal, 9(1960)1253–1265

An Incremental Topology Control Algorithm for Wireless Mesh Networks*

Mani Malekesmaeili[1,**], Mehdi Soltan[2], and Mohsen Shiva[3,***]

[1,3] School of Electrical and Computer Engineering, University of Tehran, Tehran 14399, Iran
[2] STAR Lab., Stanford University, Stanford, CA, USA
m.malekesmaeeli@ece.ut.ac.ir, msoltan@wireless.stanford.edu, mshiva@ut.ac.ir

Abstract. Wireless Mesh Networks (WMNs) with a promising future in wireless backhauling applications have attracted the attention of many network research centers and academic researches. High throughput and low delay as essential requirements of a wireless back-bone network demand new topology control algorithms as well as new optimization metrics. In this paper, a topology control algorithm compliant with the structure of WMNs is proposed, as well as a new metric combining the characteristics of different network layers. The proposed metric is compared with the conventional ones such as distance and interference where even an increase of over 100% in the capacity of the network is presented by applying this metric.

Keywords: Ad-hoc networks, collision domain, collision domain based topology control, topology control, wireless mesh networks, wireless backbone.

1 Introduction

Recently, Wireless Mesh Network (WMN) [1], as a promising newcomer in wireless access technology, has caught the attention of networking industries. Constructing a wireless backbone that provides access to Internet (or any database) is probably the most prominent application of WMNs. Large capacity and low delay as the two essential requirements in WMNs, as well as any backhaul network, demand new approaches for Topology Control (TC) and routing other than the ones available for ad-hoc wireless networks. In this paper, we focus on topology control in WMNs and introduce a new metric compliant with the needs of WMNs.

Considering the traffic pattern of WMNs, the existing TC algorithms do not suit WMNs in terms of throughput and fairness. The need for new algorithms, conforming to the requirements and characteristics of WMNs, is, therefore, perfectly felt. Other than the algorithms, the existing metrics such as the average link length, mean power, or even the number of interfering nodes, do not practically aim at fairness and throughput maximization. In conventional TC algorithms, for ad-hoc networks, direction of data streams on a link affects neither the metric nor the algorithm's procedure. While data

* This work was supported by Iran Telecommunication Research Center (ITRC).
** Student Member IEEE.
*** Member, IEEE.

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 490–497, 2007.
© Springer-Verlag Berlin Heidelberg 2007

directions are randomly distributed and are not of importance in ad-hoc networks, data streams in backbone WMNs [1] are mainly from gateways to end users. Considering this fact, directions of data streams are of great importance in WMNs; so either the existing TC algorithms should change to match the WMN applications or new TC algorithms should be developed. A change in a conventional algorithm can be a change in its metric. In this paper, we introduce a new metric called the Collision Domain (CD). In [2], CD is introduced as a tool for estimating the capacity of WMNs. It is shown that with certain modifications, CD can be used as a metric for TC to optimize the WMNs' performance. Thus, in so doing a new algorithm should also be developed.

The paper is organized as follows: in Section 2, we briefly refer to TC in ad-hoc networks. In Section 3, CD is firstly introduced (with certain changes to the original definition in [2]) and later the CD based topology control algorithm (CDTC) is presented. In Section 4, the simulation results are given, and finally the paper is concluded in Section 5.

2 Topology Control in Ad-Hoc Networks

The term of topology control has not been precisely defined in wireless networks; however, any algorithm that deals with the various aspects of the topology of a network, including nodes' positioning, network construction, link adoption, channel assignment, power control, antenna beam-forming, and even routing can be counted as topology control. Finally, a TC algorithm, directly or indirectly, provides the prerequisites for having a connected network.

There exists lots of works and papers on topology control for multi-hop wireless networks, and many different algorithms corresponding to their special requirements are developed. These requirements stem from the specific applications of the network, e.g. sensor networks or peer-to-peer, and vary from low power consumption and low number of hops to low congestion, and low delay. Full connectivity is a must in almost all of these algorithms, i.e. these algorithms aim at a full connected network with the minimum delay or the minimum average number of hops, etc. A great number of topology control algorithms for ad-hoc networks are developed to reduce transmission power in the network. Most of these algorithms take link lengths as a metric (as it is related to power with an exponent factor of 2 to 4) and claim they reduce interference implicitly. Local Minimum Spanning Tree [3], cone-based [4], and relay region based topology control [5] are among the known examples of these algorithms. It is pointed out, in [6] and [7], that reducing power does not necessarily decrease interference; they have attempted to enter interference directly into TC. These two papers and some others have each proposed a method for measuring the interference of a link. A basic model for measuring the interference of a link as a metric for TC is simply the number of active nodes in the transmission range of either ends of that link. In [8], a robust interference model is proposed for wireless ad-hoc networks. For more details on the models, and also the algorithm procedures one can refer to [6, 7, and 8]. While interference is a suitable metric for most ad-hoc networks, however, it does not match the properties of WMNs completely. Although there have been some researches in TC for WMNs, nonetheless they are mostly involved with either the channel assignment [9] or the directional antennas. The existing commercial wireless network interfaces support one channel per radio; so for a multiple channel network, the number of

network interfaces should be as many as the number of the channels in a node, thus raising the cost of its implementation. Directional antenna designs have not only increasing costs, but also their dynamic pattern forming techniques have not been fully developed yet. Considering the above, a practical TC algorithm for current WMNs, should focus on the link adoption and the power control on single channel networks. To the best of our knowledge, there has not been any attempt for TC devoted to WMNs in single channel, omni-directional networks (a common case in the existing wireless multi-hop networks). This paper may be the first of its own in topology control for single-channel WMNs. In the next section, the collision domain (CD) load as a new metric that is well suited for a backbone WMN is introduced.

3 Topology Control in WMNs

3.1 Collision Domain Load

The idea of collision domain has stemmed from a paper by Jangeun et al [2]. Authors introduced the concept of collision domain for predicting the capacity of WMNs, wherein a unidirectional traffic from nodes to gateways on the network is assumed and the capacity of a WMN is estimated. In a backbone WMN, traffic is mostly from gateways to nodes (download traffic). In the following we introduce CD [2] with certain changes to have it well-matched to the characteristics of WMNs infrastructure.

CD of a link is an area where all other nodes or links should be silent (idle) for a successful transmission on that link. The CD load of a link is the total amount of data that is to be transmitted on that domain including the link itself. This load is simply referred to as the CD in the rest of the paper. The maximum available download rate can be estimated by the maximum existing collision domain load. As downloading from gateways (file downloading, web browsing, watching videos online, and etc.) is the dominant traffic load in an access network, like WMN, the maximum available download rate can be counted as the nominal capacity of a wireless network. The lower is the maximum CD, the larger will be the capacity of the network. Rightfulness of the above statement is proved in [2] by the simulations. It is shown that the network capacity can be precisely estimated using the maximum CD. For more information on assumptions and conclusions one can refer to [2]. CD not only shows the interference but also represents the load of the interfering nodes. For a WMN backhaul network, power consumption is not of much importance, thus, a minimum power TC algorithm does not suit these networks. Considering the above, it can be concluded that CD is superior to any of the other metrics proposed so far. It states the characteristics of a link more comprehensively by combining the Media Access Control (MAC) characteristics (like collision), with the amount of load obtained from the application layer, and routing rules (the network layer characteristics).

From the above it can be inferred that CD itself is dependent on the topology and the route that transfers the load to a specific node. Therefore, the existing TC algorithms cannot be employed for a topology construction with CD as the metric. It seems that because of CD's dependency on the TC result, there exists no direct optimal solution to the problem. Heuristic methods like the Genetic Algorithms or Ant Colony can be considered as sub-optimal solutions to this kind of "result depends on procedure" problem. For limited area low density networks, with a logically small

number of available links, these methods are fine and will generate near optimal results; but for larger areas and a high number of links, as the number of variables grow, these methods will become almost impractical. Number of local minima or maxima of a real world problem (including TC) grows almost exponentially with the number of the variables. This causes an exponential increase in the complexity and run-time of a heuristic optimization problem along with a decrease in the optimality of its algorithm due to the increase in the probability that the algorithm gets stuck in a local extreme.

An important characteristic of a wireless mesh backbone (and also generally any ad-hoc network) is its self-reconfiguration, self-healing, and self-reconstruction abilities. This means that in the case that a route (e.g. a link,) or a node, (e.g. a router) fails, the network should be able to reconstruct itself and the services should not be interrupted. Therefore, a TC algorithm used in WMN should be able to be run many times, fast enough, and with the least change. This is not the case in time-consuming heuristic methods, which even for identical cases may generate totally different results.

In the next section, we introduce a new TC algorithm that uses CD as a metric. In addition to the above, the algorithm itself is devoted to WMNs. Other metrics like transmission power, interference, etc. can be used with this algorithm. The developed algorithm is sub-optimal, but runs fast enough and the result of the algorithm does not change fundamentally with small changes. The results in Section V show the superiority of CD to the other available metrics.

3.2 Collision Domain Based Topology Control (CDTC)

As pointed out in Section 2, for a TC algorithm to be practical, it should consider link adoption and power management for single channel multi-hop wireless networks. A new metric conforming to the characteristics of a wireless backhaul network based on the collision domain of links was proposed in the previous section. Here, we present a TC algorithm based on CD. It was mentioned before that CD is itself dependent on routes and topology; so, developing a topology control algorithm that minimizes the maximum collision domain load for any topography is impossible. A practical way for handling such a problem is to ignore optimality and look for the best sub-optimal practical algorithm. Our proposed algorithm is also a suboptimal one shown in Fig.1. The proposed algorithm not only determines the links to be used, but also designates the routes from any node to gateways. Therefore, other than TC it can perform routing and can eliminate the need for a routing algorithm for download traffic.

The proposed algorithm, as shown in Fig. 1, assumes an initial local-weight for non-gateway nodes. This local-weight is a nominee of the local traffic of each node; assuming all nodes request (can request) the same amount of traffic (this is fairness by some means), the weight of unity is applied to all nodes except the gateways. This weight can be refined further in an operating network using a feedback from the application layer.

For the proposed algorithm to be more compatible with specific traffic pattern protocols of higher levels, the required load can be estimated and sent to TC as its local-weight. The topology refreshes as significant changes occur in local-weights. The other optimization parameter, the transient-weight of a node, is an accumulation of the local-weight of nodes that receive their traffic through this node. This weight distinguishes CD from interference.

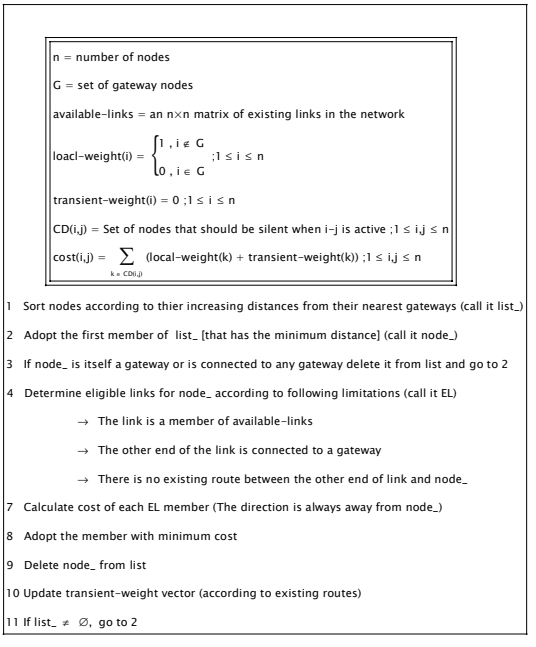


Fig. 1. Collision Domain based Topology Control (CDTC)

This algorithm is designed for WMNs and aims at a topology with the maximum available capacity. It uses CD as its optimization metric, however, any other conventional metric that is used for TC can also be used in this algorithm. For other metrics only the cost function should change. If distance is the metric, cost of a link between nodes "i" and "j" [cost (i,j)] will be simply the distance between these two nodes; and the interference can be simply calculated by omitting the transient-weight factors from the formula in Fig.1. Whatever the metric, the algorithm will generate a topology that connects each router to a proper gateway. It dictates both the links and routes to be chosen for connecting a router to its gateway, but it does not generate routes for connecting any arbitrary pair of nodes. So, peer to peer traffic that has a small share in the traffic pattern of backhaul networks can be handled using the existing ad-hoc routing algorithms. As we have only considered the unidirectional download traffic from gateways to nodes, the maximum CD in the network determines the nominal capacity of the network. It is proved in [2] that under certain conditions, which are also fulfilled in our network, the nominal capacity of a WMN can be calculated as the nominal MAC layer capacity divided by the maximum CD in the network. Thus, minimizing the maximum CD will maximize the capacity. In the section that follows, the superiority of the proposed metric to distance and interference in terms of the maximum CD is presented.

4 Simulation Results

In this section, we compare the proposed metric (CD) is compared to two other metrics, namely interference and power (distance), using the algorithm shown in

Fig. 1. In the simulations we have considered the download traffic. As mentioned in the previous sections, in a backbone WMN, download traffic dominates the other types of traffic such as the peer-to-peer traffic. It is this traffic pattern that differentiates WMNs from stationary ad-hoc multi-hop networks. So, for an algorithm to be suitable for WMNs this type of traffic should be considered. In [2], a method is proposed and confirmed to estimate the nominal capacity of a WMN. As our simulations match the assumptions of the said paper, we use the method proposed there to calculate the nominal capacity of our presented topologies.

Simulations are done for backbone WMNs, i.e. clients of the final hop are not considered. Since each router acts as an access point for its clients, and all clients of a specific router can only communicate with other clients using their access point, the responsibility of a backbone WMN is to deliver data packets of any client to their corresponding access point. This is why considering or not considering the last hop (from access point to client) does not affect the performance of a WMN. Fig. 2 shows a sample topology resulted from CDTC. 49 nodes (routers) are distributed distancing 120 to 200 meter from adjacent nodes in a square shaped area.

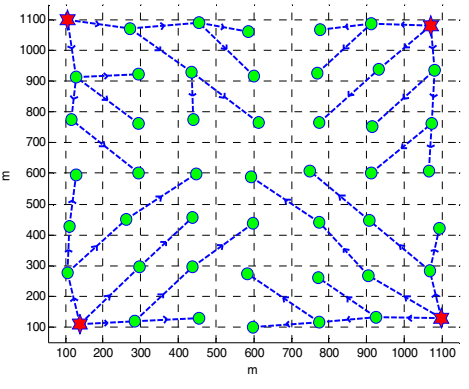
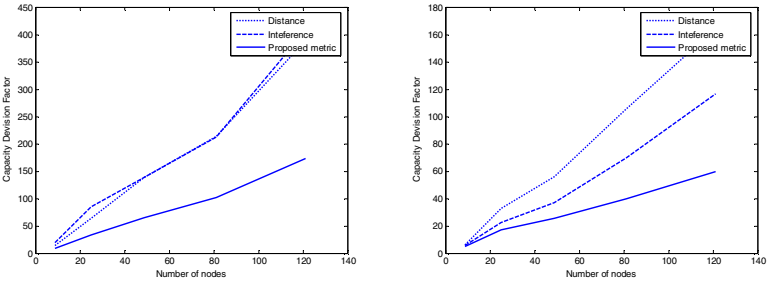


Fig. 2. CDTC result for a 40 node network with 4 gateways in the corners

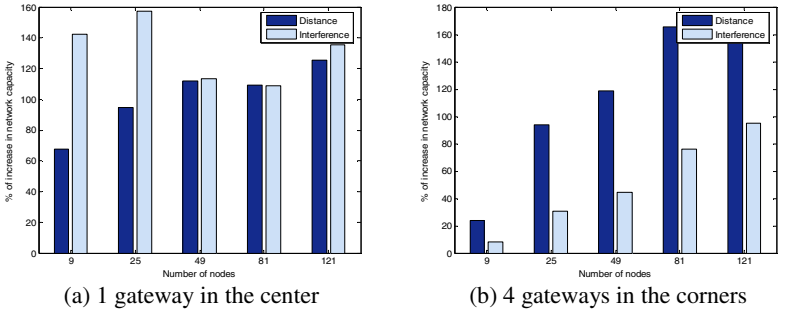
The performance of the proposed metric in terms of the maximum CD is compared to the other metrics in Fig. 3. As pointed out before, for the specific conditions of our simulations, this load is a direct indicator of the nominal capacity of the network, in which the capacity will be the nominal MAC layer capacity divided by this load. So, if the maximum CD is minimized, the capacity (the maximum download rate) of the network will be maximized. The graphs in Fig. 3 show the maximum CD for different number of nodes. The area is again square shaped, and the density of the nodes in the area is constant, i.e. increasing the number of nodes equals to increasing the area that is covered by the WMN. For each number of nodes, the maximum CD is the result of the average of 10 different node position simulations (again nodes are placed randomly distancing 120 to 200 meters from their neighbors). In Fig. 3, results are shown for two different scenarios, in (a) the center node is the gateway and in (b) the 4 corner nodes are gateways.



(a) 1 gateway in the center (b) 4 gateways in the corners

Fig. 3. Maximum CD of the three metrics

If we assume that the nominal capacity of MAC layer is B [10], we can say that the capacity (here, the maximum download rate) is B/max(CD) (as stated in [2]). So, we can calculate the increase in the capacity of the network (when CD is the metric as compared to when interference or distance are) simply as: (max ($CD_{other\ metrics}$-max(CD_{CD}))/max(CD_{CD}). In Fig. 4, the percentage of increase in capacity (of Fig. 3 networks) using CD with respect to interference and distance is shown. We can see that in both cases CD outperforms the conventional metrics, and even in some cases the capacity using CD is more than twice the capacity obtained from the other metrics. From the above graphs we can also conclude that in most cases interference works better than distance. The figures also highlight the efficiency of the algorithm in using CD as the optimization metric.



(a) 1 gateway in the center (b) 4 gateways in the corners

Fig. 4. Percentage of increase in capacity using CD compared to Interference and distance

5 Conclusion

In this paper, a new metric compliant with the needs of WMNs is proposed as well as an algorithm designed for download traffic in backbone WMNs. The proposed metric is tested within the algorithm and it is shown that the proposed metric outperforms the conventional metrics like distance and interference. Other than link adoption, the

proposed metric defines the direction of data in the links, being bi-directional with the route for each node to its corresponding neighbor indicative of the possibility of its use as a routing algorithm.

References

1. I.F. Akyildiz and W. Xudong: A survey on wireless mesh networks, Communications Magazine, IEEE, Volume 43, Issue 9, (Sept. 2005) 23 - S30.
2. J. Jangeun and M.L Sichitiu: The nominal capacity of wireless mesh networks, Wireless Communications, IEEE, Volume 10, Issue 5, (Oct 2003) 8 – 14.
3. N. Li, J. Hou, and L. Sha: Design and Analysis of an MST-Based Topology Control Algorithm, Wireless Communications, IEEE Transactions on, Volume 4, Issue 3, (May 2005) 1195 - 1206.
4. L. Li, J.Y. Halpern, P. Bahl, Y.-M. Wang, and R. Wattenhofer: A Cone-Based Distributed Topology Control Algorithm for Wireless Multi-Hop Networks, IEEE/ACM Transactions on, Volume 13, Issue 1, (Feb. 2005) 147 - 159
5. V. Rodoplu, T. H. Meng: Minimum energy mobile wireless net-works, Selected Areas in Communications, IEEE Journal on, Volume 17, Issue 8, (Aug. 1999) 1333 - 1344.
6. M. Burkhart, P. v. Rickenbach, R. Wattenhofer, and A. Zollinger. : Does Topology Control Reduce Interference?, Proceedings of the 5th ACM international symposium on Mobile ad hoc networking and computing, Session: Geometry and positioning (2004) 9 - 19.
7. K. Wu, W. Liao: Interference Efficient Topology Control in Wireless Ad Hoc Networks, IEEE CCNC 2006, Volume 1, 8-10 (Jan. 2006) 411 - 415.
8. P.V. Rickenbach, S. Schmid, R. Wattenhofer, A. Zollinger: A Robust Interference Model for Wireless Ad-Hoc Networks, Parallel and Distributed Processing Symposium, 2005. Proceedings 19th IEEE International, (April 2005) 8
9. M. K. Marina, S. R. Das: A Topology Control Approach for Utilizing Multiple Channels in Multi-Radio Wireless Mesh Networks, Broadband Networks, 2005 2nd International Conference on, Volume 1, (Oct. 2005) 381 - 390.
10. J. Jun, P. Peddabachagari, and M. L. Sichitiu: Theoretical maximum throughput of IEEE 802.11 and its applications, IEEE International Symposium on Network Computing and Applications (NCA 2003) (Apr. 2003) 249–256

TCP Adaptation for Vertical Handoff Using Network Monitoring

Faraz Idris Khan and Eui Nam Huh

Internet Computing and Security Lab
Department of Computer Engineering, Kyung Hee University, 449-701 Yongin, South Korea
{faraz,johnhuh}@khu.ac.kr

Abstract. Next Generation Network envisions convergence of different wireless networks to provide ubiquitous communication environment to the mobile user. The ubiquity is achieved by enabling a mobile user to switch to a network with better QoS (Quality of Service), by a mechanism called vertical handoff. We propose a vertical handoff management architecture which handles TCP adaptation and ensure efficient utilization of mobile terminal resource i.e. memory. The adaptation is achieved by configuring TCP socket buffer according to the available network resource i.e. bandwidth. The simulation results show that TCP utilizes the bandwidth and memory efficiently by employing our TCP adaptation strategy.

Keywords: TCP adaptation, adaptive network QoS, vertical handoff, congestion control, TCP performance.

1 Introduction

Mobile communications and wireless networks are developing at a fast pace. The increasing number of mobile subscribers and terminals are the real life evidence of the fast growing development in the area of wireless communication. As a resultant effect an increasing demand of IP based services (e.g. e-commerce) and IP related enormous applications (e.g. WWW and email), the demand for wide band data access through wireless medium is created. Integration of different heterogeneous networks is able to diminish the wideband data access problem and also capable of minimizing the deployment cost for completely new infrastructure and access.

Among different wireless access technologies, WLAN [3] and Cellular system [4] are getting momentum during last couple of years. WLAN is mostly used to serve the hotspot region like airports, cyber cafes, coffee shops, hospitals, schools etc. for its huge data rate (11 Mbps for 802.11b, 54 Mbps for 802.11a and 802.11g) and relatively low cost.

In telecommunication domain cellular system remains the most popular wireless access technologies in communication domain. The development of cellular system resulted into different generations like 1G, 2G (GSM, DAMPS), 2.5G (GPRS, EDGE) followed by 3G (UMTS). In 1992, International Telecommunication Union (ITU) has issued International Mobile Telecommunication for year 2000 (IMT-2000) which defines the basic characteristics of 3G. The advantages of cellular system are wide

Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 498–505, 2007.
© Springer-Verlag Berlin Heidelberg 2007

coverage, well known voice service whereas low data rate and cost relative to WLAN are few of the limitations.

In order to provide the communication platform for pervasive systems the capabilities of both WLAN and cellular systems will be combined. This can be achieved with the overlay architecture in which different networks overlap each other. The objective is to provide good Quality of Service (QoS) to the mobile user all the time. This is achieved by a mechanism called *vertical handoff* which enables a mobile user to switch to a different network. The application experiences degradation in QoS due to changing network parameters such as RTT and packet loss rate. A handoff decision is made to switch to a better network which provides better QoS to the mobile user. The decision of switching to another network considers user profile, device profile and application profile in order to provide seamless service to the end user [2]. There are two kinds of vertical handoff situation that might occur. Either the mobile terminal switches from a high data rate network to low data rate network or it moves from low data rate network to high data rate network. In the former case TCP congestion window does not open up fully due to which the bottleneck remains underutilized. While in the latter case the memory allocated to a TCP connection is wasted as the new network is a low data rate network. In order to provide undisrupted service to the end user there are various vertical handoff management architecture proposed in literature [1][2][4][9] which mostly considers content adaptation after vertical handoff. There are few architecture proposed such as in [9] which incorporates TCP adaptation but they require modification in TCP.

Thus in this paper we propose vertical handoff management architecture, providing undisrupted service to the end user by incorporating TCP adaptation module.The TCP adaptation module requires no modification in TCP code but requires application level configuration of socket buffer size.

The structure of the paper is as follows. Section 2 describes the related work on vertical handoff architecture and provides a brief overview of current research considering TCP performance in wireless network and TCP buffer tuning techniques. Section 3 describes the proposed vertical handoff architecture. Section 4 contains the simulation results of the TCP adaptation logic considering vertical handoff scenarios.

2 Related Work

This section is divided into three subsections. In subsection1 vertical handoff management architecture is discussed. Subsection 2 contains work related to wireless TCP considering mobility and its adaptation in handoff scenario.

2.1 Vertical Handoff Management Architecture

Helal et al [1] proposed an architecture which considers application level adaptation and handles vertical handoff mobility by employing Mobile IP. Applications participate fully in vertical handoff decision process by making its need known which helps in application level adaptation.

Balasubramaniam et. al [2] presented a vertical handover as an adaptation method supporting pervasive computing. Adaptation is carried on multimedia streams upon a

change in context which is defined by dynamic network and application profile. Vertical handoff decision is carried out by weighing the decision parameters employing AHP (Analytical Hierarchy Process) [3] and then calculating the quality of each network in order to decide the best network.

Di Caor et. al [4] proposed a strategy of optimized seamless handover applied in WiOptimo system in which the network parameters used for vertical handoff decision are adapted according to the context. A cross layer mechanism is employed to detect the status of the connection. The vertical handoff can occur in the following cases

- Switching to another network occurs due to user's preference
- If a mobile terminal is moving to the edge of the network the signal strength might degrade and a handoff might be required
- The third case that causes switching in case the current network is not good enough to provide QoS to the running application

If we summarize the above work we can see that in all of the architectures a separate module to handle TCP adaptation is not taken into consideration. We propose a separate module in vertical handoff management architecture which handles TCP adaptation.

2.2 Advances in TCP Adaptation and Mobility

TCP is the most popular transport protocol employed in internet to control the smooth transmission of data by employing a buffering mechanism. It is designed with reliable network in mind. Thus TCP performs poorly in case of wireless network which often experiences packet losses due to mobility of the terminal which invokes congestion control algorithm to lower the transmission rate of the application. Thus in literature we can find enhancements in TCP such as in [5] [6] [7] which are proposed to differentiate between packet losses due to congestion and unreliable nature of the wireless link.

In [8] an FRA algorithm is proposed which probes the new link exponentially rather than staying in congestion avoidance phase with linear probing in case the capacity increase is large. In the opposite case they use explicit handoff notification (EHN) to tackle the packet loss due to sudden reduction in capacity. In [9] a mechanism for dynamically adjusting the congestion window is proposed to adapt TCP. The scheme shrinks the window less rigorously upon detection of packet loss during vertical handoff period in order to improve TCP throughput and retain the retransmission timeout value.

2.3 Advances in TCP Buffer Tuning

There are various mechanisms proposed in literature to improve TCP throughput by dynamically adjusting the socket buffer size.

In [11] a daemon called *Work Around Daemon* handles the tuning of the sockets by using *ping* for measuring the RTT of the path and *pipechar* to estimate the capacity C of the path. The daemon cannot cater the changing BDP due to mobility of the mobile terminal.

The first proposal for automatic TCP buffer tuning was [12]. The main objective of their work is to enable fair sharing of the kernel memory among multiple TCP

connections. The sockets are configured by considering the present congestion window to calculate the BDP of the path. The mechanism will work well for a path with constant BDP rather than for varying BDP in case of vertical handoff. In some other works such as in [13] dynamic adjustment of socket buffer tuning is proposed. The shortcomings of these mechanisms are that they do not cater the changing RTT and bandwidth during vertical handoff situation.

3 System Architecture

The architecture is designed in a modular fashion with a middleware based approach. The device, application, user, current network and application QoS profile are considered to be stored in a context repository (CR).CR loads the QoS profile which is used by different networks for QoS provisioning. The decision engine (DE) monitors the context cache for triggering the handoff upon detection of context changes which can be degradation in QoS due to changes in network characteristics, disconnection from the current network etc. The feedback from DE triggers the TCP level adaptation which is handled by TCP adaptation module (TAM). Network Resource Monitoring Agent (NRMA) agent monitors the network resources such as available bandwidth, RTT which is used by TCP adaptation module. The TCP adaptation module is discussed in detail in subsection 3.1.

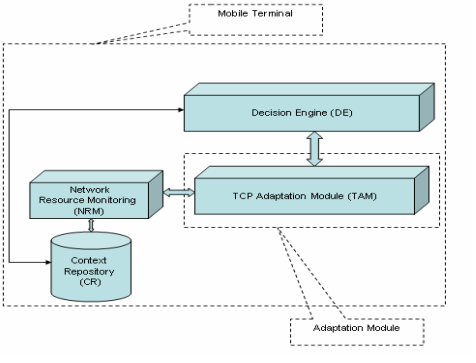


Fig. 1. System Architecture designed using a middleware based approach

3.1 TCP Adaptation Module (TAM)

The working of TAM is as follows. Let us denote the capacity of the network by C bps and the RTT by T sec. Thus in order to fully utilize the bandwidth of the network the socket buffer should be adjusted to at least $C \times T$. This will ensure that the TCP congestion window W_c will open up to saturate the network. The size of the socket buffer can be represented by S. The handoff situation can be of two types one is when a mobile terminal moves from high data rate network to low data rate network. The other is when a terminal moves from low data rate network to high data rate network. In either of the case the socket buffer size of the connection need to be adjusted

according to the available bandwidth. Mathematically the socket buffer size of the application can be calculated as follows. If we represent the current network by N_c and the new network by N_n then the algorithm of TCP adaptation is shown in figure 2.

If (a decision of vertical handoff is signaled by the decision engine)
Get available bandwidth A_{BWD} *of the network* N_n
Get RTT T of the network N_n
$S = A_{BWD} \times T$
Else
Get Capacity C of the network N_c *// during connection initiation*
Get RTT T of the network N_c
$S = C \times T$
End if

Fig. 2. TCP Adaptation Algorithm

4 Simulation

For TCP adaptation we simulated our proposed idea in *ns-2* by implementing vertical handoff scenario. In this simulation we assume that the data rate with in 3G cellular network is 144kbps with end-to-end RTT 300msec and data rate with WLAN is 2Mbps with end-to-end RTT 100 msec.

In ns-2 simulation, the total duration was of 140 seconds. The mobile terminal moves from low data rate network to high data rate network. We assume that the vertical handoff occurs at 50 sec and completes at 54 sec.

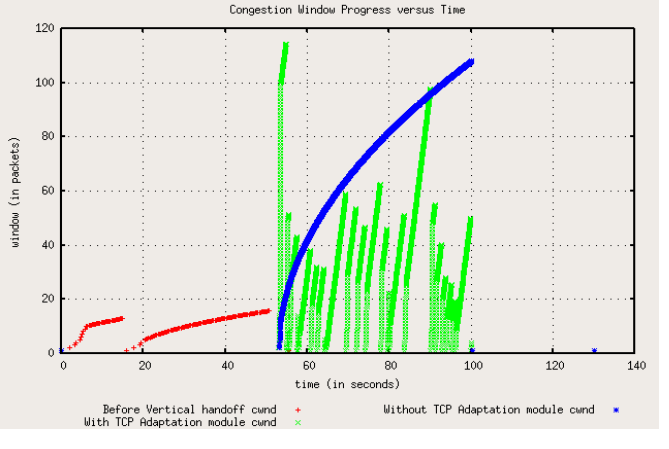


Fig. 3. Congestion Window Progress during vertical handoff from low data rate network to high data rate network

There is a 65% improvement in the bottleneck utilization of the link. Figure 3 shows TCP window progress with and without the TCP adaptation module during vertical handoff. By configuring the buffer size, the newly available bandwidth is adapted swiftly but the TCP connection and it takes 4 seconds to hit maximum bandwidth as shown in figure 3.

4.1 High Data Rate Network to Low Data Rate Network

Figure 4 and Figure 5 shows the memory utilization of a TCP connection with and with out TCP adaptation module. The memory remains largely underutilized because the socket buffer is not adjusted according to the bandwidth delay product. By applying TCP adaptation module the memory utilization of a mobile terminal is improved in our case by 40 %.

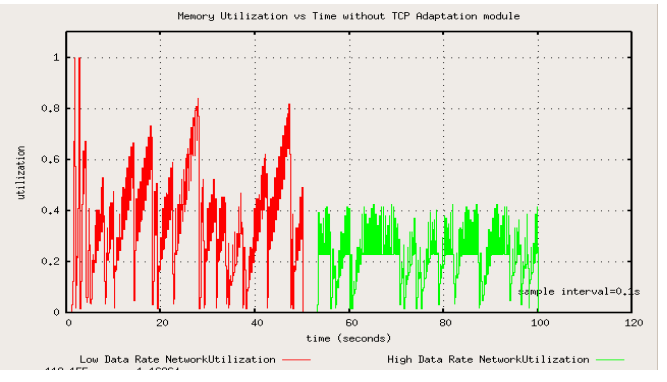


Fig. 4. Figure shows the memory utilization of the system with out TCP Adaptation module during handoff

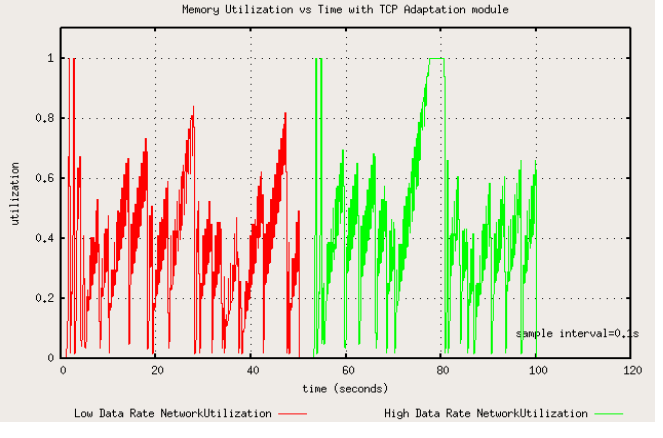


Fig. 5. Figure shows the memory utilization of the system with out TCP Adaptation module during handoff

5 Conclusion

In this paper we propose vertical handoff management architecture with TCP adaptation for pervasive systems. We have incorporated a TCP adaptation module which uses socket buffer tuning mechanism to improve the TCP performance and maintain system performance by conserving system resources i.e. memory. The adaptation mechanism is initiated by feedback from DE in cooperation with NRMA which monitors the network resource. For our future work we will investigate the CPU resource consumption after vertical handoff situation. The ultimate objective is to develop a vertical handoff management architecture which provides QoS to the end user by maintaining application performance and ensure efficient utilization of resources.

Acknowledgements. This research was supported by MIC (Ministry of Information and Communication), Korea, under ITRC (Information Technology Research Center) support program supervised by the IITA (Institute of Information Technology Advancement). (IITA-2006-C1090-0603-0040).

References

1. Helal, S., Lee, C., Zhang, Y.G., Richard III, G.G.: An Architecture for Wireless LAN/WAN Integration. *WCNC 2000* IEEE Wireless Communication and Networking Conference Chicago, 9 (2000)
2. Balasubramaniam, S., Indulska, J.: Vertical handover supporting pervasive computing in future wireless networks. Vol. 27 Computer Communications, 5 (2004) 708 – 719
3. Song, Q., Jamalipour, A.: A Network Selection Mechanism for Next Generation Networks. *IEEE ICC'05*, Seoul 5 (2005)
4. Gianni, A., Caro, D., Giordano, S., Kulig, M., Lenzarini, D.: A cross layer and autonomic approach to optimized seamless handover. IFIPS WONS France, 1 (2006)
5. Bakre, A., Badrinath, B.: I-TCP: Indirect TCP for mobile hosts. IEEE ICDCS (1995) 136-143
6. Brown, K. Singh, S.: M-TCP: TCP for mobile cellular networks. ACM Computer Comm. Review (CCR) Vol. 27(1997)
7. Balakrishnan, H., Padmanabhan, V.N., Seshan, S., Katz, R.H.: A comparison of mechanisms for improving TCP performance over wireless links. IEEE/ACM Trans. Networking Vol. 5. IEEE Computer Society (1997) 756-769
8. Chen, L.J., Yang, G., Sun, T., Sanadidi, M.Y., Gerla, M.: Enhancing QoS Support for Vertical Handoffs Using Implicit/Explicit Handoff Notification. QSHINE IEEE Computer Society (2005) 37
9. Kang, R.J., Chang, H.P., Chang R.C.: A seamless vertical handoff scheme. WICON IEEE computer society (2005) 64 - 71
10. Dunigan, T., Mathis, M., Tierney, B.: A TCP Tuning Daemon. SuperComputing: High-Performance Networking and Computing Vol. 25. IEEE computer society, Baltimore Maryland (2002) 1 - 16
11. Semke, J., Madhavi, J., Mathis, M.: Automatic TCP Buffer Tuning. ACM SIGCOMM Vol. 28(1998) 315 - 323

12. Gardner, M.K., Feng, W.C., Fisk, M.: Dynamic Right-Sizing in FTP (drsFTP): Enhancing Grid Performance in User-Space. IEEE Symposium on High-Performance Distributed Computing (2002) 42 - 49
13. Prasad, R., Jain, M., Dovrolis, C.: Socket Buffer Auto-Sizing for High-Performance Data Transfers. Journal of Supercomputing Vol. 1. (2003) 361-376

Optimization of Mobile IPv6 Handover Performance Using E-HCF Method*

Guozhi.Wei[1], Anne.Wei[1], Ke.Xu[2], and Gerard.Dupeyrat[1]

[1] University of Paris XII 61, Avenue du General de Gaulle 94010 Créteil France
[2] Tsinghua University 100084 Beijing China
{guozhi.wei,wei,g.dupeyrat}@univ-paris12.fr,
xuke@mail.tsinghua.edu.cn

Abstract. Mobile IPv6 permits the Mobile Node (MN) to maintain the continuous connectivity to the Internet when it moves from one access router to another. Due to both link switching delay and IPv6 protocol operation during the handover process, packets destined to the MN can be delayed or lost. This paper proposes a solution to improve the Mobile IPv6 handover performance over wireless network by introducing a new entity - Extension Handover Control Function (E-HCF). E-HCF could send the decisive control message to MN to accelerate the handover process and manage the traffic belonging to MN to reduce the packet loss. With a comparison between Mobile IPv6 and our E-HCF solution, we show that our solution allows us providing low-latency and low packet loss for real-time services during the handover.

Keywords: Mobile IPv6, Handover, E-HCF function, WLAN.

1 Introduction

The need of keeping connection with Internet in everywhere and at every time is more and more demanded in recently years. However, the continuous Internet connectivity and the correct routing of packets could not be guaranteed when users change their access point to the Internet. To resolve these problems, the protocols Mobile IPv4 (MIPv4) [1] and Mobile IPv6 (MIPv6) [2] are respectively proposed by the Internet Engineering Task Force (IETF). The major difference between MIPv4 and MIPv6 is that Foreign Agent (FA) is elided in the latter; moreover Mobile Node (MN), its Home Agent (HA) and its Correspondence Node (CN) must all support the IPv6 Protocol. The MIPv6 operations involve movement detection, router discovery, Care of Address (CoA) configuration, Duplication Address Detection, and Binding update. Although MIPv6 is proposed in the interest of improving the handover performance, the latency and packet loss are always two main problems, which affect the real-time application employing for mobile user.

Along with the widely implantation of Wireless LAN (WLAN) [3], the user of WLAN would like to have the more mobility. To meet this demand, the MIPv6 is needed to be better improved over the WLAN. While the infrastructure networks are already installed, such as Access Router (AR), Access Point (AP), they are hardly

* This work was supported in part by the international project PRA-SI under Grant SI04-03.

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 506–513, 2007.
© Springer-Verlag Berlin Heidelberg 2007

updated or modified. Under this background, we introduce a new entity - Extension Handover Control Function (E-HCF) in Mobile IPv6 over WLAN to improve the handover performance without changing the existing infrastructures. As the probe phase (MN launches the scanning process to find the available AP in WLAN.) [3] and Detection Address Duplication (DAD) process are the main negative influences on handover latency, E-HCF aims to improve the handover performance by reducing their effects. Moreover, to reduce the packet loss, E-HCF could buffer, and then redirect the traffic according to the MN's need during the handover process to reduce the packet loss.

The remainder of the paper is organized as follows: Section 2 presents background and related work of Mobile IP. Section 3 presents our Extension Handover Control Function (E-HCF) architecture and the detailed protocol operation associated. Section 4 deals with the performance of E-HCF handover in term of handover latency and packet loss. The numerical results show that E-HCF handover procedure reduces significantly the latency and packet loss regarding the Mobile IPv6. Finally, conclusion and future work are mentioned in section 5.

2 Background and Related Works

Actually, the main proposals accepted by IETF are Hierarchical Mobile IPv6 (HMIPv6) [4] and Fast Handover for MIPv6 (FMIPv6) [5].

HMIPv6 introduces an entity - Mobility Anchor Point (MAP) who acts somewhat like a local HA for the visiting MN. HMIPv6 classes MN mobility into micro-mobility (within the same MAP domain or intra-MAP) and macro-mobility (between the two MAP domains or inter-MAP). When intra-MAP handover occurs, MN only needs to register its new CoA with its serving MAP, therefore, HMIPv6 could limit the amount of signal required outside the MAP's domain, decrease the signal load and the signal transmission delay, and consequently, the handover performance could be improved. While HMIPv6 introduces the additional delay for establishing the bi-directional tunnel between MAP and MN, generates the more signal load for Inter-MAP handover. Furthermore, it is difficult to determine an adaptive MAP domain size for different MN's mobility pattern [6] [7].

FMIPv6 tries to reduce handover delay by providing fast IP connectivity as soon as MN attaches to a new subnet. To realize this aim, MN must launch the passive or active scanning process to discover the available AP [8]. According to the probe results, AR provides MN with the corresponding subnet prefix information, and then MN could generate a New CoA when it is still connected to its current subnet. To minimize packets loss, a bi-directional tunnel is set up between Previous AR and New AR. Utilizing this tunnel, PAR forwards the packets destined to MN's Previous CoA to its New CoA, MN could also continue to send packets to CN through PAR. Such tunnel remains active until MN completes Binding Update with its CNs. However, there are two mains shortcomings in the FMIPv6 protocol. Firstly, MN couldn't receive or send the data during the probe phase, while it lasts minimum 350 ms [9], furthermore, MN must spend time to re-switch the channel and re-associate with its Previous AP to exchange the messages with PAR; Secondly, the DAD process could

not be completely avoided if MN's New CoA isn't validated by the New AR before MN disconnects with its previous AR.

Besides the main proposals, some approaches are proposed to providing lossless handover and minimizing the handover delay ([10], [11], [12]). IEEE 802.11f standard, known as Inter-Access Point Protocol (IAPP), has been proposed in 2003[10]. IAPP specifies the information to be exchanged between APs to support MN handover. But this standard has been withdrawn on February 2006 by the IEEE 802 Executive Committee. In [11], a Pre-Handover Signaling (PHS) protocol is proposed to support the triggering of a predictive handover and to allow the network to achieve accurate handover decisions considering different constraints, such as Quality-of-Service (QoS), the user's profile and the MN service requirements. In [12], this paper proposes to use a new entity - Handover Control Function (HCF) to pre-decide MN's new CoA, consequently, MN could send Binding Update message when it still connects with its previous AP. However HCF could not avoid absolutely the IP collision and the inter-HCF handover is not discussed. Therefore, in this paper, we use E-HCF entity to reduce the layer 2 scanning delay, avoid the duplication of IP address, and support the inter-E-HCF handover.

3 Extension Handover Control Function (E-HCF) for Mobile IPv6

3.1 E-HCF Overview

We introduce a local intelligent entity, called E-HCF, which is able to control the ARs, APs and MNs of its domain. The architecture of E-HCF is shown in Fig 1.

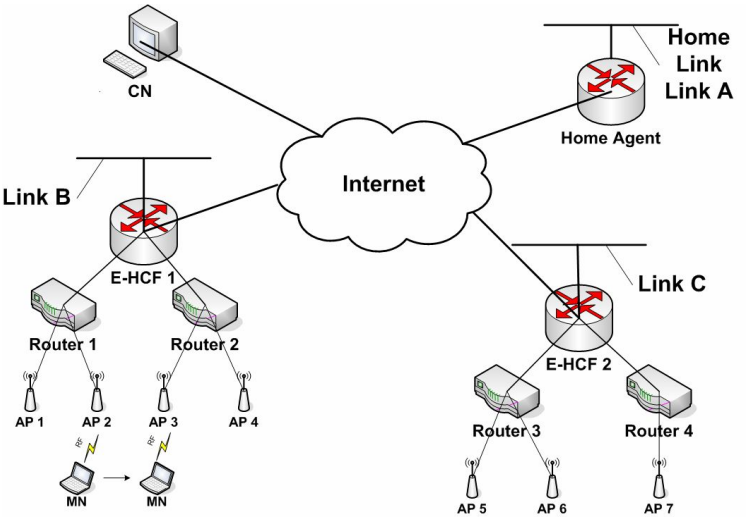


Fig. 1. Architecture of Extension Handover Control Function (E-HCF)

In order to improve the MIPv6 handover performance over WLAN, we endow E-HCF with the new functions: E-HCF could provide a list of the AP as the available

AP and a corresponding IP address as the new CoA for MN. In this way, MN needn't launch the scanning process to discovery the available AP and the Detection Address Duplication (DAD) process in the new subnet.

E-HCF could provide beforehand a list of the APs which MN could potentially connect to in accordance with the MN's location (in other words, based on the location of MN's current AP). Moreover, E-HCF defines the priority of the available APs according to the AP's charge, the network balance, the MN's trajectory, etc. If necessary, E-HCF could enquire its contiguous E-HCFs to offer their APs and relative information via the Int-E-HCFReq and Int-E-HCFRep messages [13].

E-HCF reserves a pool of IP address to form a list of available CoA for MN. Simultaneously, E-HCF generates and updates periodically a special IP address list - a list of IP address used by the nodes of its domains. By comparing these two lists, E-HCF could find the potential address collision in its domain. If the reserved IP address were used by a node, E-HCF would either withdraw this IP address from the fist list or demand the node to change its IP address. In this way, we could assure that once E-HCF distributes a reserved IP address to MN, MN could use this IP address directly in the new subnet.

We add six new messages into Mobile IPv6 protocol. E-HCF could use these messages to exchange both the intern (local) information with its ARs/MNs and the extern information with another E-HCF. These messages are the following: MN Request (MNReq), E-HCF Reply (E-HCFRep), Inter-E-HCF Request (Int-E-HCFReq), Inter-E-HCF Reply (Int-E-HCFRep), Connection Established Information (CEInf) and Handover Finished Confirmation (HFCon) messages. Due to the limited pages, the formats of these messages won't be given at here, please read our technique report [13].

3.2 E-HCF Procedure

In this sector, we give the detail handover procedure to the well understand of our proposition. As shown in Figure 2, E-HCF procedure is detailed as follows:

- MN moves in the network, once the threshold of received signal strength is overstepped, it sends a MNReq Message to E-HCF immediately to request the network information, such as the available APs, its BSSID, its channel, the prefix of the corresponding ARs, etc.
- E-HCF replies to MN with the E-HCFRep message. MN could obtain the needed information, and the prospective corresponding CoA from this message.
- Once MN receives the E-HCFRep message, MN begins to associate with the first AP of the list. If MN could establish the connection with the first AP, MN uses the proposed CoA to send the Binding Update to its HA, its CNs, and sends the CEInf message to E-HCF to notify its attachment. Otherwise, it does the same in order to connect to the better AP.
- To avoid packets loss, E-HCF commences buffering the traffic destined to MN's previous CoA when it receives the MNReq message. Once it receives the CEInf message, it sends the buffered packets to MN's new CoA. E-HCF sends HFCon message to MN when it couldn't receive the traffic destined to MN's previous CoA any more.

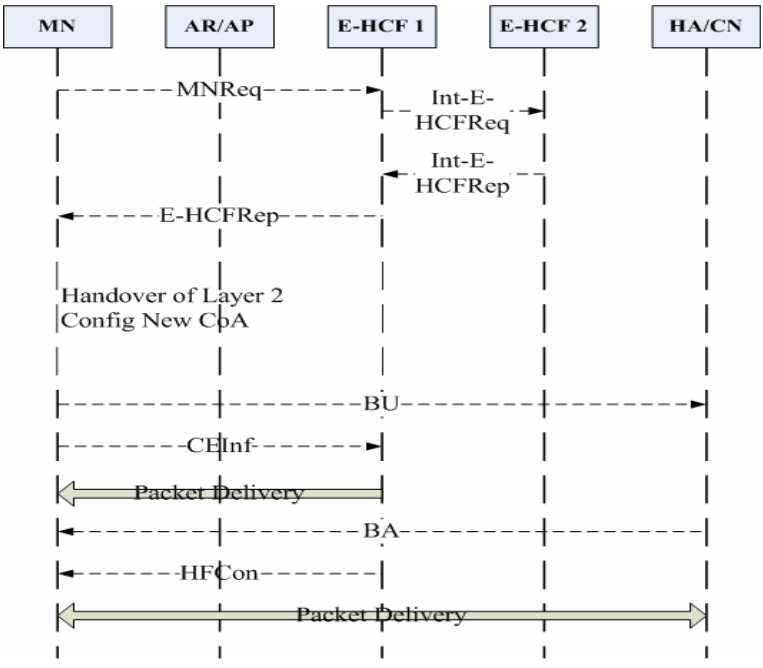


Fig. 2. E-HCF procedure

Recall that Int-E-HCFReq/Int-E-HCFRep messages are exchanged between the two E-HCFs for MN's inter-E-HCF handover. Each E-HCF maintains and updates its proper network information.

4 E-HCF Performance Estimation

The MIPv6 handover over WLAN consists of Link Layer handover and Network Layer handover. Link Layer handover includes Probe phase, Authentication phase, and Re-association phase. Network Layer handover includes Router Discovery phase, DAD phase and Binding Update phase. Displayed in Figure 3, the handover latency

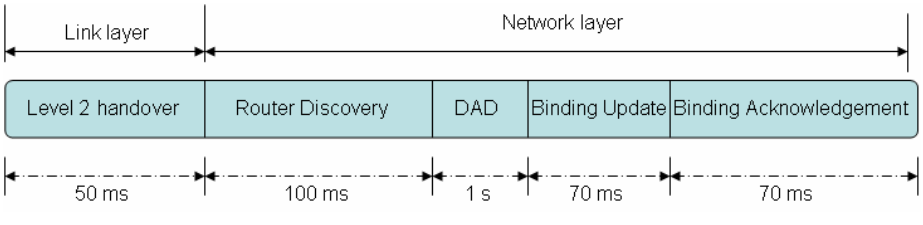


Fig. 3. Standard MIPv6 Latency

could be estimated to minimum 1290 ms [14] without including probe phase delay (between 350ms and 500ms). If we analyze each phase during handover process, we can observe that the probe phase and DAD phase delay costs 350ms and 1000ms respectively. In another word, probe phase and DAD phase are the key argument for the handover latency. In the following sections, we try to prove our proposition by mathematic analysis and compare E-HCF with MIPv6.

Before the detail latency analysis, we give the following notations:

L_{MIPv6}	Total handover latency with MIPv6
$L_{E\text{-}HCF}$	Total handover latency with E-HCF.
L_{Prob}	Latency that MN scans all neighboring AP to find the available AP
$L_{Authentication}$	Latency for authentication phase of Link Layer handover.
$L_{Association}$	Latency for association phase of Link Layer handover.
$L_{Router\ Discovery}$	Latency for Router Discovery phase of Network Layer handover
L_{DAD}	Latency for DAD phase of Network Layer handover
$L_{BU/BA}$	Latency for Binding Update phase of Network Layer handover
L_{MNReq}	Latency that MN sends a MNReq message to its E-HCF.
$L_{E\text{-}HCFRep}$	Latency that E-HCF sends E-HCFRep message to MN, including sometimes the delay for the Int-E-HCFReq message/Int-E-HCFRep message exchange.
L_{CEInf}	Latency that MN sends CEInf message to E-HCF
L_{HFCon}	Latency that E-HCF sends HFCon message to MN

In according to the procesus illustrated in above section.

The total MIPv6 handover latency L_{MIPv6} is:

$$L_{MIPv6} = L_{Prob} + L_{Authentication} + L_{Association} + L_{Router\ Discovery} + L_{DAD} + L_{BU/BA}$$

The total E-HCF handover latency L_{HCF} is:

$$L_{HCF} = L_{MNReq} + L_{E\text{-}HCFRep} + L_{Authentication} + L_{Association} + L_{BU/BA} + L_{CEInf} + L_{HFCon}$$

When we compare our E-HCF approach with MIPv6, we find that the key negative arguments of handover latency (such as L_{Prob}, $L_{Router\ Discovery}$, L_{DAD}) could be eliminated from our E-HCF proposition. Although the latency for messages exchange is introduced, latency for messages exchange would not be important.

In term of the packet loss, E-HCF commences buffering the traffic destined to MN's previous CoA when it receives the MNReq message. Once it receives the CEInf message, it sends the buffered packets to MN's new CoA. If we consider that the size of E-HCF buffer is sufficient for buffering all packets destined to MN's previous CoA, which arrived before E-HCF receives CEInf message and begins to sends the packets to MN's new CoA, the packet loss gets its value to zero. Moreover, according to our E-HCF approach, the duration between the time that E-HCF commences buffering the packets and the time that E-HCF receives CEInf message and begins to sends the packets to MN's new CoA is quite short.

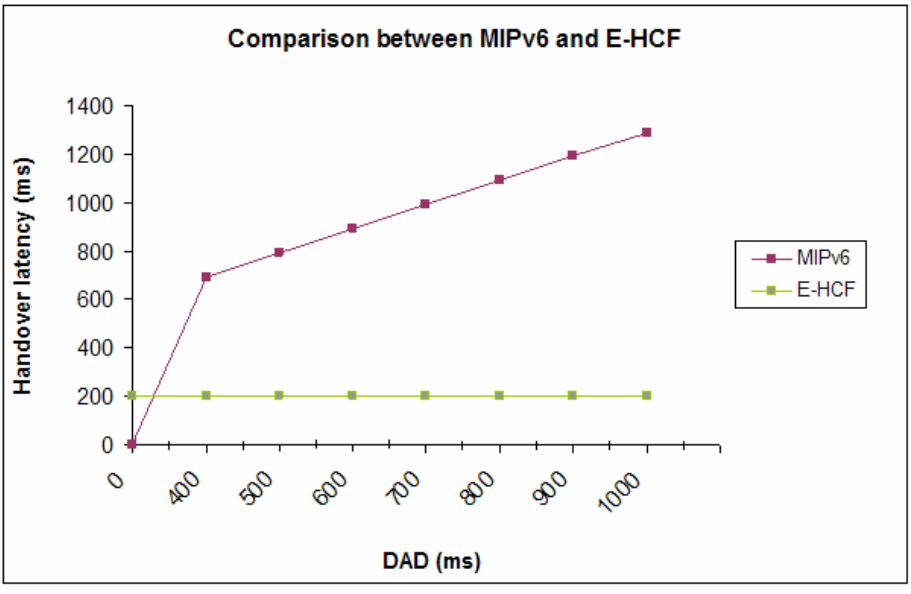


Fig. 4. E-HCF handover latency comparison in function of DAD

5 Conclusions

This paper proposes a new entity - E-HCF to improve the handover performance in Mobile IPv6 over WLAN. According to our analysis, both the handover delay and the packet loss could be significantly decreased. Moreover, E-HCF could enhance the AP's security, because AP needn't broadcast its SSID any more to permit MN to discover itself. Furthermore, E-HCF permits MN to connect to the authorized AP by adding the corresponding security parameters (such as WEP/WPA) into the encrypted E-HCFRep messages.

While how to better choose the available AP and define the AP's priority according to MN's mobility pattern is still a challenge for us. What's more, we need guarantee the QoS service for diverse MNs which have the different traffic type, the different service demand, and the different priority. Our current work focuses on resolving above problems and using the OPNET (Optimized Network Engineering Tool, a network simulator) [15] to prove our proposition. The wireless handover in MIPv6 will be studied in the future.

References

[1] C. Perkins, "IP Mobility support for IPv4", RFC 3220, IETF, January 2002.
[2] D.Johnson, C. Perkins, and J.Arkko, "Mobility Support in IPv6", RFC 3775, June 2004.
[3] B. O'Hara, A. Petrick, "IEEE 802.11 Handbook- A Designer's Companion", IEEE Press.
[4] H.Soliman, C.Castelluccia, K.Malki, and L.Bellier, "Hierarchical Mobile IPv6 mobility management (HMIPv6)", RFC 4140, August 2005.

[5] R. Koodli, Ed. “Fast Handovers for Mobile IPv6”, RFC 4068, July 2005.
[6] Bok-Deok Shin, Kyung-Jae Ha, “An Improvement of Handoff Latency by Virtual MAPs for HMIPv6”, in Proceedings of the IEEE CIT 2006 p.78
[7] Sangheon Pack, Minji Nam, and Yanghee Choi, "Design and Analysis of Optimal Multi-Level Hierarchical Mobile IPv6 Networks," Springer (Kluwer) Wireless Personal Communications (WIRE), Vol. 36, No. 2, January 2006.
[8] Mishra, M.Shin, and W.Arbaugh, “An Empirical Analysis of the IEEE 802.11 MAC layer Handoff Process”, ACM Computer Communications Review, vol. 33, no. 2, Apr. 2003.
[9] Ishwar Ramani, Stefan Savage, “ SyncScan: Practical Fast Handoff for 802.11 Infrastructure Networks”, Proceedings of the IEEE Infocom Conference, Miami, FL, March 2005
[10] “IEEE 802.11f: Recommended Practice for Multi-Vender Access Point Interoperability via an Inter-Access Point Protocol Access Distribution Systems Supporting IEEE 802.11 Operation”, IEEE Standard 802.11, Jan. 2003
[11] H.Chaouchi, P.Antunes, “Pre-handover signaling for QoS aware mobility management,” international journal of network management 14, pp.367-374, 2004;
[12] G. Z. Wei, A. Wei, K. Xu and H. Deng, “Handover Control Function Based Handover for Mobile IPv6,” In Proceedings of ICCS 2006, May 2006.
[13] Guozhi WEI, “E-HCF function and messages formats”, research report 2006-145, University of Paris XII –Val de Marne
[14] Wei Kuang Lai and Jung Chia Chiu, “Improving Handoff Performance in Wireless Overlay Networks by Switching Between Two-Layer IPv6 and One-Layer IPv6 Addressing,” IEEE Journal on Selected Areas in Communication vol. 23, No.11, November 2005.
[15] C. Zhu, O. W. W. Yang, J. Aweya, M. Oullette, and D. Y. Montuno. “A comparison of active queue management algorithms using the OPNET Modeler.” IEEE Communications Magazine, 40(6): 158-167, 2002.

HMIPv6 Applying User's Mobility Pattern in IP-Based Cellular Networks*

Teail Shin, Hyungmo Kang, and Youngsong Mun

School of Computing, Soongsil University
Sangdo-5dong, Dongjak-gu, Seoul, Korea
nullx@sunny.ssu.ac.kr, kangzzang@sunny.ssu.ac.kr,
mun@computing.ssu.ac.kr

Abstract. Hierarchical Mobile IPv6 (HMIPv6) was proposed by the Internet Engineering Task Force (IETF) for efficient mobility management of Mobile IPv6 nodes. In HMIPv6, when a mobile node moves access router in the different MAP domain, such a movement is called macro mobility handover. In this situation, the mobile node creates new RCoA and LCoA and performs registrations with the new MAP and HA. Especially, until the address registrations with an MAP and HA are completed, the mobile node cannot receive IP packet and it is hard to actualize seamless services. Therefore, we need to execute the macro mobility handover in advance to reduce the handover latency and packet loss. We propose a scheme that is able to improve the macro mobility handover efficiently in HMIPv6. To do this we will adjust the user's mobility patterns to edge access routers between different MAP domains in the handover situation. In this paper, we compare cost of the macro mobility handover of the proposed scheme with that of the original HMIPv6. As a result, we can obtain the improved macro mobility handover in HMIPv6.

1 Introduction

Mobility management is an essential technology because mobile service has to find out user's current location and deliver data correctly [1]. In this situation, the Mobile IP working group within Internet Engineering Task Force (IETF) proposed a mobile IPv6 (MIPv6) [2] protocol to support mobility management in IPv6 network. MIPv6 allows a mobile node (MN) to move while maintaining connections between the MN and correspondent nodes (CNs). To do this the MN sends Binding Update (BU) messages to its home agent (HA) and all CNs. But, handover involves the time of the new prefix discovery on the new subnet, the time for generating a new care-of address (CoA) and the time for registering the new address with the HA and the CNs. These are called the handover latency and can disrupt the real-time multimedia application service. Therefore, the extensions of MIPv6 have been proposed by IETF to offer

* This research was supported by the MIC(Ministry of Information and Communication), Korea, under the ITRC(Information Technology Research Center) support program supervised by the IITA(Institute of Information Technology Advancement) (IITA-2006-C1090-0603-0040).

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 514–521, 2007.
© Springer-Verlag Berlin Heidelberg 2007

seamless service in MIPv6 and a hierarchical MIPv6 (HMIPv6) [3] protocol manages the movement of the MN by using mobility anchor point (MAP).

In HMIPv6, macro handover is that a MN moves between access routers in the different MAP domain. In this situation, the MN creates new regional CoA (RCoA) and local CoA (LCoA) and performs registration with the new MAP and HA. Especially, the MN cannot receive IP packet and it is hard to actualize seamless services, until the address registration with an MAP and HA is completed. Therefore, we need to execute the macro mobility handover in advance to reduce the handover latency and packet loss. We propose a scheme that is able to improve the macro mobility handover efficiently in HMIPv6. To do this we will adjust the user's mobility patterns to edge access routers between different MAP domains during handover. We also use fast handover technique [4], [5]. In this paper, we compare cost of the macro mobility handover of the proposed scheme with that of the original HMIPv6. Therefore, we can improve macro mobility handover in HMIPv6.

2 Related Works

2.1 Hierarchical Mobile IPv6 (HMIPv6)

In HMIPv6, the MAP can be located at any level in a hierarchical network of routers and is intended to limit the amount of MIPv6 signaling outside the local domain. In HMIPv6, an MN can perform two types of handover. Firstly if the MN moves between access routers belonging to the same MAP domain, it is called a micro mobility handover. Secondly when the MN moves between access routers belonging to the different MAP domains, it is called a macro mobility handover and we have to analyze this situation.

When an MN enters into a new MAP domain, it receives a router advertisement (RA) message from a new access router. The MN generates a new LCoA and RCoA based on the access router's prefix contained in the RA message and the prefix of the MAP contained in the MAP option, respectively. After generation of the new addresses, the MN sends a local binding update (LBU) message to the MAP. The LBU includes new RCoA and LCoA. The MAP performs the duplicate address detection (DAD) process to verify that the MN's new RCoA is unique in the MAP domain. Then, the MAP stores the RCoA and LCoA of the MN in its binding cache. After that, the MAP intercepts the packets transmitted to the RCoA of the MN through a proxy neighbor advertisement message. The MAP sends the packets to the MN's LCoA by tunneling. If the new addresses of the MN are stored in MAP, the MN sends a binding update (BU) message to its HA to inform its new RCoA as CoA. After the registration with the HA, the MN can perform the registration with the CN.

If the MN performs macro mobility handover, it cannot receive IP packet until the address registrations with the MAP and the HA are completed. Namely, macro handover is hard to actualize seamless real-time services.

3 Proposed Scheme

3.1 User's Mobility Pattern and Proposed Scheme

In spite of development of transport systems, people spend a lot of time in limited zone such as home, school and one's place of work and have similar life pattern. Also, they always come and go between their routine places and the movement paths of each people are scarcely changed. According to this fact we can analyze mobility patterns of people and anticipate future movement spot. Especially, if Mobile IPv6 nodes use this technique, MN can know next access router in advance. As a result, when handover happens, Mobile IPv6 nodes reduce latency and protect packet loss considerably.

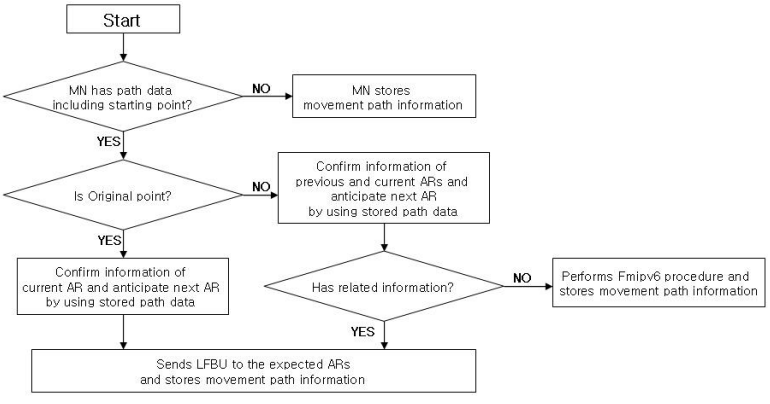


Fig. 1. Access router prediction algorithm based on user pattern

3.2 Access Router Prediction Algorithm Based on User Pattern

In this paper, we propose access router prediction algorithm based on user pattern and we assume that people register original points such as home and school or one's place of work to their mobile equipment and MNs store their path between source address and destination address every movement. By stored data, an MN keeps various user mobile paths and pattern. When an MN accesses some router, it checks whether the MN had current router information or the router is an original point. If the router is an original point and the MN has the router information, the MN gets the next router through the router link information. Although the router is not original router, if the MN has the router information, then, the MN expects the next router by comparison of previous router and current router information. Else the MN stores the movement paths and performs the basic FMIPv6 procedure. Stored data must not be duplicated. Fig. 1 shows access router prediction algorithm based on user pattern.

3.3 Procedure of the Proposed Scheme

In procedure of the proposed scheme, we assume that mobile users already came back from their own routine mobility paths. While MNs move access routers, each MN

stores mobility path data such as access router and map information. The procedure of proposed scheme is divided by two steps of LCoA and RCoA establishment, as shown in Fig. 2. When the handover will happen, the MN operates access router prediction algorithm based on user pattern and anticipate the next router. If an MN gets next router information, the MN generates an RCoA and an LCoA based on the prefixes of the MAP2 domain and the AR3. At this time, the MN sends an LCoA FBU message including the new LCoA to the AR2. The AR2 sends an LCoA HI message to the AR3 that is in the MAP2 domain for establishing new LCoA of the mobile node. The AR3 performs the DAD to verify the uniqueness of the MN's new LCoA in the AR3 subnet. If the AR3 completes the DAD about the address, it sends an LCoA HACK message to the AR2 in the MAP1 domain. By an LCoA FBACK message the MN gets information that the new LCoA of the MN is available and the first step of the fast macro mobility handover is ready.

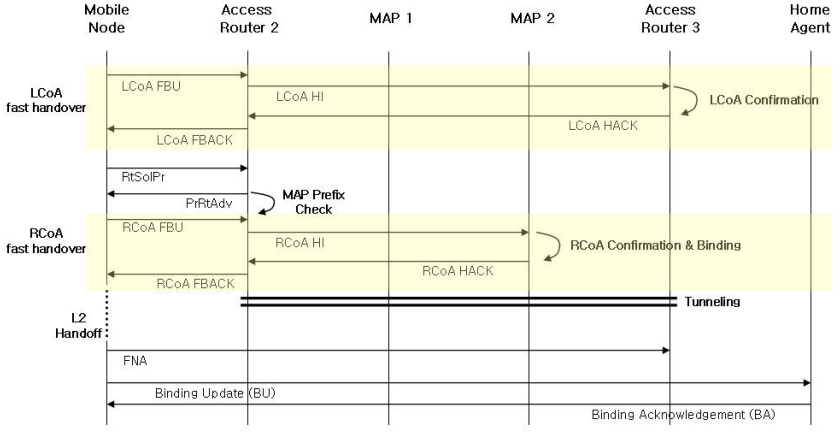


Fig. 2. The procedure of proposed method (LCoA fast handover and RCoA fast handover)

By AP signal strength, the MN will recognize that it is close to another AR and MAP domain and exchange an RtsolPr and a PrRtAdv messages. The MN compares prefix of new RCoA with MAP prefix information of PrRtAdv message and confirms the next MAP and access route. And then, the MN performs the second step of fast macro mobility handover process. The MN sends an RCoA FBU message including the new RCoA to the AR2. The AR2 sends an RCoA HI message to the MAP2 for both establishing new RCoA of the mobile node and binding RCoA and LCoA. The MAP2 performs the DAD to verify that the mobile node's new RCoA is unique. If the MAP2 completes the DAD about the address and binding, it sends an RCoA HACK message to the AR2 in the MAP1 domain and the AR2 transfers an RCoA FBACK message to the MN. Especially, in order to reduce processing load of MAP, after an MN makes certain its real movement to the MAP2, the MN operates handover with new RCoA of MAP2. During the L2 handover, all packets sent by the CN are stored by the AR3 through the tunnel along the MAP1 and MAP2 and AR3. The AR3 buffers the packets until the L2 handover of the mobile node ends. After the MN

completes the L2 handoff, it sends a FNA message to the AR3 to announce that the mobile node is on-link in AR3 subnet. The AR3 received the FNA message stops the proxy function for the mobile node's LCoA and transmits all buffered packets to the mobile node. Therefore, the mobile node can receive all packets transferred during the macro mobility handover. After this handover procedure, the mobile nodes can perform the location update with its HA and CN, too. But, if stored data and current and previous router information has not same value, an MN stores movement paths and operates basic HMIPv6 and FMIPvc6 process.

4 Performance Evaluation

4.1 System Modeling

We use the evaluation model in [6], [7] to evaluate performance of the proposed scheme. Fig. 3 shows the distance of entities and values. We define the variables for this paper in Table 1 and assume that message treatment cost is identical at each hop.

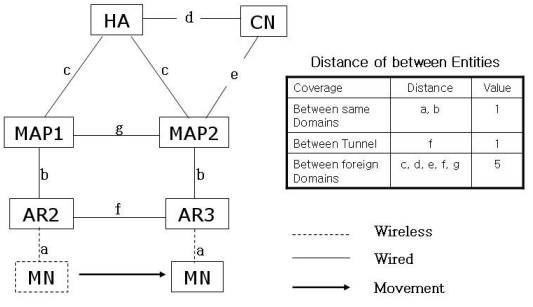


Fig. 3. System method for performance analysis

Table 1. Defined variables for computation of performance evaluation Reference

λ	the rate that correspond node (CN) transmits data packets to a mobile node (MN)
μ	the rate that MN moves from one subnet to another
p	packet to Mobility Ratio (PMR) : the mean number of packets that a MN received from a CN
l_c	the average control packet length (=200byte)
l_d	the average data packet length (=1024byte)
l	$l = l_d / l_c$
r	cost of processing control packets (all hosts)

4.2 Cost Analysis

In this paper, in order to analyze the performance of mobile networks, we consider the total cost that consists of signaling cost and packet delivery cost.

$$C_{total} = C_{signal} + C_{packet} \tag{1}$$

4.2.1 Signaling Cost

Contrary to the basic HMIPv6, the proposed scheme reduces packet loss by using fast macro handover. But, we should consider the additional signaling costs such as detection of DB and accomplishment of two pre-handover works. In the proposed scheme, the signaling cost for macro handover can be expressed as follows:

$$C_{signal} = C_{fast-l} + C_{fast-r} + C_{HA} \tag{2}$$

Signaling cost consists of the cost of LCoA Fast Handover (C_{fast-l}), RCoA Fast Handover (C_{fast-r}), and registration with HA (C_{HA}). Each signaling costs are as defined in the following expressions.

$$C_{fast-l} = 2(a + 2b + g) + 9r \tag{3}$$

$$C_{fast-r} = 5a + 2(b + g) + 16r \tag{4}$$

$$C_{HA} = 2(a + b + c) + 13r \tag{5}$$

In this paper we assume that the processing cost is r per message treatment, registration or confirmation and distance is directly proportional to process time. Each processing cost is increased or decreased in accordance with message process method.

4.3 Packet Transmission Cost

Packet transmission cost is C_{packet} that consists of packet forwarding cost (C_{fwd}) and the packet loss cost (C_{loss}).

$$C_{packet} = C_{fwd} + C_{loss} = (\lambda \times t_{delay} \times C_{dt}) + C_{loss} \tag{6}$$

where t_{delay} is waiting time that MN can't receive any message and it includes L2 handoff latency (t_{L2}), the movement time of hops (t_x), and packet processing time on host (t_r). C_{dt} is a single data packet cost delivered from CN to MN and it is calculated as $C_{dt} = l \times (g + f + h + d) + 3r$. Since the packet transmission cost is greater than the signaling cost, reducing the delay of packet transmission cost decreases the total cost. Therefore, a main focus of the proposed method is a reduction of a packet forwarding latency. t_{delay} can be calculated as follows:

$$t_{delay} = 2t_a + 2t_r + t_{L2} \tag{7}$$

The proposed method decreases configuration latency because it performs DAD and binding procedure in advance. In basic HMIPv6, t_{delay} is obtained from the following equation.

$$t_{delay-hmip} = 2(2t_a + t_b) + 8t_r + t_{L2} + 2t_{DAD}, \tag{8}$$

where t_{DAD} is DAD time which spends 10 times as large as t_r. The proposed method reduces the delay regardless of predictive or reactive mode. Especially, the suggested method in reactive mode gets high performance improvement in comparison with other methods. Packet loss cost is delay cost of real-time service caused by latency. In this paper, we define packet loss cost as a partial packet transmission cost from completion of fast handover process to home registration. It is because buffered packets are not seamless real-time service.

$$C_{loss} = \eta \times (\lambda \times (t_{delay} - t_{delay-fast}) \times C_{dt}) \tag{9}$$

In Eq. (9), η is weight of a packet loss cost. We assume that η is 1 , because real-time service does not re-transmit packets.

4.4 Numerical Results

We verify the improvement by using the rate of cost. When a mobile node moves to the new domain, the rate shows below. The rate is the total cost of proposed method about the total cost of the comparison method. In this paper, we assume that the total cost includes loss cost because the service data for the waiting time is not retransmitted in real-time service. We refer to the method in [7] and use Eq. (10) and Eq. (11).

$$t_{RT-wire}(h,k) = 3.63k + 3.21(h-1) \text{ (msec)} \tag{10}$$

$$t_{RT-wireless}(k) = 17.1k \text{ (msec)} \tag{11}$$

The k is packet length (kbyte) and h is the number of hops. In [8], [9], it provides the latency (t_{L2} =84 msec and t_r =0.5 msec) for wireless section. And we use $\lambda = p \times \mu$ in Eq. (12) to apply mobility. Eq. (12) is the rate in predictive mode of Fast Handover.

$$\lim_{p\to\infty} \frac{C_{total}}{C_{total-hmip}} = \lim_{p\to\infty} \frac{C_{fast-l} + C_{fast-r} + C_{HA} + (\lambda \times t_{delay} \times C_{dt}) + C_{loss}}{C_{signal-hmip} + C_{HA} + (\lambda \times t_{delay-hmip} \times C_{dt}) + C_{loss}} \tag{12}$$

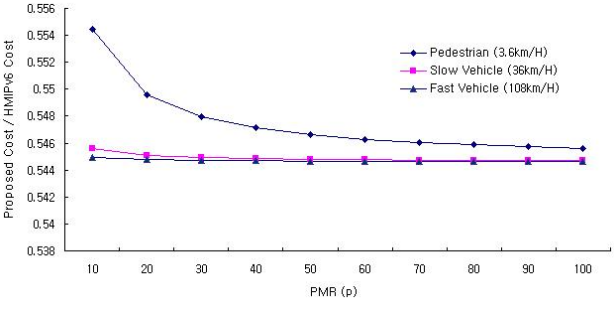


Fig. 4. The variation of the cost ratio of the proposed method per the basic HMIPv6 PMR

Through Eq. (12), we can get the results illustrated in Fig. 4. In the figure, the abscissa and ordinate show PMR value and the rate of cost decrease. The result rate converges at a spot rather than 100 point of PMR. The rate of Eq. (12) converges to 0.546 in pedestrian and vehicle (3.6km/H: $\mu = 0.01$ ~ 108km/H: $\mu = 0.5$(m/msec)).

The proposed method gets approximately 45% cost profit in comparison with the basic HMIPv6. By this result we know that the proposed method reduces handover latency and supports real-time services.

5 Conclusions

During macro mobility handover, the MN creates new RCoA and LCoA and performs registration with the new MAP and HA. Especially, until the address registration with an MAP and HA completed, the MN cannot receive IP packet and it is hard to actualize seamless services. In order to solve handover latency problem, we need to analyze user mobile patterns. People spend a lot of time in limited zone such as home, school and one's place of work and have similar life pattern. Also, they always come and go between their routine places and the movement paths of each people are scarcely changed. According to this fact we can analyze mobility patterns of people and anticipate future movement spot. If Mobile IPv6 nodes use this technique, an MN can know next access router in advance. As a result, when handover happens, Mobile IPv6 nodes reduce latency and protect packet loss considerably. By performance evaluation, we compare cost of the macro mobility handover of the proposed scheme with the original HMIPv6 and obtain the improved macro mobility handover in HMIPv6.

References

1. I. F. Akyildiz, et al., "Mobility Management in Next-Generation Wireless Systems." Proceedings of the IEEE, August 1999.
2. Charles E. Perkins and David B. Johnson, "Mobility Support in IPv6." RFC 3775, IETF
3. Soliman, H., Catelluccia, C., Malki, K. and Bellier, L., "Hierarchical Mobile IPv6 Mobility Management." Work in progress, 2004.
4. Koodli, R., et al., "Fast Handovers for Mobile IPv6." RFC 4068, IETF, 2005.
5. Narten, T., Nordmark, E. "Neighbor Discovery for IP version 6 (IPv6)." RFC2461, IETF
6. Sangheon Pack and Yanghee Choi, "Performance Analysis of Fast Handover in Mobile IPv6 Networks," in proc. IFIP PWC 2003, Venice, Italy, September, 2003.
7. R. Jain, T.Raleigh and C, Graff, M. Bereschinsky, "Mobile Internet Access and Qos Guarantees using Mobile IP and RSVP with Local Register," in Proc. ICC'98 Conf,. pp. 1690-1695, Atlanta.
8. Thomas, R., H. Gilbert and G. Mazzioto, "Influence of the mobile station on the performance of a radio mobile cellular network," Proc. 3rd Nordic Sem., paper 9.4, Copenhagen, Denmark, Sep. 1988.
9. Jon-Olov Vatn, "An experimental study of IEEE 802.11b handover performance and its effect on voice traffic", SE Telecommunication Systems Laboratory Department of Microelectronics and Information Technology (IMIT), July 2003.

Performance Analysis and Comparison of the MIPv6 and mSCTP Based Vertical Handoff*

Shi Yan[1], Chen Shanzhi[2], Ai Ming[1], and Hu Bo[1]

[1] State Key Laboratory of Networking and Switching,
Beijing University of Posts and Telecommunications, 100876
[2] China Academy of Telecommunications Technology, Beijing, 100083
catshiy@gmail.com, {chenshanzhi,aimingam}@yahoo.com.cn,
hubo@bupt.edu.cn

Abstract. Vertical handoff is the key technology supporting session mobility in the future heterogeneous network environments. Based on the asymmetry feature of vertical handoff, this paper analyzes the handoff procedure of MIPv6 and mSCTP in forced and unforced handoff scenarios. Qualitative and quantitative analysis and comparison of the handoff performance including handoff delay, handoff packet loss and signaling overhead are given accordingly. Besides, the main factors influencing vertical handoff performance are pointed out and possible performance improvement schemes are discussed.

Keywords: Vertical handoff, MIPv6, mSCTP, Handoff performance.

1 Introduction

Handoff is one of the key technologies in mobility management. To support session mobility in the future ubiquitous and heterogeneous network environments, the research focus turns from horizontal handoff between homogeneous access technologies to vertical handoff between heterogeneous access technologies.

According to the handoff direction, vertical handoff can be classified into upward and downward handoff. Upward handoff is from the network with small coverage to the network with large coverage, e.g. the handoff from WLAN to GPRS. The reverse handoff is called downward. Different handoff directions bring notable asymmetry in handoff scenarios, procedures and performance. There are two typical scenarios – forced and unforced handoff. Forced handoffs are usually triggered by lower layer events incurring the change of interface availability. Since only one interface is available, the handoff is obligatory to avoid communication interruption. Unforced handoffs are triggered by users actively according to user policies, preferences or perceived QoS. And multiple interfaces are available simultaneously. Downward handoff must be unforced, while upward handoff may be forced or unforced.

In order to ensure the independency of access technologies, vertical handoff is implemented at network layer or above. MIP (Mobile IP)[1][2] and mSCTP (mobile

* This work is supported by the National High-Technology (863) Program of China under Grant No. 2006AA01Z229, 2005AA121630 and 2003AA121530.

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 522–529, 2007.
© Springer-Verlag Berlin Heidelberg 2007

Stream Control Transport Protocol)[3][4] are two typical vertical handoff protocols. MIP operates at network layer and provides server-based handoff, while mSCTP operates at transport layer and provides end-to-end based handoff. [5-10] propose the vertical handoff schemes with simple performance analysis. Some handoff performance comparison of MIPv4 and mSCTP is made in [11]. But in all these works, the asymmetry feature, forced handoff scenario and error-prone feature of wireless links are not considered adequately. This paper analyzes the vertical handoff performance of MIPv6 and mSCTP. Qualitative and quantitative results are given. The asymmetry feature of vertical handoff and lossy feature of wireless links are considered. The rest of this paper is organized as follows: section 2 gives the qualitative analysis based on the unified analysis model. Section 3 gives quantitative results. Section 4 concludes this paper, the main performance influencing factors are pointed out and possible performance improvement schemes are discussed.

2 Vertical Handoff Performance Analysis

Generally, vertical handoff performance can be evaluated by handoff delay, handoff packet loss and signaling overhead. Handoff delay refers to the connection interrupted interval. Packet loss is often evaluated by the interval incurring loss. Signaling overhead is defined as the traffic caused by the signaling messages exchange during the handoff. In wireline networks, it is denoted by the product of signaling message packet size and transmission distance (hops). Considering the limited bandwidth on wireless links, the product should be multiplied by a factor α , where $\alpha > 1$.

2.1 Unified Analysis Model

The handoff performance analysis is based on different subprocedure combinations and sequences. The main subprocedures include: (1) L2 handoff: the link layer connection change procedure. (2) System discovery and movement detection: MN recognizes the change of network connectivity. (3) Access authentication: the necessary procedure for handoff across heterogeneous access networks belonging to different domains or operators. (4) Address configuration: the network assigns an IP address to MN. (5) Handoff protocol signaling subprocedure: the signaling exchange for implementing the handoff, e.g. the location registration procedure in MIPv6, the setting primary address procedure in mSCTP. (6) Assistant subprocedure: the handoff assistant operations, e.g. the unavailable IP address deletion in mSCTP.

2.2 Vertical Handoff Procedure and Qualitative Performance Analysis

A. Vertical Handoff Based on MIPv6

The forced vertical handoff procedure based on MIPv6 is illustrated in Fig.1, which includes several serial subprocedures: L2 handoff, movement detection, access authentication, address configuration and MIPv6 location registration. The introduced delay are denoted by T_{L2HO} , $T_{MoveDet}$, T_{Auth} , $T_{AddrCon}$ and T_{Regis} . We will analyze $T_{MoveDet}$ and T_{Regis} , which are protocol-related. The others are determined by specific access technologies.

The movement detection adopts L3 detection [2]. MN continuously listens to the RA (Router Advertisement) messages multicasted by the routers periodically with interval Δ_{RA}. If MN has not received any RA from the default router in Δ_{RA}, it can determine one RA miss. After several RA misses (here assumes three for a reasonable tradeoff between efficiency and robustness), MN can confirm an IP connectivity change. [12] defines the interval as 3-4s at least and 1350-1800s at most. In order to support mobility efficiently, [2] decreases it to 30-70ms. Hence, the movement detection duration can be given by $T_{MoveDet} = 3 \bullet \Delta_{RA}$.

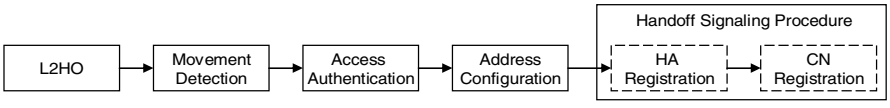


Fig. 1. Forced Vertical Handoff Procedure Based on MIPv6

The location registration procedure includes HA registration and CN registration. [2] dictates that MN is permitted to send BU (Binding Update) to CN only after it receives BA (Binding Acknowledgement) from HA, that is, HA registration and CN registration are serial procedures.

In the unforced vertical handoff based on MIPv6, MN has multiple available networks simultaneously. As illustrated in Fig. 2, system discovery, authentication and address configuration can be completed when MN continues communication through the former network interface. Only after the handoff decision, the MIPv6 location registration procedure is triggered for link switch.

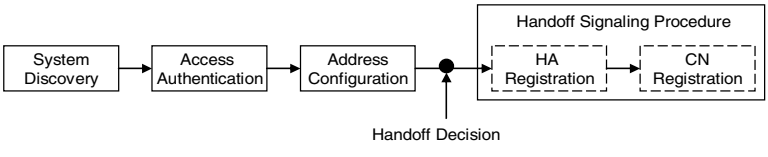


Fig. 2. Unforced Vertical Handoff Based on MIPv6

The detailed performance for MIPv6-based forced and unforced handoff is summarized in Table 1. The packet loss is denoted by the duration incurring loss.

B. Vertical Handoff Based on mSCTP

mSCTP supports vertical handoff because of its multi-homing feature and DAR (Dynamic Address Reconfiguration) extension [13]. With the movement of MN, mSCTP-based handoff includes three DAR operations: adding new acquired IP address, changing the primary IP address and deleting the unavailable IP address. All these DAR operations are implemented through a couple of ASCONF and ASCONF_ACK messages carrying different parameters. Fig. 3 illustrates the procedures of the downward unforced handoff and upward forced handoff.

The forced vertical handoff based on mSCTP includes three serial subprocedures: movement detection, setting primary IP address and deleting unavailable IP address.

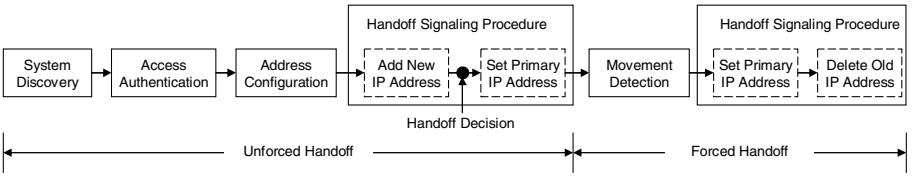


Fig. 3. Unforced and Forced Handoff Based on mSCTP

Because of the multi-homing feature, the new primary IP address is the existing and authenticated address in the association. Hence, authentication and address configuration are not necessary.

mSCTP uses the path failure detection function [3] for movement detection. An error counter is maintained for the primary IP address. It will be incremented each time the retransmission timer T3-rtx expires. When the error counter is incremented successively and exceeds the threshold PMR (Path Maximum Retransmissions), this address is regarded as unreachable and the movement is detected. The initial value of T3-rtx is RTO^{T3}_{mSCTP}=1s. Hence, the movement detection duration is $T_{MoveDet} = \sum_{i=0}^{PMR} 2^i RTO^{T3}_{mSCTP}$.

The key signaling procedure for handoff based on mSCTP is changing the primary IP address. The old IP deletion belongs to the assistant procedure. It incurs signaling overhead but it should not be included in the handoff delay.

The unforced handoff includes several subprocedures: system discovery, authentication, address configuration, adding IP address, setting primary IP address. The former ones are completed when MN continues communication through the old interface. Only after handoff decision, the primary address is changed for path switch.

The detailed performance of mSCTP is summarized in Table 1. It is notable that the packet loss in unforced handoff is 0 because mSCTP supports soft handoff.

Table 1. Vertical Handoff Performance Comparison for MIPv6 and mSCTP

Scenarios		MIPv6	mSCTP
forced	Handoff Delay	$D^{forced}_{MIPv6} = T_{L2HO} + 3 * \Delta_{RA} + T_{Auth} + T_{AddrCon} + T_{HA_Regis} + T_{CN_Regis}$	$D^{forced}_{mSCTP} = \sum_{i=0}^{PMR} 2^i RTO^{T3}_{mSCTP} + T_{SetPrimAddr}$
	Packet Loss	$LD^{forced}_{MIPv6} = D^{forced}_{MIPv6}$	$LD^{forced}_{mSCTP} = D^{forced}_{mSCTP}$
	Signaling Overhead	$C^{forced}_{MIPv6} = (L_{BU} + L_{BA}) \bullet (\alpha + d_{MN-HA} - 1) + (L_{BU} + L_{BA}) \bullet (\alpha + d_{MN-CN} - 1)$	$C^{forced}_{mSCTP} = 2 \bullet (L_{ASCONF} + L_{ASCONF-ACK}) \bullet (\alpha + d_{MN-CN} - 1)$
unforced	Handoff Delay	$D^{unforced}_{MIPv6} = T_{HA_Regis} + T_{CN_Regis}$	$D^{unforced}_{mSCTP} = T_{SetPrimAddr}$
	Packet Loss	$LD^{unforced}_{MIPv6} = D^{unforced}_{MIPv6}$	$L^{unforced}_{mSCTP} = 0$
	Signaling Overhead	$C^{unforced}_{MIPv6} = (L_{BU} + L_{BA}) \bullet (\alpha + d_{MN-HA} - 1) + (L_{BU} + L_{BA}) \bullet (\alpha + d_{MN-CN} - 1)$	$C^{forced}_{mSCTP} = 2 \bullet (L_{ASCONF} + L_{ASCONF-ACK}) \bullet (\alpha + d_{MN-CN} - 1)$

2.3 Analysis of the Handoff Signaling Delay

Handoff signaling procedure is unaviodable in any scenario. In this part, we will analyze the signaling delay (T_{Regis} in MIPv6 and $T_{SetPrimAddr}$ in mSCTP) in detail. Considering the error-prone feature of wireless links, the delay incurred by the error

recovery mechanism should be taken into account. Refering to and extending the analysis method in [14], the signaling delay analysis model is shown in Fig. 4. Here the delay consists of five parts: T_{MN}, T_{SAP} and $T_{CN}(T_{HA})$ denote the average processing delay at MN, SAP (Service Access Point, e.g. BS or AP), CN or HA, $T_{wireless}$ and $T_{wireline}$ denote the average transmission delay over wireless and wireline links.

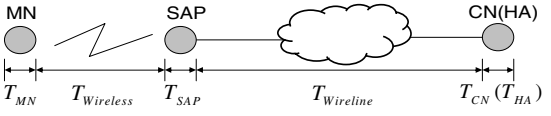


Fig. 4. The Handoff Signaling Delay Analysis Model

Assuming an M/M/1 queuing model for MN, SAP, HA and CN, according to the queuing theory, the average processing delay at each entity is: $T_{Entity} = 1/(\mu_{Entity} - \lambda_{Entity})$, where λ and μ denote the signaling message arriving rate and processing rate. Assuming the distance between SAP and HA or CN is w (hops) and the average transmission delay per hop is $\overline{t}$, we have $\overline{T}_{wireline} = w \bullet \overline{t}$. The average transmission delay over wireless links will be discussed in the following parts.

A. Average Transmission Delay over Wireless Link in MIPv6

[2] defines timeout retransmission mechanism to ensure the reliable transmission of BU and BA messages. The initial value and the threshold of the retransmission timer are 1s and 32s [2], i.e., the max retransmission is $N_{MIPv6} = 5$.

Assuming the FER (Frame Error Rate) at link layer is p, the bandwidth of the wireless link is B bps and the inter-frame time is τ s, for the IP packet with size L, the frame numbers contained in it is $k = L/(B \bullet \tau / 8)$.

For the HA registration procedure, the transmission error of either BU or BA will incur retransmission. Success registration implies both BU and BA are transmitted successfully, the probability of which is $p_S = (1-p)^{k_{BU}+k_{BA}}$. Therefore, with the maximum retransmission number N_{MIPv6}, the success registration probability is given by: $P = p_s + (1-p_s) \bullet p_s + (1-p_s)^2 \bullet p_s + \cdots + (1-p_s)^{N_{MIPv6}-1} \bullet p_s$. Correspondingly, the average transmission delay over wireless link caused by HA registration is given by:

$$\begin{aligned} \overline{T}^{MIPv6}_{wireless_HA_Reg} &= p_s \bullet T^{tran}_{MIPv6} + (1-p_s)\, p_s \bullet (T^{tran}_{MIPv6} + RTO_{MIPv6}) + \cdots \\ &\quad + (1-p_s)^{N_{MIPv6}-1}\, p_s \bullet (T^{tran}_{MIPv6} + (2^0 + 2^1 + \cdots + 2^{N_{MIPv6}-2}) RTO_{MIPv6}) \end{aligned} \tag{1}$$

where $T^{tran}_{MIPv6} = (D + (k_{BU} - 1)\tau) + (D + (k_{BA} - 1)\tau)$ is the transmission delay of BU and BA over the wireless link and D is the end-to-end frame propagation delay.

Obviously, for CN registration, we have the similar result.

B. Average Transmission Delay Over Wireless Link in mSCTP

mSCTP forms packets from the ASCONF and ASCONF-ACK control chunks and here we do not consider the bundling. mSCTP also uses retransmission mechanism for reliable transmission [13]. The timer is T4-rtx with initial value 1s. Considering

the multi-homing feature, the retransmission is implemented on the secondary path if it exists. In addition, an individual retransmission timer is maintained for each path.

For the forced handoff, the secondary path does not exist and the retransmission will be completed over the primary path. The maximum retransmission times should be PMR=5 [3]. Therefore, assuming the same FER for primary and secondary paths, we have similar conclusion to MIPv6.

For the unforced handoff, there exists the secondary path for retransmission and the maximum retransmission times should be AMR (Association Maximum Retransmission) =10 [3]. Considering the individual retransmission timer for each path, the average wireless transmission delay for unforced handoff can be given by:

$$\begin{aligned}\overline{T}_{wireless}^{mSCTP} &= p_s \bullet T_{mSCTP}^{tran} + (1-p_s)p_s \bullet (T_{mSCTP}^{tran} + RTO_{mSCTP}^{T4}) + (1-p_s)^2 p_s \bullet (T_{mSCTP}^{tran} + RTO_{mSCTP}^{T4} + RTO_{mSCTP}^{T4}) \\ &+ (1-p_s)^3 p_s \bullet (T_{mSCTP}^{tran} + RTO_{mSCTP}^{T4} + RTO_{mSCTP}^{T4} + 2RTO_{mSCTP}^{T4}) + \cdots \end{aligned} \tag{2}$$

3 Numerical Results

Based on the analysis in section 2 and referring to the typical parameters defined in the protocol recommendations [2][3][4][13] as well as related researches [6][15][17], the numerical results are given in the following part.

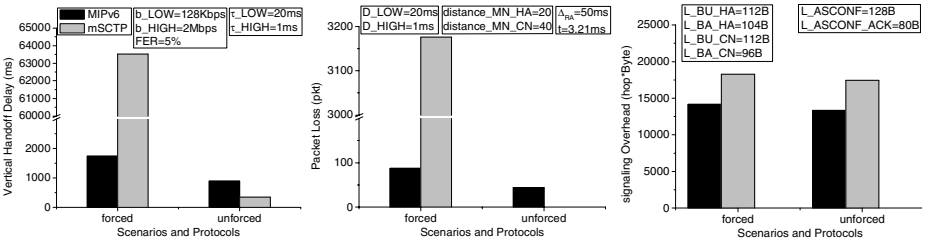


Fig. 5. Vertical Handoff Performance Comparison of MIPv6 and mSCTP in Upward Forced and Downward Unforced Handoff Scenarios

3.1 Vertical Handoff Performances Comparison

Fig. 5 shows the performance of MIPv6 and mSCTP in forced and unforced handoffs.

(1) For the same protocol, the handoff delay and packet loss in unforced handoff is better than forced handoff obviously. This is due to two reasons. Firstly, movement detection, authentication and address configuration contribute to the delay in forced handoff, but in unforced handoff they are completed before handoff is triggered and do not incur handoff delay. Secondly, signaling messages are transmitted over the higher performance link in downward unforced handoff

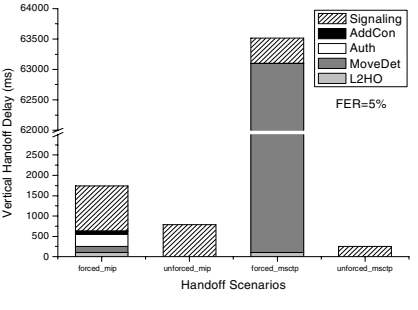


Fig. 6. Components of Vertical Handoff Delay

whereas over the lower performance link in upward forced handoff. This causes the difference in the signaling delay. Fig. 6 illustrates the components of the handoff delay in each scenario. We can see the effects of the above two reasons apparently. (2)Although authentication and address configuration are not necessary in mSCTP-based forced handoff, its handoff delay and packet loss is rather worse than MIPv6. This is due to the long duration for movement detection. mSCTP is designed for wire-line networks initially and many parameter values are not suitable for wireless and mobile networks. E.g., the initial values of the retransmission timers are defined as 1s, which incurs large delay for movement detection. Reversely, MIPv6 decreases the interval between two successive RAs for better mobility supporting. This reduces the movement detection delay efficiently. (3) In the unforced handoff, MIPv6 has larger handoff delay. This is because MIPv6 requires the serial HA and CN registration. (4) As for handoff packet loss, it is proportional to the handoff delay in most scenarios. In mSCTP-based unforced handoff, the loss is 0 because mSCTP supports soft handoff.

3.2 Handoff Signaling Delay Comparison

Handoff signaling delay is the unavoidable component in any scenario. In unforced handoff, it is the major part of handoff delay. The signaling delay with the variation of FER is shown in Fig. 7, and Fig. 8 illustrates the detailed composition.

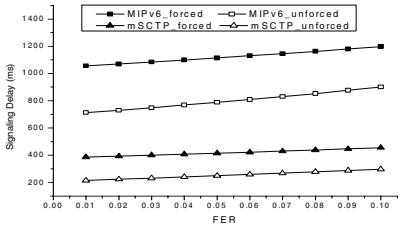


Fig. 7. Handoff Signaling Delay vs. FER

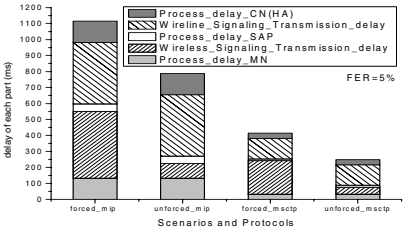


Fig. 8. Components of Signaling Delay

For both MIPv6 and mSCTP, the signaling delay of upward forced handoff is obviously lager than that of downward unforced handoff. We can see in Fig.8, the main difference exists in the wireless part transmission delay. This is because the packet loss rate at upper layer and the end-to-end frame propagation delay of lower bandwidth link are all larger than higher bandwidth link. In addition, the signaling delay of MIPv6 is lager than mSCTP due to the serial HA and CN registrations.

4 Conclusions

Performance Analysis and comparison of the MIPv6 and mSCTP based vertical handoff in different scenarios are given in this paper, considering the asymmetry in handoff directions, scenarios, procedures. Numerical results show that the handoff performances, especially the forced handoff performance based on standard protocols are not satisfying. For the more and more delay-sensitive realtime services with maximum acceptable interruption about 200ms, such a performance will be unacceptable. Therefore, it is stringent to improve the handoff performance.

The main reasons include: (1) in forced handoff, the serial procedures such as movement detection, authentication and address configuration incur considerable delay; (2) the low bandwidth and error-prone feature of wireless link introduce high message transmission delay for error recovery; (3) some parameters defined in the standard protocols can not support handoff in wireless mobile networks efficiently.

The possible methods to improve the vertical handoff performance include: (1) In forced handoff, cross-layer interaction should be introduced to provide lower layers information to upper layer handoff protocols. This is helpful to avoid long movement detection delay and realize proactive handoff control rather than reactive handoff. The serial relationships can be changed, e.g. pre-authentication, pre-address-configuration, pre-registration etc. MIH (Media Independent Handoff) [16] defined by IEEE 802.21 WG provides an efficient cross-layer interaction mechanism. (2) Some parameters defined in the protocol standards should be modified for more efficient mobility supporting. (3) Special measures should be adopted to reduce the handoff packet loss. Soft handoff is another issue in handoff performance optimization.

References

1. Perkins, C.: IP Mobility Support for IPv4. RFC 3344 (2002)
2. Johnson, D.: Mobility Support in IPv6. RFC 3775 (2004)
3. Stewart, R., Xie, Q., et al.: Stream Control Transmission Protocol. RFC 2960 (2000)
4. Riegel, M., Tuexen, M.: mobile SCTP. draft-riegel-tuexen-mobile-sctp-06 (2006)
5. Vogt, C., Zitterbart, M.: Efficient and Scalable, End-to-End Mobility Support for Reactive and Proactive Handoffs in IPv6. IEEE Communications Magazine, Vol. 44, No.6 (2006)
6. Choi, H.-H., Song, O., Cho, D.-H.: A Seamless Handoff Scheme for UMTS–WLAN Interworking. Proceedings of IEEE GlobeCom (2004) 1559-1564
7. Bernaschi, M., Cacace, F.: Vertical Handoff Performance in Heterogeneous Networks. Proceedings of ICPPW (2004) 100-107
8. Lee, C.-W., Chen, L.-M., Chen, M.-C., Sun, Y.-S.: A Framework of Handoffs in Wireless Overlay Networks Based on Mobile IPv6. IEEE Journal on Selected Areas in Communications, Vol. 23, No. 11 (2005) 2118-2128
9. Ma, L., Yu, F., Leung, V.C.M.: A New Method to Support UMTS/WLAN Vertical Handover Using SCTP. IEEE Wireless Communications, Vol.11, No.4 (2004) 44-51
10. Ma, L., Yu, F., Leung, V.C.M.: SMART-FRX: A Novel Error-Recovery Scheme to Improve Performance of Mobile SCTP during WLAN to Cellular Forced Vertical Handoff. IEEE WCNC (2005) 1377-1382
11. Song, J.-K.: Performance Evaluation of Handoff between UMTS/802.11b based on Mobile IP and Stream Control Transmission Protocol. Master Thesis (2005)
12. Narten, T., Nordmark, E., Simpson, W., Soliman, H.: Neighbor Discovery for IP Version 6. draft-ietf-ipv6-2461bis-06 (2006)
13. Stewart, R., Xie, Q., Tuexen, M., Maruyama, S., Kozuka, M.: SCTP Dynamic Address Reconfiguration. draft-ietf-tsvwg-addip-sctp-17 (2006)
14. Banerjee, N., Basu, K., Das, S.K.: Hand-off Delay Analysis in SIP-based Mobility Management in Wireless Networks. Proceedings of IPDPS (2003)
15. Banerjee, N., Wu, W., Basu, K., Das, S.K.: Analysis of SIP-based mobility management in 4G wireless networks. Computer Communications, Vol. 27, No. 8 (2004) 697-707
16. Draft IEEE Std. P802.21/D00.05: Media Independent Handover Services. (2006)
17. He, X.-Y., Liu, Q., Lei, Z.-M.: Location Pre-query Scheme to Internet Mobility Support. Journal of Software, Vol. 15, No. 2 (2004) 259-267

Reliability of Wireless Sensor Network with Sleeping Nodes*

Vladimir V. Shakhov[1] and Hyunseung Choo[2,**]

[1] Institute of Computational Mathematics and Mathematical Geophysics of SB RAS
Novosibirsk 630090, Russia
shakhov@skku.edu
[2] School of Information and Communication Engineering, Sungkyunkwan University
Suwon 440-746, South Korea
choo@ece.skku.ac.kr

Abstract. Energy efficiency is an important technology of sensor networks development. In this paper we offer a method of sensors lifetime maximizing under required level of network reliability. The optimizing statement is offered. As example, we consider homogeneous sensor networks with topology like tree. Other example concerns ZigBee technology. It shown that heterogeneous network can be improved by power elements reassignments. If a sensor network topology is fixed then the most critical nodes can be detected and improved.

Keywords: Wireless Sensor Networks, Reliability, and Sleeping Time.

1 Introduction

A cost of sensor components is a critical consideration in the design of practical sensor networks. A cost of sensor network increases with sensor battery power. It is often economically advantageous to discard a sensor rather then sensor recharging. By this reason a battery power is usually a scare component in wireless devices. On the other hand, sensor lifetime depends on battery lifetime. Thus, energy efficiency is an important direction of sensor networks investigations [1].

A widely employed energy-saving approach consists in placing sensors in sleeping mode. During sleeping time a sensor can be used for data transmission or other services like environment monitoring. Hence, network reliability degrades. But sensor power consumption is very low under the mentioned mode. Hence, sensors lifetime increases. Thus, it is necessary to find a trade-off between sleeping time and network reliability. Many researchers are currently engaged in

* This research was supported by the MIC (Ministry of Information and Communication), Korea, under the ITRC (Information Technology Research Center) support program supervised by the IITA (Institute of Information Technology Assessment),IITA-2006-(C1090-0603-0046).

** Corresponding author.

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 530–533, 2007.
© Springer-Verlag Berlin Heidelberg 2007

energy-saving mechanisms developing that are based on sleeping mode application. Investigations are focused on scheduling schemes for sensors lifetime maximization [2] or protocol for network throughput increasing [3]. An interrelation of sensor battery lifetime and network reliability has not been considered. To fill this gap we offer a method of sensors lifetime maximizing under required level of network reliability.

The rest of paper is organized as follows. In section 2 we discuss a concept of reliability and state an optimizing problem for sensor lifetime maximization. In 3.1, we propose the problem solution for homogenous sensor networks. In 3.2 we produce some details for ZigBee technology (IEEE 802.15.4.) and consider the problem for heterogeneous. Section 4 is the final conclusion.

2 Sensor Reliability and Network Reliability

As it was said above a sensor network is a set of low-cost and low-power wireless sensor. Let us designate the sensor lifetime as T and the total sensor sleeping time as S. Thus, the duration of sensor active state is $T - S$. Assume that a sensor is randomly switched from active state to sleep state and inversely. From here, the probability p of single sensor availability can be calculated as

$$p = \frac{T - S}{T} = 1 - \frac{S}{T}. \tag{1}$$

If a sensor network is modeled by a graph then the probability p is a reliability of graph vertex. Thus, sensor networks with sleeping nodes are modeled by a graph with unreliable vertices. Actually, a wireless channel is also unreliable due to interference, fading, shading and so on. But a distance between sensors is not large. So, not reduce generality, we assume that edges of graph above are absolutely reliable. The probability of graph (network) connectivity is known as network reliability, R. Let n is the total number of sensors in a network. It can be shown in the trivial way that network reliability is a polynomial of degree n in the variable p. Let us remark we consider the connection for active nodes. Thus, the mentioned reliability reflects the connectivity probability of non-sleeping nodes.

The wished sensor lifetime T is usually given. It is defined by applications. But if sleeping time is maximized then the capacity of sensor battery is reduced. It also reduces a cost of sensor networks. But from (1) we can see that large sleeping time is a reason of network unreliability. Thus, it is necessary to minimize sensor availability saving given level of network reliability, C. Here $C \in [0, 1]$. Continuing this line of reasoning, we get the following optimization problem

$$p \to min \tag{2}$$

$$R(p) \geq C \tag{3}$$

Let us remark the problem is formulated for homogeneous network nodes. The parameter p is same for all sensors.

3 Reliability Calculations

3.1 Reliability of Homogeneous Sensor Network

In the optimization statement above the criterion is a monotone continuous growing function in p. It is obvious that a reliability polynomial monotonically grows when p runs over the set $[0 \dots 1]$. Thus, the solution of considered optimization problem is reached on area boundary defined by the function $R(p)$. Therefore, the optimal p is a root of equation

$$R(p) = C.$$

Let us consider the following example. Two terminals are connected by n sequential sensors. Thus, the sensor network topology is a tree then the function of reliability (3) equals $R(p) = p^n$, here we do not take into account failures of terminals.

If we solve the equation $p^n = C$ then we find an optimal solution of the problem (2). Thus, the optimal parameter p is

$$p_{opt} = C^{\frac{1}{n}}.$$

From here, the optimal sleeping time equals

$$S = T(1 - C^{\frac{1}{n}}).$$

In case of huge number of sensors and arbitrary network topology the exact solution can be impractical. Numerical methods can be used.

3.2 Reliability of Heterogeneous Sensor Network

The approach above with insignificant modification can be used for heterogeneous nodes. Now the function of reliability in (2) is a function on a few variables. In some practical cases the optimal topology has to be heterogeneous. Actually, if a sensor network topology is fixed then the most critical nodes can be detected and improved. We demonstrate it on the following example. Let us consider a Personal Area Network (PAN) technology known as ZigBee. A ZigBee network consists of three types of logical units [4]. There are ZigBee Coordinator (ZC), ZigBee Router (ZR) and ZigBee End Device (ZED). ZC is PAN coordinator. The coordinator functions depend on network applications. ZED is data terminal equipment. ZR provides a connection of ZEDs and ZC.

Assume ZC and ZED are connected by three ZigBee Routers A,B, and C. Not reduce generality we consider the case of absolutely reliable ZC and ZEDs and unreliable ZR. Let a network topology be as shown on the adjacency matrix.

	ZC	**A**	**B**	**C**	**ZED**
ZC	0	1	0	0	0
A	1	0	1	1	0
B	0	1	0	0	1
C	0	1	0	0	1
ZED	0	0	1	1	0

Let p_A be the probability of active status of ZigBee Router A, p_B be the probability of active status of ZR B, and p_C be the probability of active status of ZR C.

The network reliability is

$$R(p_A, p_B, p_C) = p_A(1 - (1 - p_B)(1 - p_C)).$$

Hence the equations (2),(3) look as follows

$$p_A + p_B + p_C \to min$$

$$p_A(p_B + p_C - p_B p_C) \geq C$$

The expressions above can be simplified. From the reasons of symmetry we get $p_B = p_C$. Now the reader will have no difficulty in finding the optimal solution. It is easy to see that the optimal p_A exceeds the probabilities p_B and p_C. Thus, for considered network topology the optimal sleeping times of network nodes are not the same. Hence, the optimal network structure are not homogeneous.

4 Conclusion

In this paper, we investigated one of the methods of cost reduction for sensors network. The offered technique allows to minimize battery capacity and calculate the optimal sleeping time of sensors. At the same time the required level of reliability is kept. The problem of sensor network reliability is reduced to an optimizing problem with monotone and continuous functions. Solutions of the problem can be calculated by Lagrange approach. In some cases a heterogeneous assignment of power is preferable. Actually, if a sensor network topology is fixed then the most critical nodes can be detected. And a network topology can be improved by power elements reassignments. In the future it is reasonable to investigate the problem of reliability maximization under given cost of sensors.

References

1. Akyildiz, I., Su, W., Sankarasubramaniam, Y., Cayirci, E.: A survey on sensor networks. Communications Magazine, Vol. 40, Issue 8. IEEE (2002) 102 - 114
2. Subramanian, R., Fekri, F.: Sleep scheduling and lifetime maximization in sensor networks: fundamental limits and optimal solutions. Proceeding of IPSN. IEEE (2006) 218 - 225.
3. Chiasserini, C.-F., Garetto, M.: An Analytical Model for Wireless Sensor Networks with Sleeping Nodes. Transactions on Mobile Computing, Vol. 5, Issue 12. IEEE (2006) 706 - 1718.
4. Vargauzin, V.: Wireless sensor networks based on IEEE 802.15.4, TeleMultiMedia, Vol. 6. Telesputnik (2005) (in Russian).

Energy Efficient Forwarding Scheme for Secure Wireless Ad Hoc Routing Protocols

Kwonseung Shin, Min Young Chung, and Hyunseung Choo*

School of Information and Communication Engineering
Sungkyunkwan University
440-746, Suwon, Korea +82-31-290-7145
{manics86,mychung,choo}@ece.skku.ac.kr

Abstract. Ad hoc networks can be used in scenarios such as military and emergency relief, and have the advantage of establishing communications independently. Securing the ad hoc routing protocol is necessary to prevent attacks in these applications. However, securing protocols for mobile ad hoc networks presents challenges due to energy consuming processes such as authentication of hosts and verification of messages. In this paper, we propose an efficient forwarding mechanism for secure ad hoc routing protocols, based on public key cryptography.

Keywords: Energy Efficiency, Public Key, and Secure Ad-hoc Routing.

1 Introduction

A mobile ad hoc network [1] is a set of wireless mobile hosts with the capability of cooperatively establishing communications independently. Due to this advantage, ad hoc networks can be used in military and emergency relief scenarios. In these and other ad hoc networking applications, security in the routing protocols is necessary to guard against attacks, such as malicious routing misdirection [2]. However, securing protocols for mobile ad hoc networks [3,4,5] presents challenges due to energy consuming processes such as authentication of hosts and verification of messages. Therefore, to extend the lifetime of wireless mobile ad hoc hosts with limited battery capacity, an energy efficient secure ad hoc routing protocol must be designed.

Most secure ad hoc routing protocols depending on the Certificate Authority (CA) provide authentication, non-repudiation, and integrity to prevent attacks such as modification, impersonation, and fabrication. Authenticated Routing for Ad Hoc Networks (ARAN) [6] is a secure ad hoc routing protocol using the CA. Initially each node receives a certificate, and public key of the CA, from the CA securely. In reactive routing protocols, when a source node wants to send a packet to a destination node, it broadcasts a route request packet if it has no route information about the destination node. However, in ARAN, the following route discovery process is performed to provide secure routing:

* Corresponding author.

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 534–537, 2007.
© Springer-Verlag Berlin Heidelberg 2007

1. The source node signs a route request message using its private key and appends its own certificate. Then, it broadcast the message to all nodes within its radio frequency coverage area.
2. The one-hop neighbor of the source ignores the duplicate requests occurring when the RREQ is received multiple times. Otherwise, it verifies the certificate of the source by using the public key of the CA. Then it checks the signature of the source by using the public key extracted from the certificate of the source. If the verification is successful, the intermediate node set up a reverse path for the source, signs the RREQ, and appends its own certificate to the message. Then, it rebroadcasts the RREQ.
3. Intermediate nodes that receive the RREQ verify the source and node that has sent RREQ directly if the message has not already been received. Then, they use the public keys extracted from two certificates attached in the RREQ to validate the signatures. The intermediate nodes remove the signature and certificate of the node that sent RREQ directly if the verification is successful. Then, the nodes sign the message and rebroadcast it with their own certificate.
4. Until the RREQ message is flooded to all the nodes in the network, process 3 is repeated.
5. When the destination node eventually receives the RREQ, it creates a RREP and unicasts it toward the source node with its signature and certificate. The flooding process of RREP is similar to the processes 2 and 3 described above.

ARAN provides end-to-end security using cryptographic certificates, thereby preventing most attacks. However, appending the certificate each time each node sends route request packets, which is flooded to all nodes in the network, causes considerable network overhead. In this paper, we propose an efficient forwarding mechanism that reduces network overhead significantly by decreasing the frequency of included certificates.

2 The Proposed Scheme

In our scheme, each node that receives the route request packet that has a certificate verifies the certificate using the public key of the CA. Then, it stores the certificate to its own repository when the verification is successful. The sender does not include its certificate in messages when other nodes in the network have the certificate. On the other hand, it must include the certificate when it infers that other nodes in the network do not have its certificate. The decision of including certificate is made by states of nodes. Each node can be in on of two states, a *Participation* state or a *Non-Participation* state.

Fig. 1 represents the two states of nodes. The detailed rules regarding to the states are described as follows:

***Non-Participation* state:** Nodes are initially in a *Non-Participation* state. When each node initiates a new route, it includes its certificate before broadcasting the route request message. Then, it enters the *Participation* state. A

node sends a message with its certificate without changing state when the message is not a route request message.

***Participation* state:** When a node enters the *Participation* state, it starts a timer T_B. The node changes to the *Non-participation* state by stopping T_B when the T_B expires. In the *Participation* state, a node does not include its certificate when it sends a message, because it infers that the other nodes in the network have its certificate. A node initiates timer T_B when it broadcasts a route request message to initiate a new route. However, it does not initiate the timer when it broadcasts a route request message which is sent from other nodes.

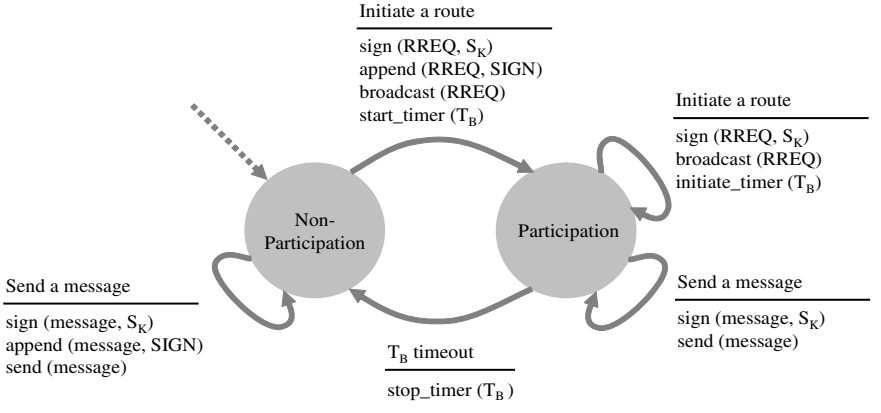


Fig. 1. The two states of nodes

Each node authenticates the sender using the public key of the CA and verifies the message using the public key included in the certificate when it receives the route request message that has a certificate from the other node. When the authentication and verification is successful, the node maintains the certificate in its repository. The certificate of a node in the repository expires when the route request message from the node is not received during T_M. Otherwise, the certificate of the node will be maintained permanently, though the node does not exist in the network. When a node receives a message that does not have certificates for authentication, it searches the certificate needed to authenticate the sender in its repository. When the certificate is found, it then verifies the signature of the received message without authentication of the sender, because the certificate in its repository has already been verified. Obviously, T_M must be a larger value than T_B, and be set by considering the variation of propagation delay of route request messages that are caused by the movement of the node or the variation of network traffic.

Although T_M is set properly, some nodes may receive the message with insufficient certificates for verification. For example, the node that newly participates in the network has no certificates of other nodes in the network. Therefore, the

new node is not able to verify the sender or message when it receives a message that does not have complete set of certificates needed for verification. In this case, to obtain the certificates needed, the new node sends a Certificate REQuest (CREQ) message to neighbor node that has sent a message informing of insufficient certificates. The neighbor node is able to send certificates to the new node because it has verified the message successfully. The CREQ is signed using the private key of the sender, and the certificate is included only if the node is in the *Non-Participation* state. The node that receives the CREQ, verifies the sender and CREQ message. Then, it sends the Certificate REPly (CREP) message, including certificates requested by the CREQ. The CREP is also signed using a private key, and the certificate is included only if the sender of CREP is in the *Participation* state. Therefore, the new node that has no certificate is able to verify the message with insufficient certificates securely.

3 Conclusion

In this paper, we proposed an energy efficient forwarding scheme for secure ad hoc routing protocols using the CA. It reduces network overhead by decreasing the frequency of sending messages with certificates. By employing additional repository in each node, the energy efficiency of proposed scheme is increased significantly compared to conventional ARAN, when the nodes in the network initiate new routes frequently. In addition, the process of verification of hosts is also reduced, by eliminating the process of authentication that has already been performed.

Acknowledgments. This research was supported by the MIC(Ministry of Information and Communication), Korea, under the ITRC(Information Technology Research Center) support program supervised by the IITA(Institute of Information Technology Assessment), IITA-2006-(C1090-0603-0046)

References

1. Corson, S., Macker, J.: Mobile Ad hoc Networking (MANET): Routing Protocol Performance Issues and Evaluation Considerations. Network Working Group, Request for Comments: 2501. http://www.ietf.org/rfc/rfc2501.txt (1999)
2. Hu, Y-C., Perrig, A.: A survey of secure wireless ad hoc routing. Security & Privacy Magazine, Vol. 2. IEEE (2004) 28-39
3. Papadimitratos, P., Haas, Z.J.: Secure Routing for Mobile Ad Hoc Networks. Proc. SCS Communication Networks and Distributed Systems Modeling and Simulation Conf. (2002)
4. Capkun, S., Hubaux, J.-P.: BISS: building secure routing out of an incomplete set of secure associations. Proc. 2nd ACM Wireless Security (2003) 21-29
5. Zapata, M.G., Asokan, N.: Securing Ad Hoc Routing Protocols. Proc. ACM Workshop on Wireless Security (2002) 1-10
6. Sanzgiri, K., LaFlamme, D., Dahill, B., Levine, B.N., Shields, C., Belding-Royer, E.M.: Authenticated Routing for Ad hoc Networks. Journal on Selected Areas in Communications, Vol. 23. IEEE (2005) 598-610

Sender-Based TCP Scheme for Improving Performance in Wireless Environment

Jahwan Koo, Sung-Gon Mun, and Hyunseung Choo*

School of Information and Communication Engineering, Sungkyunkwan University
Chunchun-dong 300, Jangan-gu, Suwon 440-746, South Korea
jhkoo@songgang.skku.ac.kr, {msgon,choo}@ece.skku.ac.kr

Abstract. Conventional TCP in wireless environment has a lack of ability to differentiate the packet losses caused by network congestion from those caused by wireless link error. When the packet losses are due to wireless link error, TCP performance degradation occurs because of its congestion control mechanism. This paper propose a sender-side TCP scheme for discriminating the cause of packet losses. Simulation results show that it outperforms existing TCP congestion mechanisms such as WestwoodNR, TCP New Jersey, and TCP Reno in terms of goodput performance in wireless links.

Keywords: TCP, congestion control, wireless environment.

1 Introduction

In wireless environment, conventional TCP is unable to differentiate the cause of packet losses, resulting in severe performance degradation [1]. Many researchers attempt to improve TCP performance in wireless networks [2]. Current research work has been classified into two directions: the end-to-end TCP modifications such as congestion control mechanism and link layer approaches that include intermediate router mechanisms. However, TCP modifications do not distinguish the cause of packet loss whether it is based on network congestion or wireless link error. Link layer approaches require more time and cost to be deployed in real wired and wireless networks. Therefore, discrimination of the cause of packet loss using the modified end-to-end TCP mechanism, without supporting the intermediate router mechanisms, is very important.

This paper propose WestwoodVT (WestwoodNR based on TCP Vegas buffer Thresholds), a sender-side TCP scheme, which discriminates the cause of packet loss and operates according to the cause when packet retransmission is required. It uses the flow control concept of TCP Vegas [3] so as to discriminate the cause. If the cause is due to network congestion, it retransmits packets using the available bandwidth estimator of WestwoodNR [4], which is an efficient TCP scheme in wired and wireless networks. Meanwhile, if the cause is due to wireless link error, it ignores the loss and responses as if the packet loss has never been occurred.

* Corresponding author.

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 538–541, 2007.
© Springer-Verlag Berlin Heidelberg 2007

2 Discriminating the Cause of Packet Loss

In this section, we present a novel scheme for improving TCP performance in wireless environment. WestwoodVT investigates the buffer state of network nodes between sender and receiver in order to discriminate the cause of packet loss. The discrimination inherits from the concept of TCP Vegas. WestwoodVT estimates the maximum *Expected* transmission rate and *Actual* transmission rate when received every ACK. The *Expected* transmission rate is given by where *WindowSize* is the size of the current congestion window, equal to the number of bytes in transit, *BaseRTT* is the minimum value of all measured RTT. In addition, *Actual* transmission rate is given where *WindowSize* is equal to *WindowSize* of Equation 1, *RTT* is the current calculated RTT.

$$Expected = \frac{WindowSize}{BaseRTT}, Actual = \frac{WindowSize}{RTT} \tag{1}$$

Additionally, WestwoodVT defines $\triangle$ is calculated by the difference of *Actual* from *Expected* transmission rate. The value of $\triangle$ presents the amount of current existing data in buffer of network nodes. Namely, $\triangle$ indicates the state of the current network. This is given by:

$$\triangle = (\frac{WindowSize}{BaseRTT} - \frac{WindowSize}{RTT}) \times BaseRTT \tag{2}$$

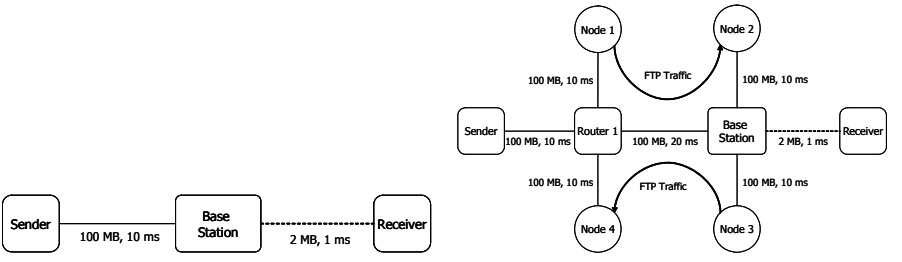
In addition, WestwoodVT defines the buffer thresholds, α and β. These values indicate to the lower and upper bound of network nodes's buffer. The discrimination of the cause of packet loss is operated in congestion avoidance (CA) phase. In CA phase, the sender estimates $\triangle$ using Eq. 1, and 2 when it receives 3-DUPACK. WestwoodVT then compares $\triangle$ to the buffer thresholds α and β.

If $\triangle$ is smaller than the threshold α, WestwoodVT assumes that the current buffer state of network nodes is loose, and decides that the packet loss is due to wireless link error. If $\triangle$ is greater than the threshold β, the current buffer state is tight. At this point, WestwoodVT assumes that the packet loss is caused by network congestion. If $\triangle$ is greater than α and smaller than β, WestwoodVT does not decide the cause of the packet loss. Due to the ambiguousness of the discriminating, WestwoodVT maintains the current state and then retransmits lost packets as the standard sending. Therefore, WestwoodVT postpones the decision of cause to the next it receives 3-DUPACK for the more accurate discriminating.

3 Performance Evaluation

The simulation environments are depicted in Figure 1(a) and Figure 1(b). In Figure 1(a), a single TCP connection running a long-live FTP application delivers data from sender to receiver, with various wireless link errors: 0.1%, 1%, 2%, 3%, 4%, and 5%, respectively. In Figure 1(a), all network nodes retain the queue size of 20 MSS. Furthermore, we construct a more complex topology, such as Figure 1(b). In this case, all node's queues are also set to 20 MSS. A single TCP

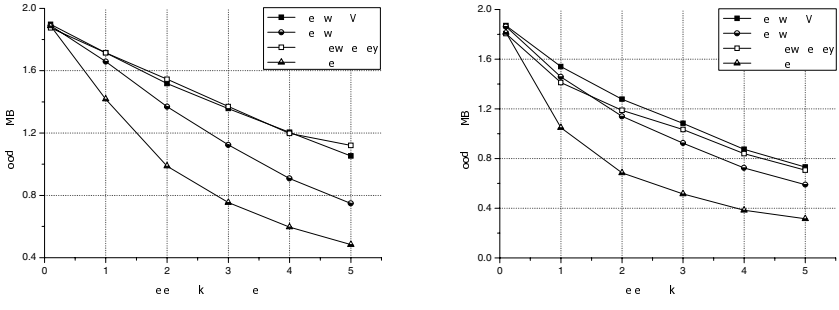
connection running a long-live FTP application delivers data from the sender to the receiver with various wireless link errors: 0.1%, 1%, 2%, 3%, 4%, and 5%, respectively. At the same time, cross-traffic flows are generated: forward traffic flows from node1 to node2 via router1 and base station, and backward traffic flows from node3 to node4 via base station and router1, respectively.



(a) Simulation topology

(b) Complex topology with background traffic

Fig. 1. Simulation topology



(a) Wireless link error rates

(b) Wireless link error rates with background traffic

Fig. 2. Goodput performance

We setup the lower bound of WestwoodVT, α to 14. if $\triangle < 14$, the cause is due to wireless link error. The upper bound, β is fixed on 16, if $\triangle > 16$, the cause is due to network congestion. Finally, we run the simulation for WestwoodVT, WestwoodNR, TCP New Jersey, and TCP Reno, respectively. Figure 2(a) demonstrates the result of simulation, an average of 10 times, over 200 seconds.

As the results of the first simulation, WestwoodVT is superior over WestwoodNR and TCP Reno in every wireless link error rates. With wireless link error rate of 1%, WestwoodVT outperforms WestwoodNR by 3% and TCP Reno by 21%. WestwoodVT outperforms WestwoodNR by 11% and TCP Reno by 54%

with wireless link error rate of 2%. Further, WestwoodVT is almost identical to TCP New Jersey, until the wireless link error rate reaches 4%. WestwoodVT achieves a maximum of 41% and 118% improvements in goodput over WestwoodNR and TCP Reno, in various wireless link errors rates, respectively.

Figure 2(b) depicts results of the second simulation over 200 seconds, an average of 10 times. The buffer thresholds, α and β of WestwoodVT are configured to 14 and 16, respectively. These values are based on the queue size of network nodes. The performance of TCP Reno is worst. WestwoodVT achieves 6% $\sim$ 24% improvements in goodput over WestwoodNR, and it is overwhelming TCP New Jersey, with every wireless error rates. In this simulation, TCP New Jersey experiences the decreasing of throughput with background traffics. However, WestwoodVT keeps its performance constantly, regardless of wireless link error rates or background traffics.

4 Conclusion

In this paper, we have proposed WestwoodVT which utilizes the sender-based transmission window control mechanism for discriminating the cause of packet loss. It checks the buffer state of network nodes between sender and receiver using TCP Vegas operating mechanism and discriminates the cause of packet loss based on the buffer state. When the packet loss is due to network congestion, WestwoodVT uses the available bandwidth estimator of WestwoodNR for packet retransmissions. Otherwise, it retransmits packets without congestion control mechanism.

Acknowledgements

This research was supported by the MIC, Korea, under the ITRC support program supervised by the IITA, IITA-2006-(C1090-0603-0046).

References

1. Xylomenos, G., Polyzos, G. C.:TCP Performance Issues over Wireless Links. IEEE Communications Magazine, Vol. 39. (2001) 52–58
2. Tian, Y., Xu, K., Ansari, N.:TCP in Wireless Environments: Problems and Solutions. IEEE Radio Communications, Vol. 43. (2005) S27–S32
3. Brakmo, L. S., O'Malley, S. W., Peterson, L. L.:TCP Vegas: New Techniques for Congestion Detection and Avoidance. ACM/SIGCOMM Computer Communication Review, Vol. 24, (1994) 24–35
4. Casetti, C., Gerla, M., Mascolo, S., Sanadidi, M. Y., Wang, R.:TCP Westwood: Bandwidth Estimation for Enhanced Transport over Wireless Links. ACM/Mobicom, (2001) 287–297

Design and Implementation of DLNA DMS Through IEEE1394*

Gu Su Kim[1], Chul-Seung Kim[1], Hyun-Su Jang[1], Moon Seok Chang[2], and Young Ik Eom[1]

[1] School of Information and Communication Engineering, Sungkyunkwan University, 300 Chunchun-dong, Jangan-gu, Suwon, Kyunggi-do, 440-746, Korea
{gusukim,hallower,jhs4071,yieom}@ece.skku.ac.kr
[2] Advisory S/W Engineer Kernel Performance Server Group, MS 9571 11400 Burnet Rd. Austin, TX 78758, USA
moon@austin.ibm.com

Abstract. With the technological growth of the AV(Audio and Video) equipments and the popularization of video contents, IEEE1394 for the transmission of AV data has spread fast. However, DLNA(Digital Living Network Alliance), which is a standard of the home network middleware, does not consider the transmission of AV data over IEEE1394. In this paper, we propose a scheme for transmission of AV contents over IEEE1394 in DLNA environments. In addition, by the validation test of DLNA compatibility, we show that our proposed scheme can support AV contents sharing without regard to the manufactures and transmission media.

Keywords: AV, DLNA, IEEE1394.

1 Introduction

With the progress of various IT technologies and the proliferation of high speed access networks and digital convergence, the concern on home network environments has increased and many home network middleware technologies and standards have been presented [1]. Digital Living Network Alliance(DLNA) presents the Home Network Device Interoperability Guidelines v1.0 for sharing the AV (Audio and Video) contents [2,3]. However, these guidelines support only HTTP-GET method as the basic protocol for transmission of AV contents and do not consider IEEE1394 that is widely being used in devices of home network environments [4,5]. Therefore, the AV devices using IEEE1394 transmit AV contents through manufacturer specific methods in DLNA environments and there are problems when different manufacturer devices share the AV contents.

In this paper, we propose IEEE1394 DMS(Digital Media Server) architecture for DLNA environments and show the implementation of the proposed IEEE1394

* This research was supported by MIC, Korea under ITRC IITA-2006-(C1090-0603-0046).

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 542–545, 2007.
© Springer-Verlag Berlin Heidelberg 2007

DMS architecture and test whether our IEEE1394 DMS prototype has the compatibility with DLNA standard.

The rest of the paper is organized as follows. In Section 2, we briefly show the design of our proposed IEEE1394 DMS architecture. Section 3 presents the experimental results focusing on the DLNA compatibility. Finally, Section 4 concludes with a summary.

2 The Architecture of IEEE1394 DMS

The proposed IEEE1394 DMS architecture uses the IEEE1394 as the basic transmission interface and consists of Content Creation Subsystem, Content Management Subsystem, Content Storage Database, and Content Transfer Subsystem. Figure 1 shows the IEEE1394 DMS architecture.

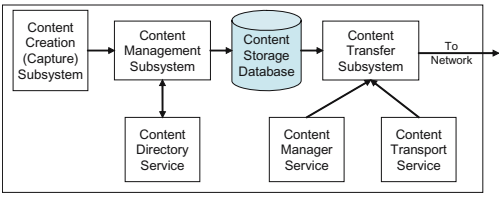


Fig. 1. IEEE1394 DMS architecture

We constructed the software stack of our IEEE1394 DMS architecture which consists of IEEE1394 interface, IEEE1394 device driver, IEEE1394 library, IPv4 over IEEE1394 driver, TCP/IP stack, UPnP DA(Device Architecture)/AV stack, and DMS applications.

DLNA device using the IEEE1394 interface can get the information for the transmission of AV contents through the device description. The device description expresses that the device has the GUID number of IEEE1394 interface and plays the role of DLNA DMS and provides the service of UPnP DA stack and Contents Directory, Connection Manager, and AV Transport Service of UPnP AV stack [6,7]. In addition to these basic information, Our URI for IEEE1394 DMS using IEC61883 protocol provides additional information as follows: the bandwidth of IEEE1394, the GUID of Isochronous Resource Manager(IRM) assigning the isochronous channel, the contents type, the GUID of IEEE1394 DMS, and PCR number.

3 Implementation and Experiment Results

We implemented IEEE1394 DMS on Redhat 9 Linux platform and used IEEE1394 driver for Linux[8] and Eth1394 module, which is IPv4 over IEEE1394 driver[9]. Also, we used UPnP stack of twonkyvision [10], whose DMS is compatible with DLNA.

```
event: -> device list changed
question=SUBSCRIBE /ContentDirectory/Event HTTP/1.1
HOST: http://192.168.0.1:9000/DeviceDescription.xml
CALLBACK: <http:///(null)/>
NT: upnp:event
TIMEOUT: second-30000

question=SUBSCRIBE /AVTransport/Event HTTP/1.1
HOST: http://192.168.0.1:9000/DeviceDescription.xml
CALLBACK: <http:///(null)/>
NT: upnp:event
TIMEOUT: second-30000

event: -> device list changed
>List of Media Servers:
[0] [dlna-test-linux] 1394 DMS
>List of Media Renderer:
[action server: - NONE -]
[active renderer: local]
```

Fig. 2. URI for IEC61883 protocol

Figure 2 shows the result that IEEE1394 DMP confirms IEEE1394 DMS and its service, and IEEE1394 DMS was registered as a media server.

Figure 3 shows the process that DMP selects and plays an AV content. In Figure 3, the first paragraph shows the result that DMP performs AVT: SetAVTra nsportURI action including URI (iec61883://00601d00000006bd:63/disk/video/ O37.mpg?DLNA.ORG_PN=MPEG_PS_NTSC;DLNA.OR G_OP=01) for selection of the content. The second paragraph shows the response of the SetAVTransportURI action and the third paragraph shows AVT:Play action. The last paragraph shows the response from the AVT:Play action.

```
<?xml version="1.0" encoding="utf-8"?><s:Envelope xmlns:s="http://schemas.xmlsoap.org/soap/envelope/"
s:encodingStyle="http://schemas.xmlsoap.org/soap/encoding/"><s:Body><u:PrepareForConnection xmlns:u="urn:schem
-upnp-org:service:ConnectionManager:1"> <RemoteProtocolInfo>iec61883:000000000000000:MPEG2_TS:00601d00000006
;0<RemoteProtocolInfo>  <PeerConnectionManager></PeerConnectionManager> <PeerConnectionID>0</PeerConnectionID>
<Direction>INPUT</Direction>    </u:PrepareForConnection>       </s:Body></s:Envelope>
-------------
*** Con ID = 0, *** AVT ID = 0, *** Rcs ID= -1
*** PrepareForConnection 0
[active server: [dlna-test-linux] 1394 DMS ]
[active renderer: local]
x

-------------
<?xml version="1.0" encoding="utf-8"?><s:Envelope s:encodingStyle="http://schemas.xmlsoap.org/soap/encoding/"
lns:s="http://schemas.xmlsoap.org/soap/envelope/"><s:Body><u:SetAVTransportURI xmlns:u="urn:schemas-upnp-org:s
vice:AVTransport:1"><InstanceID>0</InstanceID><CurrentURI>iec61883://00601d00000006bd:63/disk/video/037.mpg?DL
.ORG_PN=MPEG_PS_NTSC;DLNA.ORG_OP=01</CurrentURI><CurrentURIMetaData /></u:SetAVTransportURI></s:Body></s:Envel

-------------
<?xml version="1.0" encodingW"utf-8"?>
<s:Envelope xmlns:s="http://schemas.xmlsoap.org/soap/envelope/" s:encodingStyle="http://schemas.xmlsoap.org/so
/encoding/">
<s:Body>
<u:SetAVTransportURIResponse xmlns:u="urn:schemas-upnp-org:service:AVTransport:1" />
</s:Body>
</s:Envelope>

-------------
<?xml version="1.0" encoding="utf-8"?><s:Envelope s:encodingStyle="http://schemas.xmlsoap.org/soap/encoding/"
lns:s="http://schemas.xmlsoap.org/soap/envelope/"><s:Body><u:Play xmlns:u="urn:schemas-upnp-
org:service:AVTransport:1"><InstanceID>0</InstanceID><Speed>1</Speed></u:Play></s:Body></s:Envelope>
-------------
<?xml version="1.0" encoding="utf-8"?>
<s:Envelope xmlns:s="http://schemas.xmlsoap.org/soap/envelope/" s:encodingStyle="http://schemas.xmlsoap.org/so
/encoding/">
<s:Body>
<u:PlayResponse xmlns:u="urn:schemas-upnp-org:service:AVTransport:1" />
</s:Body>
</s:Envelope>
[active server: [dlna-test-linux] 1394 DMS ]
[active renderer: local]
```

Fig. 3. URI for IEC61883 protocol

In order to verify that our implemented IEEE1394 DMS has compatibility with DLNA, we tested our system through DLNA Conformance Test Tool(CTT) [11]. Our test items are as follows: device description, URI, DMS, contents directory service, connection manager service, GetProtocolInfo of connection manager service, contents browsing, ProtocolInfo, and MM URI item.

In the CTT verification, our system satisfied above items except for GetProtocolInfo and ProtocolInfo items. GetProtocolInfo and ProtocolInfo items are failed in our tests because the current DLNA specification supports only HTTP-GET protocol and does not support IEC61883 protocol. In order to solve these failures in CTT test, it is needed that the future DLNA 2.0 will have to define IEEE1394 as transmission protocol and CTT will have to reflect the test items for IEEE1394.

4 Conclusion

In this paper, we proposed the architecture of IEEE1394 DMS for sharing of AV contents in DLNA envrionments and implemented the device description for IEEE1394 DMS, AV Transport Service, and URI for content specification over IEEE1394. Also, we showed that our proposed scheme is compatible with DLNA by testing through DLNA CTT. Our proposed IEEE1394 DMS allow devices only with IEEE1394 interface such as TV, STB, and DVD player, to share AV contents on DLNA environments.

References

1. T. Nakajima and I. Satoh, "A software infrastructure for supporting spontaneous and personalized interaction in home computing environments," Personal and Ubiquitous Computing, Springer-Verlag, Vol. 10, No. 6, pp. 379 - 391, Sep. 2006.
2. DLNA Homepage, http://www.dlna.org/.
3. DLNA, Home Networked Device Interoperability Guidelines v1.0, Jun. 2004.
4. IEEE Std 1394a-2000, IEEE Standard for a High Performance Serial Bus–Amendment 1, Mar. 2000.
5. IEEE Std 1394b-2002, IEEE Standard for a High-Performance Serial Bus–Amendment 2, 2002.
6. UPnP Forum, UPnP Device Architecture 1.0, May 2003.
7. UPnP Forum, UPnP AV Architecture 0.83, Jun. 2002.
8. IEEE1394 for Linux, http://www.linux1394.org/
9. IPv4 over IEEE1394 (RFC2734), http://www.ietf.org/rfc/rfc2734.txt?number=2734
10. http://www.twonkyvision.com/index.html
11. DLNA Conformance Test Tool, http://www.dlna.org/members/ctt/

Efficient Measurement of the Eye Blinking by Using Decision Function for Intelligent Vehicles

Ilkwon Park[1], Jung-Ho Ahn[2], and Hyeran Byun[1]

[1] Dept. Of Computer Science, Yonsei University, Seoul, 120-749, Korea
[2] Dept. Of Computer & Media Engineering, Kangnam University, South Korea
{ikheart,jungho,hrbyun}@yonsei.ac.kr

Abstract. In this paper, we propose an efficient measurement of the eye blinking for drowsy driver detection system that is one of the driver safety systems for the intelligent vehicle. However, during the real driving in the daytime, driver's face is exposed to various illuminations. It makes too difficult to monitor driver's eye blinking. Therefore, we propose efficient formation of the cascaded form of Support Vector Machines (SVM) as eye verification to boost the accuracy of eye detection. Furthermore, for an efficient measurement of eye blinking, we newly define decision function that is based on the measure of eyelid movement and the weight generated by the eye classifier. In the experiments, we can show the reliable performance for our own test data acquired during a real driving in the various illumination conditions. Furthermore, through our proposed method, we use detected eye blinking for Drowsy Driver Detection System.

Keywords: Eye Blinking Detection, Eye Detection, Drowsy Detection, Cascaded SVM.

1 Introduction

In the safety driving application related with driver's drowsiness, many computer vision-based systems have been successfully applying to real application because it is quite not only accurate but also nonintrusive method for a driver [4], [5], [8].

However, this type of approach is very sensitive to change in illumination [9]. To solve these problems, some systems used the pupil reaction against infrared light pupil reaction as the difference between pupil which appears bright and that which appears dark [4], [5], [8], [9]. However, this kind of research has mainly been performed in the indoor studio without the interruption of infrared illumination from sunlight.

In this paper, we extend our previous approach that tolerates various illuminations changing not only in the daytime but also at night. In the proposed our system, we propose the eye verification by using two-level cascaded SVM that is newly configured. Moreover, we newly define decision function based on eyelid-movement and the weight generated by the eye classifier for detecting eye closed state. Open eye state and closed eye states are distinguished by using the decision function. The remainder of this paper is organized as follows. In Section 2, we describe a robust eye detection method. In Section 3, we configure the classifiers with the two level cascade form of SVM (Two-Level Cascaded SVM) for the eye verification. In Section 4, we

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 546–549, 2007.
© Springer-Verlag Berlin Heidelberg 2007

test the decision function for the measure of eye blinking. Experimental results and the estimation of our system are provided in Section 5, and finally we make a conclusion in Section 6.

2 Extraction of Eye Candidates

The input frame includes both the driver's face and the background which are parts of the vehicle as well as scenery outside the car. In order to measure eye blinking, the position of eyes should be detected. Eye corner filter is a fast and reliable preprocessing method for eye detection. In this paper, the eye corner filter is partially used to detect suitable candidates of eye region as [3].

As input frame is convoluted with the filter, we can detect the positive areas over the threshold. The threshold is determined to take 10% from high rank of the result by a convolution with total eye area. Each positive area detected by eye corner filter is cropped with 5 pixels margins from the boundary of positive area as eye candidates if each positive area has a reasonable ratio of horizontal length to vertical length. Cropped eye candidates are normalized the size of 41 by 21 for the eye verification.

3 Eye Detection Using Two-Level Cascaded SVM

In this section, we proposed a novel efficient verification scheme using SVM. In general, eye verification deals with the two-class problem of the eye vs. the non-eye. However, the variation within eye class is large due to the inhomogeneity of open and closed eye data. This fact drops down the verification performance.

Therefore, we considered three classes: open eye, closed eye and non-eye and if filtered eye candidates are classified as open or closed eye, we verified it as the eye. For computational efficiency, we evaluated the SVM with the sequential evaluation algorithm [9] that we called the cascaded SVM. Considering the three classes, we designed the verification procedure with the two-level cascaded SVM. First classifier represents a series of classifiers for open eyes and non-eyes and Second classifier represents a series of classifiers for closed eyes and non-eyes.

4 Eyelid Movement Estimation

In this section, we propose the decision function for eyelid movement estimation. We can determine open and closed eyes using a decision function based on the state of eyelid as follows. Fig.1 illustrates the angles between eye corners and the top of upper eyelid. we define the open degree θ of the eye that is calculate absolute summation of $\theta_L(x)$ and $\theta_R(x)$ as and the marginal value of the open eye detector f^O described in the previous section.

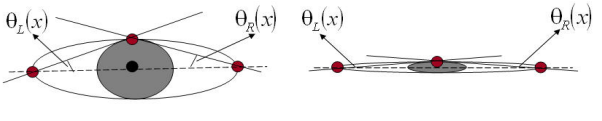


Fig. 1. The measurement of eyelid movements

the open eye classifier f_D is defined as

$$f_D(\mathrm{x}) = w_1\theta(\mathrm{x}) + w_2 f^O(\mathrm{x}) \cdot \tag{1}$$

where weights w_1 and w_2 specify the relative importance of $\theta(\mathbf{x})$ and $f^O(\mathbf{x})$. Here, the value of $f^O(\mathbf{x})$ is defined by $f_m{}^O$ $(\mathbf{x})$ if the patch $\mathbf{x}$ has negative value at some layer $m(<Nz)$ during the sequential evaluation in first classifier, or $f_{Nz}{}^O$ $(\mathbf{x})$ otherwise. The threshold T for the decision function f_D can be defined as

$$T = w_1\theta_T + w_2\varepsilon_T \cdot \tag{2}$$

In (2), θ_T is a person-specific threshold for the open degree of the eye. It is determined by a quarter of maximum degree of open eyes during first three seconds in the beginning. The constant ε_T is a person-independent threshold that is usually accepted as a small value, say 0.05. Using (1) and (2), we determine the patch $\mathbf{x}$ as the open eye if decision function is over the threshold T.

5 Experiments

In this experiment, we assume that eye detection and the measurement of eye blinking strongly depend on illumination change during real driving, regardless of nondrowsy state or drowsy state. Proposed system consists of three parts; eye detection part, verification part, and eyelid movement estimation part, we define novel decision function for eye blinking adapted for each driver.

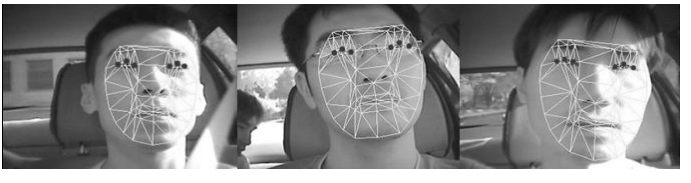


Fig. 2. Results of the eye detection process under various illumination conditions

Fig. 2. represents the results of the eye detection process. In the Fig. 2., two eye corner points and a top point of eyelid are also accurately detected in the various illuminations.

6 Conclusion

For robust eye detection, we newly designed two-level cascade form of SVM. We also proposed the decision function for the measure of eye blinking. This function could make more robust decision for the measure of eye blinking. Through our experiments, in the illumination circumstance as a daytime, our proposed method showed the outperforming the conventional method that use a difference between bright pupil and dark pupil by using pupil reaction against infrared light. In the future work, we will integrate this proposed algorithm with our Drowsy Driver Detection System (DDDS).

Acknowledgments

This research was supported by the Ministry of Information and Communication, Korea under the Information Technology Research Center support program supervised by the Institute of Information Technology Assessment, IITA-2005-(C1090-0501-0019).

References

1. Burges, C.J.C.: Simplified support vector decision rules. Intl. Conf. on Machine Learning, Vol. 1, (1996)71-77
2. Dinges, D., Grace, R.: PERCLOS: A Valid Psycho-physiological Measure of Alertness as Assessed by Psychomotor Vigilance, (1998), TechBrief FHWA-MCRT- 98-006
3. D'Orazio, T., Leo, M., Cicirelli, G., Distante, A.: An algorithm for real time eye detection in face images, In Proc.17th Int. Conference on Pattern Recognition, Vol. 3, (2004)278 – 281
4. Hamada, T., Ito, T., Adachi, K., Nakano, T., Yamamoto, S.: Detecting method for drivers' drowsiness applicable to individual features, In Proc. Intelligent Transportation Systems, Vol. 2 (2003),1405 – 1410
5. Hayami, T., Matsunaga, K., Shidoji, K., Matsuki, Y.: Detecting drowsiness while driving by measuring eye movement - a pilot study, In Proc. 5th Int. Conference on Intelligent Transportation Systems, (2002) 156 – 161
6. Hayashi, K., Ishihara, K., Hashimoto, H., Oguri, K.: Individualized drowsiness detection during driving by pulse wave analysis with neural network, IEEE, the Proceedings of Intelligent Transportation Systems, (2005) 901 – 906
7. Horne, J., Reyner, L.: Vehicle accidents related to sleep – a review, Occupational and Environmental Medicine, Vol. 56, (1999), 289-294
8. Ji, Q., Yang, X.J.: Real-time eye, gaze, and face pose tracking for monitoring driver vigilance, Real-Time Imaging, Vol. 8, Issue 5, (2002) 357 – 377
9. Park, I., Ahn, J.H., Byun, H.: Efficient Measurement of Eye Blinking under Various Illumination Conditions for Drowsiness Detection Systems, 18th International Conference on Pattern Recognition, Vol. 1, (2006) 383 – 386

Dynamic Bandwidth Allocation Algorithm Based on Two-Phase Cycle for Efficient Channel Utilization on Ethernet PON*

Won Jin Yoon, Woo Jin Jung, Tae-Jin Lee, Hyunseung Choo, and Min Young Chung**

School of Information and Communication Engineering
Sungkyunkwan University
300, Chunchun-dong,Jangan-gu, Suwon, Kyunggi-do, 440-746, Korea
+82-31-290-7990
{yoon1007,jwj0107,tjlee,choo,mychung}@ece.skku.ac.kr

Abstract. Ethernet passive optical network (EPON) is a low-cost and high speed solution to the bottleneck problem. We study previous inter-optical network unit (ONU) scheduling algorithm, interleaved polling with stop (IPS).The IPS algorithm needs an idle time period between two consecutive frame transmission cycles on uplink. In this paper, we propose a two-phase cycle dynamic bandwidth allocation (TCDBA) algorithm to increase the channel utilization on uplink through the elimination of the idle period. And, we also evaluate the performance of the proposed algorithm by simulations and confirm its effectiveness.

Keywords: Ethernet passive optical network (EPON), multipoint control protocol (MPCP), dynamic bandwidth allocation (DBA).

1 Introduction

In order to adapt data traffic increasing due to the advent of various multimedia applications, EPON has been introduced[1][2]. In EPON, all data to be transmitted are encapsulated in Ethernet frames. Thus, worldwide-spread Ethernet-based LANs can be cost-effectively interconnected through EPON. An EPON is a point-to-multipoint optical network consisting of one optical line terminator (OLT), several ONUs, and a passive optical splitter/combiner[1]. The EPON uses multipoint control protocol (MPCP) to exchange control messages between an OLT and ONUs for the time slot request and allocation[2].

Inter-ONU scheduling algorithms are responsible for arbitrating the transmissions of ONUs. However, the EPON dose not establish standardization of data

* This work was supported by the Korea Research Foundation Grant funded by the Korean Government(MOEHRD)(KRF-2005-042-D00248) and by the MIC(Ministry of Information and Communication), Korea, under the ITRC(Information Technology Research Center) support program supervised by the IITA(Institute of Information Technology Assessment), IITA-2006-(C1090-0603-0046).

** Corresponding author.

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 550–553, 2007.
© Springer-Verlag Berlin Heidelberg 2007

transmission scheduling algorithm between ONUs. In order to fairly distribute uplink resource to all ONUs, the IPS algorithm has been proposed[3]. In the IPS algorithm, an OLT allocates the time duration based on time slot requests of entire ONUs. However, the IPS needs the computation time to calculate time durations to be assigned to the entire ONUs. During the computation time, any ONU can not transmit data to the OLT. In order to use this waste time duration, Assi et al. proposed fast gate DBA (FGDBA)[4]. However, in FGDBA, if the request volume of all the ONUs is larger than the predetermined threshold value, the operation of the FGDBA is similar to that of the IPS. Hence, it can not efficiently use the waste time duration under heavy traffic loads.

In this paper, in order to fully utilize uplink in EPON, we propose a TCDBA algorithm. In TCDBA, each cycle is divided into two phases. During the first phase, OLT can receive data from all ONUs while computing the available time durations to be allocated in the second phase.

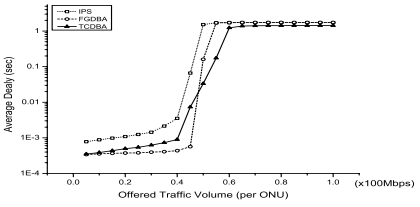
2 TCDBA (Two-Phase Cycle Dynamic Bandwidth Allocation) Algorithm

In TCDBA, a transmission cycle is divided into two phases, phases 1 and 2, in order to efficient use uplink. The time duration of phase 1 is set as the fixed value which is larger than the sum of DBA computation time and the maximum round trip time (RTT), T_{mrtt}, between the OLT and ONUs. However, the time duration of phase 2 is varied as the request information of all the ONUs. OLT can inform all the ONUs of their available time durations before the corresponding ONUs start their data transmissions on uplink in the phase 2. Hence, the phase 2 starts after the instant of ending phase 1.

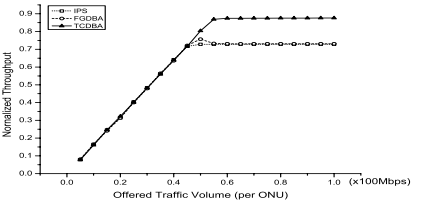
Let T_{phase1} be the sum of the DBA computation time, T_{comp}, and T_{mrtt}. and it is equally divided into N time slots, A^{p1}. As soon as the OLT receives a request volume V_i (bytes) from ONU i (i=1,...,N), the OLT determines the remaining request volume, V_i^{p2}, for ONU i. V_i^{p2} is calculated as $max(0, V_i - A^{p1})$. After the OLT informs all the N ONUs of A^{p1}, it starts DBA computation based on V_i^{p2} for all i (i=1,...,N) to allocate the available time slots in the upcoming phase 2. To compute the available time slots, we first decide the length of the current transmission cycle, T_{cycle}, as follows.

$$T_{cycle} = min(T_{max}, \sum_{i}^{N} (2 \cdot T_g + \frac{V_i \times 8}{C})), \tag{1}$$

where T_g is a guard time interval between the two consecutive time durations, T_{max} is the length of the maximum-permitted transmission cycle and C is the link capacity between the OLT and all the ONUs. Thus, the time duration of phase 2, T_{phase2}, becomes the difference between T_{cycle} and T_{phase1}. For each ONU, the size of the minimum guaranteed time slot, B_{min}, in the upcoming phase 2 is given by $(T_{phase2}/N) - T_g$. Based on B_{min} and V_i^{p2}, the OLT classified all the ONUs into two subset, M and M^c. M is the set of ONUs with $V_i^{p2} \leq B_{min}$



(a) Average packet transmission delay for IPS /FGDBA/TCDBA



(b) Normalized throughput for IPS/FGDBA /TCDBA

Fig. 1. Simulation results for IPS /FGDBA/TCDBA

and M^c is the complement of M. For ONU i in M, the OLT allocates time slot, A_i^{p2}, in the phase 2. A_i^{p2} is the smaller one between B_{min} and V_i^{p2}. To allocate the remaining time to all the ONUs in M^c, the OLT calculates the total excessive time ($B_{total}^{excess} = \sum_{i \in M} (B_{min} - V_i^{p2})$). For ONU j in M^c, the excessive time, B_j^{excess}, is given by

$$B_j^{excess} = \frac{B_{total}^{excess} \times V_j^{p2}}{\sum_{j \in M^c} V_j^{p2}}. \tag{2}$$

Finally, the OLT allocates the time slot used in phase 2 to ONU j in M^c as follows.

$$A_j^{p2} = B_{min} + B_j^{excess}. \tag{3}$$

3 Performance Evaluation

We use an EPON architecture with 16 ONUs and an OLT that expends 200μs for the DBA computation. The link capacity between an OLT and ONUs is 1Gbps. The maximum cycle time is set to 2ms and each ONU has a 10Mbytes buffer. The guard time between two consecutive time slots is set to 1μs. The inter-frame gap between two Ethernet frames is 12 bytes. We, also, assume that the round trip time between the OLT and all the ONUs is the same as 200μs. Each end-user generates high class packets with 70 bytes as a Poisson process, and medium/low class packets as a self-similar traffic pattern [5].

The average delay of IPS, FGDBA, and TCDBA varying offered traffic volume ρ per ONU is shown in Fig. 1(a). The average packet transmission delay of the TCDBA is always less than that of the IPS, because the IPS can not use the uplink during the sum of T_{comp} and T_{mrtt}. Since the TCDBA can fully use the idle time, it transmits more upstream data traffic than other algorithms. Especially, for 45Mbps$\leq \rho$, its delay is smaller than that of FGDBA and IPS. Moreover, for ρ in [40Mbps, 60Mbps] the average delay of FGDBA drastically increases as ρ increases. The reason is that some ONUs can not transmit their upstream data during idle time, because their request volume is larger than the predetermined threshold value. The waste of this period in FGDBA will make the increase of the average packet transmission delay. However, in the TCDBA, since the change of the traffic load dose not influence on the use of idle time, the average transmission delay more slowly increases. Fig. 1(b) shows the normalized throughput of IPS, FGDBA and TCDBA. Compared with the IPS and FGDBA, the proposed TCDBA, the maximum throughputs are about 71.5%, 75%, and 87.4%, respectively. Under overload condition, i.e., 60Mbps$< \rho$, the maximum throughput of TCDBA is improved about 15.9% compared with that FGDBA.

4 Conclusion

This work proposed the TCDBA algorithm in order to fully use the uplink and improve the throughput. From the simulation results, TCDBA shows a higher performance than other algorithms in terms of low delay and high throughput. For further studies, the performance evaluation of the TCDBA algorithm is required for supporting differentiated services.

References

1. M. P. McGarry, M. Maier, and M. Reisslein : Ethernet PONs: A Survey of Dynamic Bandwidth Allocation (DBA) Algorithms. IEEE Journal of Lightwave Tech. Vol. 22 No. 11 (2004) 2483–2497
2. IEEE 802.3 Ethernet in the First Mile Study Group[Online]. Available: http://www.ieee802.org/3/efm/public/index.html
3. J. Xie, S. Jiang, and Y. Jiang : A Dynamic Bandwidth Allocation Scheme for Differentiated Services in EPONs, IEEE Optical Comm. Vol. 42 Issue 8 (2004) pp. S32–S39
4. C. M. Assi, Y. YE, S. Dixit, and M. A. Ali : Dynamic Bandwidth Allocation for Quality-of-Service Over Ethernet PONs, IEEE Journal on Sel. Areas in Comm. Vol. 21 No. 9 (2003) pp. 1467–1477
5. D. Sala and A. Gummalla : PON Functional Requirements: Services and Performance [Online]. Available: http://grouper.ieee.org/groups/802/3/efm/public/jul01/presentations/sala_1_0701.pdf

Performance Evaluation of Binary Negative-Exponential Backoff Algorithm in Presence of a Channel Bit Error Rate*

Bum-Gon Choi, Hyung Joo Ki, Min Young Chung**, and Tae-Jin Lee

School of Information and Communication Engineering
Sungkyunkwan University
300, Chunchun-dong, Jangan-gu, Suwon, Kyunggi-do, 440-746, Korea
{gonace,ki0724,mychung,tjlee}@ece.skku.ac.kr

Abstract. The IEEE 802.11 standard basically uses a DCF (Distributed Coordination Function) to access a wireless channel. However, the DCF uses wireless resource ineffectively when there are many contending stations and a high bit error rate. To enhance the performance of the wireless LAN, Ki et al. proposed a Binary Negative-Exponential Backoff (BNEB) algorithm. We found that the performance of the BNEB algorithm was better than the conventional DCF. However, erroneous channel environment was not considered. In our work, we propose an analytical model for the BNEB algorithm in the presence of transmission error and compare the performance of the DCF with the BEB algorithm to that with the BNEB algorithm. From the result, the BNEB algorithm yields better performance than the DCF when the bit error rate (BER) $\leq 10^{-5}$.

Keywords: DCF, BNEB, BER, Contention Window.

1 Introduction

The IEEE 802.11 medium access control (MAC) employs the distributed coordination function (DCF) [1]. The DCF is a contention-based channel access function adopting a carrier sense multiple access with collision avoidance (CSMA/CA) for frame transmission during the contention period. However, in the DCF, the more the number of stations uses the wireless resources, the more collision occurrences are possible. To solve these problems, much research on the performance of the IEEE 802.11 DCF has been conducted.

Bianchi presented an analytical model and showed that the proposed model was very accurate [2][3]. The performance of the DCF in the presence of transmission error was evaluated in [4][5]. For effective management of wireless resources,

* This work was supported by a grant No.R01-2006-000-10402-0 from the Basic Research Program Korea Science and Engineering Foundation of Ministry of Science & Technology and by the MIC(Ministry of Information and Communication), Korea, under the ITRC(Information Technology Research Center) support program supervised by the IITA(Institute of Information Technology Assessment), IITA-2006-(C1090-0603-0046).

** Corresponding author.

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 554–557, 2007.
© Springer-Verlag Berlin Heidelberg 2007

Ki et al. proposed a Binary Negative-Exponential Backoff (BNEB) algorithm [6]. The BNEB algorithm increases the contention window to the maximum window size when stations experience a collision and decreases the contention window size by half when a transmission is successful. In [6], the results showed that the BNEB had better performance than the DCF with the BEB algorithm in ideal channel conditions. However, the performance of the BNEB algorithm may depend on the presence of transmission error. Thus, we propose an analytical model for the BNEB algorithm and evaluate the performance of the BNEB algorithm in presence of transmission error.

The rest of this paper is organized as follows. Section 2 describes an analytical model of the BNEB algorithm in the presence of transmission error under the saturation condition. In Section 3, we describe our verification of our analytical model by simulations and compare the throughput of the DCF with the BNEB algorithm to that with BEB algorithm. Finally, we conclude in Section 4.

2 Analytical Model for BNEB

Our analysis assumes that there are n stations having frame(s) to transmit and each station has frame(s) after successful transmission. In addition, it is possible that a station fails to transmit a frame because of channel noise. For a station, $s(t)$ and $b(t)$ are defined as the random process representing the backoff stage and the backoff counter at time t. The BNEB algorithm can be modeled as a bi-dimensional discrete-time Markov chain $(s(t), b(t))$. Let $b_{i,j} = \lim_{t\to\infty} P\{s(t) = i, b(t) = j\}$ and p be the transmission failure probability that a station experiences a collision or transmission error in a slot. Then, the state transition probabilities and state relations are given in [6]. And p is given by:

$$p = 1 - (1-\tau)^{n-1}(1-BER)^{l+H}, \tag{1}$$

where the BER, l and H represent the channel bit error rate, packet payload size and the packet header size.

From [6] and p, we can obtain $b_{0,0}$:

$$b_{0,0} = \frac{1}{\frac{(Wp+1)+\frac{1-p}{2}[W(\frac{1-p}{2})^m-1]}{p(1+p)} + \frac{W+1}{2}\left(\frac{1-p^L}{1-p}\right)}. \tag{2}$$

Let τ be the probability of attempting to transmit a frame. Then we have

$$\tau = \sum_{i=-m}^{L} b_{i,0} = \left(\frac{1}{p} + \frac{1-p^L}{1-p}\right) b_{0,0}. \tag{3}$$

Using τ, a transmission success probability P_s, a collision probability P_c and a transmission error probability P_{er} are calculated as

$$P_s = \frac{n\tau(1-\tau)^{n-1}}{P_{tr}}(1-PER), \tag{4}$$

$$P_c = 1 - \frac{n\tau(1-\tau)^{n-1}}{P_{tr}}, \tag{5}$$

$$P_{er} = \frac{n\tau(1-\tau)^{n-1}}{P_{tr}} PER, \tag{6}$$

where PER is the packet error rate and P_{tr} is the probability that there is at least one transmission.

$$P_{tr} = 1 - (1-\tau)^n. \tag{7}$$

Finally, we can obtain the normalized saturation throughput of the BNEB algorithm in the presence of a channel bit error rate as follows.

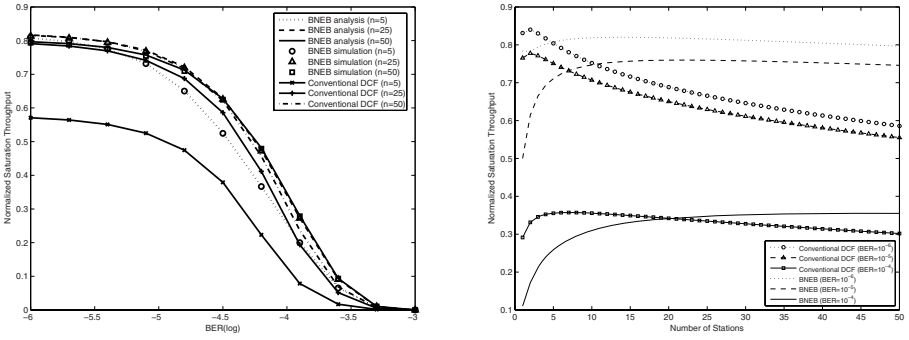
$$S = \frac{P_{tr} P_s l}{(1-P_{tr})\sigma + P_{tr} P_s T_s + P_{tr} P_c T_c + P_{tr} P_{er} T_{er}}, \tag{8}$$

where σ is the duration of a backoff slot, T_s, T_c and T_{er} are the average time intervals that the medium is sensed busy due to a successful transmission, a collision or an error transmission, respectively.

3 Performance Evaluations

To evaluate the performance of the DCF with the BEB and BNEB algorithms in the presence of a transmission error, we used the table in [6] for the MAC parameters. The normalized saturation throughput of the DCF with the BEB and BNEB algorithms is shown in Figure 1. Figure 1(a) illustrates that the analytical results of the BNEB algorithm are close to the simulation results and depicts the throughput difference between the DCF with the BEB and BNEB algorithms as the BER increases. In the result, the number of contending stations does not affect the saturation throughput of the BNEB algorithm in contrast with the DCF with the BEB algorithm when $BER \leq 10^{-5}$. When $BER > 10^{-5}$, the possibility of transmission failure by a transmission error is larger than that of transmission failure by a collision. Therefore, the difference of the DCF with the BEB and BNEB algorithms is decreased and the throughput of both algorithms reaches to 0 as BER increases.

Figure 1(b) shows the saturation throughput of the DCF with the BEB and BNEB algorithms as the number of contending stations increases when the BER is fixed at 10^{-4}, 10^{-5} and 10^{-6}. When there are many contending stations, because the collision is more frequent than the transmission error, the BNEB algorithm shows better performance than the DCF. However, the backoff time strongly affects the saturation throughput of two algorithms when there are less contending stations. In the small number of contending stations and high BER, since the backoff time of the BNEB algorithm is larger than that of the DCF with the BEB algorithm, the performance of the BNEB algorithm is less than that of the DCF with the BEB algorithm. However, because most wireless devices operate where $BER \leq 10^{-5}$ [7], in most cases, the performance of the BNEB algorithm is better than that of the DCF with the BEB algorithm.



(a) varying channel BER. (b) variable number of stations.

Fig. 1. Comparisons of normalized saturation throughput between the DCF with the BEB and BNEB algorithms

4 Conclusion

In this paper, we described our evaluation of the performance of a Binary Negative-Exponential Backoff (BNEB) algorithm in the presence of transmission error by means of an analytical model and simulations of the saturation condition. With low BER, the performance of the BNEB algorithm was better than the DCF with the BEB algorithm because of effective collision resolution. Because of ineffective management of backoff time, the throughput of the BNEB algorithm was smaller than DCF with BEB algorithm with high BER. However, we can expect that the BNEB algorithm will show better performance than DCF because most wireless devices operate when $BER \leq 10^{-5}$.

References

1. IEEE standard for Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications. ISO/IEC 8802-11: (1999(E)) Aug. 1999
2. Bianchi, G.: IEEE 802.11-Saturation Throughput Analysis. IEEE Communications Letters, Vol. 2, No. 12, (1998) 318-320.
3. Bianchi, G.: Performance Analysis of the IEEE 802.11 Distributed Coordination Function. IEEE Journal on Selected Areas in Communications, Vol. 18, No. 3, (2000) 535-547.
4. Chatzimisios, P., Boucouvalas, A. C., Vitsas, V.: Performance Analysis of IEEE 802.11 DCF in Presence of Transmission Errors. Proceedings of IEEE International Conference on Communications, Vol. 7, (2004) 3854 - 3858.
5. Chatzimisios, P., Boucouvalas, A. C., Vitsas, V.: Influence of Channel BER on IEEE 802.11 DCF. Electronics Letters, Vol. 39, No. 23, (2003) 1687 - 1689.
6. Ki, H. J., Choi, S.-H., Chung, M. Y., Lee, T.-J.: Performance Evaluation of Binary Negative-Exponential Backoff Algorithm in IEEE 802.11 WLAN. MSN 2006, LNCS 4325, 294-303.
7. http://support.dell.com/support/edocs/network/p62005/en/specs.htm

A Rough Set Based Anomaly Detection Scheme Considering the Age of User Profiles

Ihn-Han Bae

School of Computer and Information Communication Eng.,
Catholic University of Daegu, GyeongSan 712-702, Korea
ihbae@cu.ac.kr

Abstract. This paper presents an efficient rough set based anomaly detection method that can effectively identify a group of especially harmful internal attackers – masqueraders in cellular mobile networks. Our scheme uses the trace data of wireless application layer by a user as feature value. Based on this, the use pattern of a mobile's user can be captured by rough sets, and the abnormal behavior of the mobile can be also detected effectively by applying a roughness membership function with the age of the user profile.

Keywords: Anomaly Detection, Rough Set, User Profile, Age.

1 Introduction

The nature of mobile computing environment makes it very vulnerable to an adversary's malicious attacks. The use of wireless links renders the network susceptible to attacks ranging from passive eavesdropping to active interfering. Unlike wired networks where an adversary must gain physical access to the network wires or pass through several lines of defenses at firewalls and gateways, attacks on a wireless network can come from all directions and target at any node [1]. In this paper, we propose the intrusion detection technique that the normal profile of each mobile is constructed by using rough sets and anomaly activities are effectively detected by using a rough membership function with the age of user profile. When an intrusion occurs, the attacker masquerading the legitimate user trends to have a different use pattern. Therefore, we can detect anomaly by comparing the use patterns. Therefore, we can detect anomaly by comparing the use pattern.

2 Related Work

The number of intrusions into computer system is growing because new automated intrusion tools are appearing every day. Although there are many authentication protocols in cellular mobile networks, security is still a very challenging task due to the open radio transmission environment and the physical vulnerability of mobile devices. Zhang, Y. proposed new model for Intrusion Detection System (IDS) and response in mobile, ad-hoc wireless networks. Sun, B. proposed a mobility-based anomaly detection scheme in cellular mobile networks. The scheme uses cell IDs traversed by a

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 558–561, 2007.
© Springer-Verlag Berlin Heidelberg 2007

user as the feature value [2]. In [3], based on fuzzy view of rough sets instead of exact rules, a soft computing approach that uses fuzzy rules for anomaly detection is proposed. Bae, I. H. proposed a rough set based anomaly detection scheme without considering the age of user profiles [4].

3 Rough Set

Let U be a finite set of objects called Universe, and R⊆U×U be an equivalence relation on U. The pair A=(U, R) is called approximation space, and equivalence classes of the relation R are called elementary sets in A.

For x∈U, let $[x]_R$ denote the equivalence class of R, containing x. For each X⊆U, X is characterized in A by a pair of sets – its lower and upper approximation in A, defined as:

$$\underline{A}X = \{x \in U \mid [x]_R \subseteq X\}$$
$$\overline{A}X = \{x \in U \mid [x]_R \cap X \neq \varnothing\}$$

The objects in $\underline{A}X$ can be with certainty classified as members of X on the basis of knowledge in R, while the objects in $\overline{A}X$ can be only classified as possible members of X on the basis of knowledge in A. The set $BN_A X = \overline{A}X - \underline{A}X$ is called the A-boundary region of X, and thus consists of those objects that we cannot decisively classify into X on the basis of knowledge in A.

Rough set can be also characterized numerically by the following coefficient called the accuracy of approximation, where *Card* denotes the cardinality [5].

$$\alpha_A(X) = \frac{Card\underline{A}X}{Card\overline{A}X}$$

4 Anomaly Detection Considering the Age of User Profiles

In this section, we introduce the construction of a rough set based anomaly detection scheme that considers the age of user profiles. For each mobile user, the features of the user activities from the wireless application layer are captured. We use the traveled cell ID, the duration of the service and the service class of a user as the feature value. The feature values are stored in the user's feature profile database.

The equivalence classes from the user's feature profile database are computed by using rough sets. For a mobile user, based on both the user activity information and the equivalence class information, a deviation number is computed by a roughness membership function with the age of the user profile, where the deviation number represents the degree that the user behavior is deviated from the normal behavior. When a user activity occurs, if the deviation number is greater than the deviation threshold that is a system parameter, an alert message occurs, otherwise the user activity identified as normal. The whole scheme is illustrated in Fig. 1.

The user feature profile database used in our scheme is shown in Table 1, where REQ#, CELL, DUR, CLASS and AGE represent the service request number, the traveled cell ID, the requested service duration, the requested service class and the age

of the data, respectively. As the age of profile data is smaller, the profile data is newer. The profile data are continuously come and faded (aging) away.

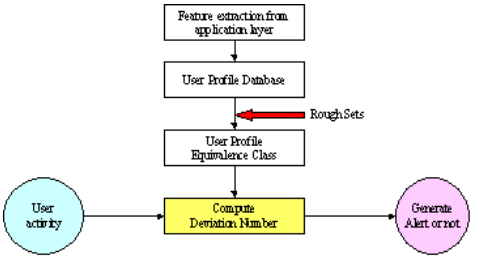


Fig. 1. The structure of rough set based anomaly detection scheme

Table 1. User feature profile information system

REQ#	CELL	DUR	CLASS	AGE
1	a	1	α	3
2	a	1	α	3
3	a	2	α	3
4	b	2	β	3
5	b	2	β	3
6	b	2	β	2
7	b	2	β	2
8	b	3	γ	2
9	c	3	γ	2
10	c	3	γ	2
11	a	1	α	1
12	a	1	α	1
13	a	2	α	1
14	b	2	β	1
15	b	3	β	1

The fuzzy inclusions are represented by the inequalities of membership functions. Further, we will allow certain errors as long as they are within the radius of tolerance. The fuzzy inclusion is computed by Eq. (1), roughness membership function.

$$\mu(X,Y) = \frac{\sum_{age=1}^{g}\left[Card((\bar{R}X \cup \bar{R}Y) - (\bar{R}X \cap \bar{R}Y)) \times (1 - w_{age})\right]}{Card(\bar{R}X \cup \bar{R}Y)} \tag{1}$$

where g represents the number of age grades, w_{age} represents the weighted value of the age, and X and Y represent condition attribute and decision attribute, respectively.

We assume that the deviation threshold (ε) is 0.58. In case that a user activity (c, 2, α) is occurred, let X={c, 2} and Y={α}, $\bar{R}X = \{3,4,5,6,8,9,10,14,15\}$ and $\bar{R}Y = \{1,2,3,11,12,13\}$, so that $\mu(X,Y) = 8.4/14 = 0.6$. Accordingly, the user activity is identified as anomalous because that $\mu(X,Y) > \varepsilon$, and an alert message is generated.

5 Performance Evaluation

We use the following two metrics: detection rate, false alarm rate to evaluate the performance of our proposed anomaly detection scheme.

We present and analyze the simulation results at different deviation threshold. We assume as follows. In cases of age 1 and age 2, if a user activity had two records those the two values of the user activity attribute values were matched with the feature values of the user profile data, the user activity is normal. Also, if a user activity had one record that two values of the user activity attribute values were matched with the feature values of age 1 within the user profile data and one or more records those two values of the user activity attributes values were matched with the feature values of age 2 or age 3 within the user profile data, the user activity is normal. All user activities of other cases are abnormal. Table 2 shows the parameter values for the simulation.

The performance of the proposed scheme is compared with that of a rough set based anomaly detection scheme without aging [4]. Simulation results of the detection rate and the false alarm rate over deviation threshold are illustrated in Fig. 2. The detection rate of our scheme (*Detection* (*age*)) is better than that of the rough set based anomaly detection scheme without aging (*Detection* (*no age*)) regardless of deviation thresholds.

Table 2. Simulation parameters

Parameter	Value
The number of user activities	1,000
The traveled cell IDs	Random (1,4)
The type of service durations	Random (1,4)
The type of service classes	Random (1,3)
The weight of ages, (w_1, w_2, w_3)	(0.6, 0.3, 0.1)

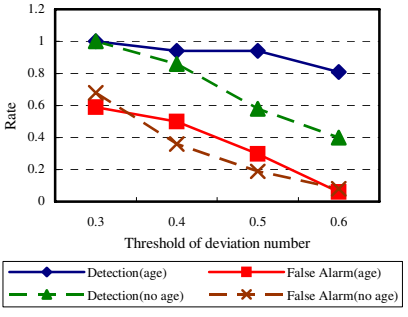


Fig. 2. Detection rate and false alarm rate at different deviation threshold

6 Conclusions

In this paper, we propose the intrusion detection technique that the normal profile of each mobile is constructed by using rough sets and anomaly activities are effectively detected by using a rough membership function with the age of user profile. From the results, we know that the performance of our scheme is better than that of the rough set based anomaly detection scheme without aging regardless of deviation thresholds.

References

1. Zhang, Y., Lee, W., Huang, Y-A.: Intrusion Detection Techniques for Mobile Wireless Networks, ACM Wireless Networks Journal, 9(5) (2003): 545-556
2. Sun, B., Yu, F., Wu, K., Leung V. C. M.: Mobility-Based Anomaly Detection in Cellular Mobile Networks, Workshop on Wireless Security, (2004): 61-69
3. Lin, T. Y.: Anomaly Detection - A Soft Computing Approach, Proceedings in the ACM SIGSAC New Security Paradigm Workshop, (1994): 44-53
4. Bae, I. H.: Design and Evaluation of a Rough Set Based Anomaly Detection Scheme for Mobile Networks, advances in Natural Computation and Data Mining, (2006): 262-268
5. Pawlak, Z.: Rough Sets: Theoretical Aspects of Reasoning about Data, Kluwer Academic Pub., (1991)

Space-Time Coded MB-OFDM UWB System with Multi-channel Estimation for Wireless Personal Area Networks

Bon-Wook Koo, Myung-Sun Baek, Jee-Hoon Kim, and Hyoung-Kyu Song

uT Communication Research Institute, Sejong University, Seoul, Korea
bebop@sdc.sejong.ac.kr, sabman@sju.ac.kr, bid4u@sdc.sejong.ac.kr, songhk@sejong.ac.kr

Abstract. In this paper, we apply multiple antennas to MB-OFDM UWB system for high performance. With an emphasis on a preamble design for multi-channel separation, we address the channel estimation in MB-OFDM system with multiple antennas. By properly designing each preamble for multiple antennas to be orthogonal in time domain, the channel estimation can be applied to the MB-OFDM proposal for IEEE 802.15.3a standard in the case of more than 2 transmit antennas. By using the multi-antenna scheme and proposed channel estimation technique, the reliability and performance of MB-OFDM system can be improved.

Keywords: MB-OFDM, CAZAC, multiple antennas, channel estimation.

1 Introduction

Ultra-wideband (UWB) technology is selected as a solution for the IEEE 802.15.3a standard [1] for low cost and high performance wireless entertainment network able to support streaming multimedia content and full motion video. In the standard of IEEE 802.15.3a, the data rate must be high enough (greater than 110 Mb/s) to satisfy a set of multimedia industry needs for wireless personal area networks (WPAN) communication. The standard also address the quality of service (QoS) capabilities required to support multimedia data types [1]. Therefore, higher rate and reliable transmission are required to satisfy the condition.

In this paper, as a solution for higher rate and reliable transmission, we apply MIMO architectures using space-time block code (STBC) to MB-OFDM UWB system. As an application of the MIMO architecture, a preamble structure for employing STBC with more than 2 transmit antennas is designed to be orthogonal in the time domain, and the channel estimation performance based on an investigated preamble structure is highlighted.

2 Space-Time Coded MB-OFDM System

In the WPAN system based on MB-OFDM, the whole available UWB spectrum between 3.1-10.6GHz is divided into 14 sub-bands with 528MHz bandwidth [2].

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 562–565, 2007.
© Springer-Verlag Berlin Heidelberg 2007

The transmission rate of the MB-OFDM is between 53.3-480Mbps. In each sub-band, a normal OFDM modulated signal with 128 subcarriers is used. The main difference between the MB-OFDM system and other narrowband OFDM systems is the way that different sub-bands are used in the transmission with several hopping patterns.

As mentioned previously, for high performance of the system, we apply multiple antennas (2 and 4 transmit antennas and 1 receive antenna) to MB-OFDM UWB system. In this paper, we use the STBC proposed by Alamouti [3].

3 Preamble Design for Multiple Antennas of MB-OFDM

In the STBC MB-OFDM system, the channel estimation is carried out for every transmit antenna of the all sub-band. For the channel estimation of each antenna, we design preambles using the Zadoff-Chu sequence, one of the constant-amplitude zero-autocorrelation (CAZAC) sequence, [4]. The sequence C_k is as follows :

$$C_k = \exp[j\frac{M\pi k^2}{N}] \tag{1}$$

where N is a length of preamble, and $k = [0, 1, ..., N-1]$. M is an integer relatively prime to N, and we consider the case of M is 1.

C_k has the property of a periodic autocorrelation function that is zero everywhere except at a single maximum per period. In the MB-OFDM system, The length of the channel estimation sequence is 128 except zero-padded suffix and guard interval. By using C_k of 64-symbol length ($N = 64$), we design the extended 128-length Zadoff-Chu sequence with zero-padding. The extended Zadoff-Chu sequence is as follows:

$$S_{8m+n+1} = \begin{cases} C_{4m+n} & for \;\; m : even \;\; including \;\; zero \\ 0 & for \;\; m : odd \end{cases} \tag{2}$$

where $m \in \{0, 1, 2, ..., 15\}$, $n \in \{0, 1, 2, ..., 7\}$. Fig. 1 shows the structure of preambles. In order to apply 4 transmit antennas, 4 preambles can be designed by cyclic shift. For 2 transmit antennas, P1 and P3 preambles are used, and for 4 transmit antennas, all preambles are used.

From [5], we deduce generalized relation to determine the number of distinguishable paths D as follows:

$$1 \leq D \leq \frac{L}{N_t} \tag{3}$$

where L indicates the symbol-length of sequence, and N_t is the number of transmit antenna.

The orthogonality of preambles is broken and the preamble is not suitable for MB-OFDM specification, when the system just uses 64-symbol length sequence. However, extended sequences will keep the property of the orthogonality at channel model (CM) 1 and 2 when the system uses 2 and 4 transmit antennas. It is noted that D is 64 for N_t=2 and is 32 for N_t=4 from (3). Using the

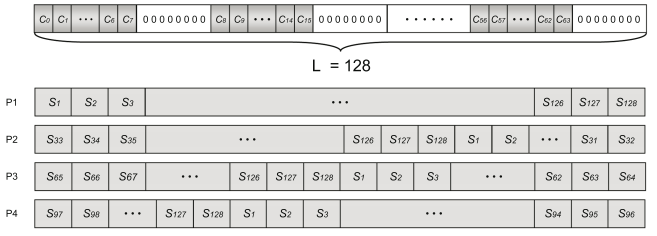


Fig. 1. The structure of each 128 length preamble for 4 transmit antennas system

orthogonality, the receiver can execute the channel estimation and separate channel impulse response (CIR) of each transmit antenna.

4 Performance Evaluation and Discussion

The performance of the proposed preamble is evaluated in terms of MSE and BER in this section. Fig. 2 (a) shows the MSE performance of 2 and 4 transmit antennas at CM 1-4 [6]. In the case of 2 and 4 transmit antennas of CM 1 and 2 transmit antennas of CM2, the system can keep the orthogonality of preambles. However, in the other cases, MSE performances are very poor because the orthogonality of preamble is broken because of the reason that mentioned through (3) in section 3.

Fig. 2 (b) shows the effect of the number of transmit antennas on the BER performance. Simulations are executed in conditions which are 1, 2 and 4 trans-

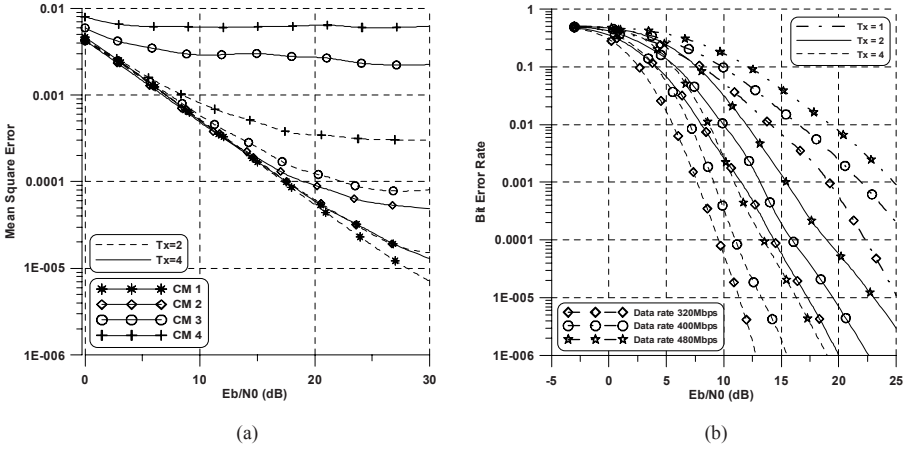


Fig. 2. (a) The MSE performance of proposed preambles applied to 2 and 4 transmit antennas with MMSE estimator. (b) The BER performance 1, 2 and 4 transmit antennas.

mit antennas at data rate 320, 400 and 480Mbps. Simulation results show that multi-antenna systems have the better performance than single antenna system and also as the number of antennas increases, the system has the better BER performance.

5 Conclusions

In this paper, we apply space-time architecture to MB-OFDM system based on WPAN for high capacity transmission and propose the new preamble structure for channel estimation which is required in MIMO architecture. Through the MSE performance, simulation results have shown that the proposed sequence can be adopted to multi-antenna MB-OFDM system. The BER performance shows that the reliability of STBC MB-OFDM system is improved efficiently by increasing the number of antennas. As the new preamble is applied, it has been shown that the MB-OFDM system with multi-antenna can achieve the high transmission capacity.

Acknowledgement

This work is financially supported by the Ministry of Education and Human Re-sources Development(MOE), the Ministry of Commerce, Industry and Energy(MOCIE) and the Ministry of Labor(MOLAB) through the fostering project of the Lab of Excellency and is supported by MIC Frontier R&D Program in KOREA.

References

1. "IEEE 802.15 WPAN high rate alternative PHY Task Group 3a(TG3a)[Online]," Available: http://www.ieee802.org/15/pub/TG3a.html
2. MultiBand OFDM Alliance(MBOA) Special Interest Group(SIG), WiMedia Alliance, Inc.(WiMedia) "MultiBand OFDM Physical Layer Specification,"Release 1.1, July 2005.
3. Siavash M. Alamouti, "A Simple Transmit Diversity Technique for Wireless Communications," *IEEE Journal on Selected Areas in Communications*, vol. 16, no. 8, pp. 1451-1458, October 1998.
4. David C. Chu, "Polyphase Codes With Good Periodic Correlation Properties," *IEEE Transactions on information theory*, vol. 18, no. 4, pp. 531-532, July 1972.
5. Dong-Jun Cho, Young-Hwan You, and Hyoung-Kyu Song, "Channel Estimation with Transmitter Diversity for High Rate WPAN Systems," *IEICE Trans. Commun.*, vol. E87-B, no. 11, Nov. 2004.
6. IEEE P802.15-02/490rl-SG2a, "Channel Modeling Sub-committee Report Final," February 2003.

Performance Enhancement of Multimedia Data Transmission by Adjusting Compression Rate*

Eung Ju Lee[1], Kyu Seol Lee[2], and Hee Yong Youn[2,**]

[1] S/W Lab, Telecomm Systems Division
Samusung Electronics, Suwon, Korea
eungju.lee@samsung.com
[2] School of Information and Communications Engineering
Sungkyunkwan University, Suwon, Korea
{rinaco,youn}@ece.skku.ac.kr

Abstract. The rapid growth of wireless communication technology has spurred various mobile applications. In this paper we propose a scheme which can improve the PSNR of multimedia data transmission using the RTP/RTCP. It is achieved by adjusting the compression rate according to the packet loss rate, delay, or jitter in the network. The NS-2 simulation reveals that the proposed approach shows a significant improvement over the conventional scheme of fixed compression rate. Particularly, it turns out that adjusting based on jitter is more effective than with the other two factors.

Keywords: Compression rate, multimedia data, PSNR, RTP/RTCP, transmission interval.

1 Introduction

Resulting from rapid growth of wireless internet, various services are available in mobile communication environment [1]. Because of the mobility of hosts, surrounding buildings, and variation of terrain, the mobile communication channel suffers from doppler shifts in the carrier due to relative motion between the terminals. It also suffers from fast spatial fading due to regions of constructive and destructive interference between the signals arriving along different propagation paths. These problems may cause variation of signal attenuation or channel fading. During a long deep fade, data packets could be severely corrupted or completely lost [2]. Consequently, there is a need for an efficient data transmission mechanism which improves the quality of services with wireless networks [3,4].

HTTP and FTP are not the protocols designed for speed but stability using TCP layer. However, the TCP has several disadvantages with respect to multimedia data transmission; high processing overhead, network transmission delay, and lack of multimedia functionalities. Because of such disadvantages, the real-time transport

* This research was supported by the Ubiquitous Autonomic Computing and Network Project, 21st Century Frontier R&D Program in Korea and the Brain Korea 21 Project in 2007.
** Corresponding author.

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 566–569, 2007.
© Springer-Verlag Berlin Heidelberg 2007

protocol (RTP) was designed operating in the UDP layer [5]. It provides end-to-end network transport functions suitable for the applications transmitting real-time data such as audio, video, or simulation data over multicast or unicast network. The RTP does not address resource reservation nor guarantee quality-of-service for real-time applications. The application transporting multimedia data by the RTP uses two ports. One is for data transportation and the other is for control. The protocol used for control is called as real-time control protocol (RTCP), showing how many packets have been transported and received.

Currently, data are transmitted to mobile handheld devices at a fixed compression rate without considering operation condition such as error rate, delay, and variation in bandwidth. Therefore, it is difficult to provide the service of high quality [6]. In this paper we propose a scheme solving the problem by adjusting the compression rate based on either packet loss rate, delay, or jitter with the RTP/RTCP. It significantly improves the PSNR of the transmission compared to the conventional scheme. NS-2 simulation verifies the efficiency of the proposed approach. Particularly, it turns out that adjustment based on jitter is more effective than the other two factors.

2 The Proposed Scheme

There exist various obstacles interfering with data transmission in wireless network environment. Therefore, an efficient transmission mechanism which can dynamically cope with harsh environment is needed to provide high quality service. For this we develop a scheme which transmits multimedia data using the RTP/RTCP, while properly adjusting the compression rate according to the network condition. Figure 1 shows the flow of data in the proposed scheme.

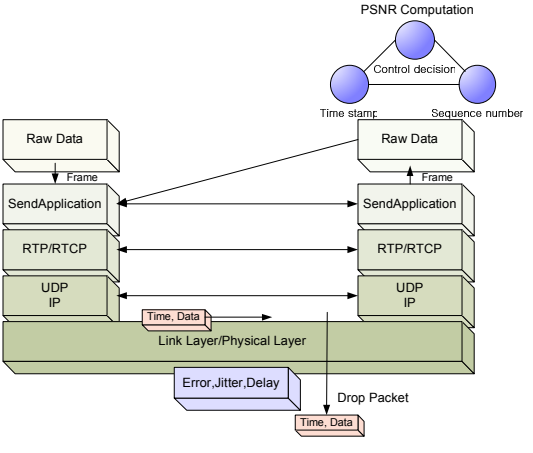


Fig. 1. The flow of data in the proposed scheme

For data transmission, the data compression rate or transmission interval can be adjusted. By checking the sequence number and timestamp of the packets arriving at the receiver, packet loss, delay, and jitter can be identified. The compression rate is then

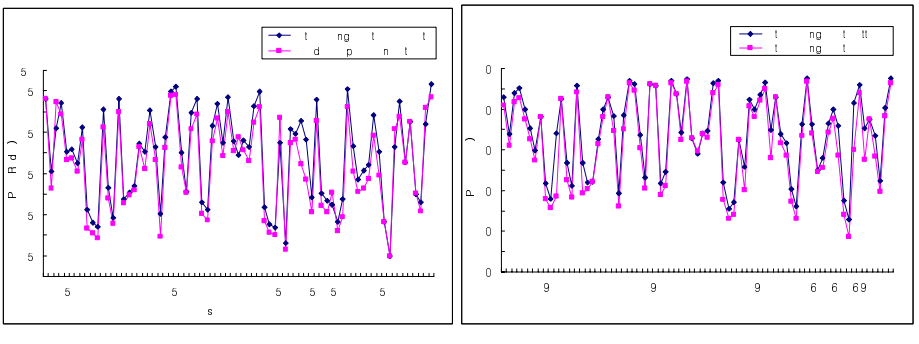
decided accordingly. When the packet loss rate, delay, and jitter are high, the receiver orders the sender to raise the data compression rate. The quality of motion picture can be improved by adopting this approach. Also, packet loss can be prevented by increasing the transmission interval. Through proper adjustment of the operation parameters according to the network condition, users can receive better services. Packet loss will increase if the bandwidth is not enough. Then the sender needs to reduce the amount of data transmitted at a time and raise the compression rate by increasing the quantization level of the transcoder. Table 1 summarizes the adjustment of compression rate and transmission interval according to the network condition represented by packet loss, delay, and jitter.

Table 1. Adjustment of compression rate (CR) and transmission interval (TI)

Packet loss \ delay, jitter	No change	Increase	Decrease
No change	No adjustment	TI++	TI--
Increase	CR--	CR --/ TI ++	CR --/ TI --
Decrease	CR++	CR ++/TI ++	CR ++/ TI --

For simulation we choose an about 100Mbyte motion picture with frequent movement of receiver node to recognize picture collapse and frame loss. We implement and evaluate the proposed scheme using ns-2. We adopt the link error model given by ns-2 and assume that the channel error rate is 10%.

The operation flow is as follows. The raw data generated in the application layer are encoded to exclude the influence of PSNR of the compression. A timestamp and sequence number are inserted in the encoded data in the RTP/RTCP layer. The data are encapsulated by the UDP header for real-time transmission, and then sent through the network. The data received are passed through the UDP layer up to the RTP/RTCP layer and then the application layer. It orders the sender to adjust the transmission interval and compression rate using the timestamp and sequence number obtained from the RTP by sending an RTCP packet every 0.1 second.



(a) Error rate (b) Jitter and delay

Fig. 2. The comparison of the PSNR values with changed compression rate

The simulation measures the average PSNR values while varying the compression rate as the error rate, delay, and jitter value change. Note that varying transmission interval may not be suitable for real-time data transmission. Therefore, we consider varying only the compression rate.

Figure 2 shows the simulation results for each of the three cases of compression rate adjustment according to error rate, delay, and jitter, respectively. Observe from Figure 2(a) that the PSNR with changing compression rate based on error rate is statistically higher than with a fixed rate. Figure 2(b) shows that changing the rate based on jitter allows higher PSNR compared with the delay-based change. The average PSNR of the fixed compression rate, changing compression rate according to the error rate, delay, and the jitter are 26.7dB, 29dB, 31dB, and 34.4dB, respectively. The result indicates that adjusting compression rate based on jitter allows maximum throughput.

3 Conclusion

In this paper we have proposed a scheme for multimedia data transmission that effectively copes with unstable environment causing packet loss, delay, and variation in the bandwidth by adjusting the compression rate and transmission interval. The simulation using ns-2 reveals that the proposed approach significantly improves the efficiency of transmission of motion picture compared with the conventional data transmission employing a uniform compression rate. Especially, adjusting the compression rate based on jitter is found to be more effective than error rate or delay. Varying transmission interval may not be suitable for real-time data transmission.

Future work will include the investigation of controlling transmission interval along with data compression rate for achieving a better performance. We will also consider the channel error rate, delay, and variation in bandwidth not separately but in a unified way for better services in wireless mobile network environment.

References

1. Youn, J., Xin, J., Sun, M.-T.: Fast video transcoding architectures for networked multimedia applications. in Proc. of IEEE Int'l Symp. on Circuits and Systems, Vol. 4. (2000) 25-28
2. Zhang, D., Shijagurumayum, S.: Personalized Content Delivery to Mobile Devices. IEEE International Conference on Systems, Man & Cybernetics, (2003) 2533 -2538
3. Liu, S., Bovik, A.C.: A fast and memory efficient video transcoder for low bit rate wireless communi-cations. IEEE Int'l Conf. on Acoustics, Speech, and Signal Processing, Vol. 2. (2002) 1969-1972
4. Wang, L., Luthra, A., Eifrig, B.: Adaptive rate control for MPEG transcoder. in Proc. of Int'l Conf. on Image Processing, Vol. 4. (1999) 266-270
5. Schulzrinne, H.: RTP: A transport Protocol for Real-Time Application. RFC 1889, (1996)
6. Kasai, H., et. al.: Rate control scheme for low-delay MPEG-2 video transcoder. in Proc. of Int'l Conf. on Image Processing, Vol. 1. (2000) 964-967

A Feasible Approach to Assigning System Components to Hybrid Task Sets in Real-Time Sensor Networking Platforms

Kyunghoon Jung[1], Byounghoon Kim[1], Changsoo Kim[2], and Sungwoo Tak[1,*]

[1]School of Computer Science and Engineering, Pusan National University, San-30, Jangjeon-dong, Geumjeong-gu, Busan, 609-735, Republic of Korea
swtak@pusan.ac.kr
[2]Pukyong National University, Dept. of Computer Science, Busan, Republic of Korea
cskim@pknu.ac.kr

Abstract. This paper studies an efficient periodic and aperiodic task decomposition technique for real-time sensor networking platforms from two perspectives. First, wireless sensor networking platforms available in literature have not addressed the concept of real-time scheduling for hybrid task sets where both types of periodic and aperiodic tasks exist. Second, individual system components in real-time sensor networking platforms must be assigned to either a periodic or aperiodic task in order to provide the concept of fully optimal real-time scheduling given real-time requirements of user applications. The proposed approach guarantees the feasible scheduling of hybrid task sets while meeting deadlines of all the periodic tasks. A case study based on real system experiments is conducted to illustrate the application and efficiency of the proposed technique. It shows that our method yields efficient performance in terms of guaranteeing the deadlines of all the periodic tasks and aiming to provide aperiodic tasks with good average response time.

Keywords: Networking Platform, Communication Software Architecture, Real-Time Scheduling.

1 Introduction

In order to realize real-time wireless sensor networking platforms in many different kinds of critical applications, three fundamental software mechanisms for real-time processing have to be studied. First is the principal division of tasks and functionalities in a sensor node, such as kernel architecture and communication protocol stacks. Second, the real-time sensor software architecture on a single node has to be extended to a network architecture, where the division of tasks between systems, not only on a single node, becomes the relevant question – for example, how to structure periodic or aperiodic tasks for communication protocol components. The last is the question of how to design periodic and aperiodic task decomposition technique which yields efficient performance in terms of guaranteeing the deadlines of all the periodic tasks and

* Corresponding author.

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 570–573, 2007.
© Springer-Verlag Berlin Heidelberg 2007

aiming to provide aperiodic tasks with good average response time. Existing senor networking platforms, such as TinyOS [1] and DCOS [2], have been focused on solving the efficient exploitation of limited resources. Envisioned application scenarios applied in sensor networking platforms need to activate real-time sensing as well as provide good average response time with task communication. In this paper, we propose an efficient periodic and aperiodic task decomposition technique for real-time sensor networking platforms from two perspectives. First, wireless sensor networking platforms available in literature have not addressed the concept of real-time scheduling for periodic and aperiodic tasks. Second, individual system components in real-time sensor networking platforms must be assigned to either a periodic or aperiodic task in order to provide the concept of fully optimal real-time scheduling given real-time requirements of user applications.

2 Real-Time Sensor Networking Platform

Fig. 1 shows hybrid task sets consisting of periodic and aperiodic tasks in the proposed real-time sensor networking platforms. In User tasks, sensing tasks which activate sensors and transmit/receive the sensing data periodically are included in a set of periodic tasks. Application tasks perform application specific operations and may be either periodic or aperiodic according to their own specific operations. The network protocol is classified into IP, TCP, UDP, ARP, and ARP Update tasks. IP, TCP, UDP and ARP tasks are included in a set of aperiodic tasks because of not having the period of their execution. The TCP task is divided into two subtasks, the TCP-IN task for transmitting packets and the TCP-OUT task for receiving packets. Most characteristics of the ARP task are aperiodic except updating the ARP table every 20 minutes. The ARP Update task updates the ARP table as a periodic task.

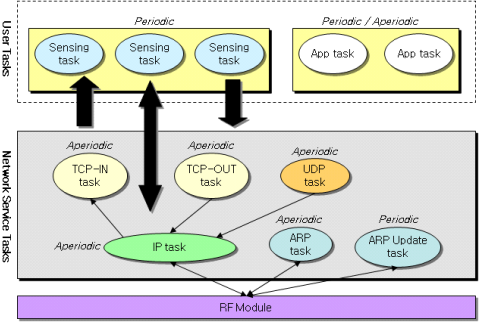


Fig. 1. Architecture of real-time sensor networking platforms

The proposed technique consists of three primary components, a set of periodic and aperiodic tasks, and PATS (periodic and aperiodic tasks scheduler). The periodic task set component contains Sensing, Application and ARP Update tasks. The PATS component is composed of four sub-modules which are SC (Schedulability Checker), STMAT (Slack Time Manager), ATS (Aperiodic Task Scheduler), and PTS (Periodic

Task Scheduler) modules. The SC module decides if a given periodic task set can be schedulable. The STMAT evaluates the processor idle time called slack time. It consists of two sub-modules, STCAT (Slack Time Calculator) and STAAT (Slack Time Allocation). The STCAT computes the slack time and stores it in the STT (Slack Time Table) and the STAAT assigns the Slack time to the aperiodic task chosen by the ATS. The ATS schedules the aperiodic tasks, allocated by the STMAT, in the FIFO policy. The PTS schedules a set of periodic and aperiodic tasks according to the priority-driven policy. We make full use of the RM (Rate Monotonic) algorithm for periodic tasks and the Slack Stealing algorithm for aperiodic tasks [3-4].

Schedulability Checker

We use the schedulability test addressed in [3]. The basic terms are as follows. C_i is the execution time of the periodic task τ_i, and T_i is the period of τ_i. Then a workload of τ_i, $W_i(t)$, is defined as follow

$$W_i(t) = \sum_{j=1}^{i} C_j \cdot \lceil t/T_j \rceil, \; t = \{k \cdot T_j \mid j = 1,...,i; k = 1,...,\lfloor T_i / T_j \rfloor\}. \tag{1}$$

Let $L_i(t)$ be the processor utilization of τ_i. By (1), $L_i(t) = W_i(t)/t$. Let L_i be the actual processor utilization of τ_i and L be the total processor utilization of task sets. Then L can be computed as follow

$$L = \max_{\{1 \le i \le n\}} L_i = \max_{\{1 \le i \le n\}} \min_{\{0 < i \le T_i\}} L_i(t) = \max_{\{1 \le i \le n\}} \min_{\{0 < i \le T_i\}} \sum_{j=1}^{i} C_j \cdot \lceil t/T_j \rceil / t \tag{2}$$

If $L \le 1$, then the task set is decided to be schedulable.

Slack Time Manager

We use the method of [4] for computing and allocating the slack time. Let $\alpha_i(t)$ be the total execution time of the aperiodic tasks which have the priorities higher or equal to the priority i. Let $P_i(t)$ be the ready time of the periodic tasks in [0, t] and $I_i(t)$ be the processor idle time or the execution time of the tasks of which priorities are lower than the priority i. Let A_{ij} be the maximum of the slack times in $[0, C_{ij}]$ where C_{ij} is the completion time of τ_{ij}. Then, the total processor time $W_i(t)$ is computed as follows

$$W_i(t) = \alpha_i(t) + P_i(t) + I_i(t) = \min_{\{0 \le t \le D_{ij}\}} \{(A_{ij} + P_i(t))/t\} = 1 \text{ for } 0 \le t \le D_{ij} \tag{3}$$

$$A^*(t) = min_{\{1 \le i \le n\}} A_i(t), \; C_{ij-1} \le t < C_{ij} \text{ for } A_i(t) = A_{ij} \text{ and } C_{ij-1} \quad t < C_{ij}, \; j \quad 1 \tag{4}$$

$A^*(t)$ represents the maximum of the slack times.

Aperiodic and Periodic Task Scheduler

The ATS can schedule the aperiodic task only when the value in the STT is only positive. The PTS schedules the periodic and aperiodic tasks by the priority-driven policy. The priority of the periodic task is assigned by the RM and the aperiodic task is given by the system designer. PTS would select and schedule the task of the highest priority of priority tables for periodic and aperiodic tasks.

3 Experiments and Conclusion

All of the experiments in this paper were carried out using the CC2420DB wireless sensor mote. The CC2420DB hardware contains an IEEE802.15.4 compliant RF transceiver CC2420, an Atmega128L microcontroller, one internal sensor (a temperature sensor) and two external sensors (light and humidity sensors) we installed. The software platform incorporating the proposed task decomposition technique is implemented in this paper by modifying and improving uC/OS-II v2.52 [5] and uIP v0.9 [6]. For measuring the performance of our method, three experimental models, Case 1, 2 and 3, are evaluated. The Case 1 model is the original uC/OS-II incorporating the original uIP. The Case 2 model is the original uC/OS-II incorporating the TCP/IP software communications architecture which is decomposed into task-per-protocol. We develop the task-based TCP/IP software communications architecture by exploiting the uIP v0.9. The Case 3 model is the modified uC/OS-II incorporating the proposed task decomposition technique containing the task-based TCP/IP software communications architecture. We evaluate the performance of Case 1, 2, and 3 respectively by executing each one of them on two CC2420DB motes communicating with each other. The unique IP address is assigned to each of two CC2420DB motes. A case study based on real system experiments is conducted to illustrate the application and efficiency of the proposed technique. It shows that our periodic and aperiodic task decomposition technique yields efficient performance in terms of guaranteeing the deadlines of all the periodic tasks and aiming to provide aperiodic tasks with good average response time. Our research gives valuable insight into how to deploy real-time sensor networking platforms that has an unique software architecture with task-based components.

Acknowledgement

This work was supported by the Korea Research Foundation Grant funded by the Korean Government (MOEHRD) (The Regional Research Universities Program/Research Center for Logistics Information Technology).

References

1. Subramonian V., Huang-Ming H., Datar S., Chenyang L.: Priority scheduling in TinyOS - a case study. Washington University Technical Report (WUCSE-2003-74), December (2003)
2. Hofmeijer T.J., et. al.: DCOS, A real-time light-weight data centric operating system. The IASTED International Conference on Advances in Computer Science and Technology(ACST 2004), St. Thomas, US Virgin Islands, November (2004)
3. Lehoczky J., Sha L., Ding Y.: The rate monotonic scheduling algorithm: exact characterization and average case behavior. Real Time Systems Symposium, December (1989), 166-171
4. Lehoczky J.P., Ramos-Thuel S.: An optimal algorithm for scheduling soft-aperiodic tasks in fixed-priority preemptive systems. Proc. of 13th Real-Time Systems Symposium, December (1992), 110-123
5. Labrosse J.J.: uC/OS-II the real-time kernel. CMP Books, Kansas, (2002)
6. Dunkels A.: Full TCP/IP for 8Bit architectures. In Proc. of the First ACM/Usenix International Conference on Mobile Systems, Applications and Services (MobiSys 2003), San Francisco, USA, May (2003), 85-98

Efficient Routing Scheme Using Pivot Node in Wireless Sensor Networks

Jung-Seok Lee[1], Jung-Pil Ryu[2], and Ki-Jun Han[1,*]

[1] Department of Computer Engineering, Kyungpook National University, Daegu, Korea
leejs@netopia.knu.ac.kr, kjhan@knu.ac.kr
[2] Center for Technology fusion in Construct, Yonsei University, Seoul, Korea
goldmunt@yonsei.ac.kr

Abstract. We propose an efficient routing for wireless sensor networks, called *pivot routing*. Pivot routing establishes the data paths while query is propagated. Source node selects a data path and transmits data packets to the sink via pivot nodes. Because subsequent query and data packets are transmitted along pivot nodes, control message overhead decreases considerably. Network lifetime also increases since the energy consumption of communication is reduced at sensor node. Simulation results show good aspects of control overhead and network lifetime.

Keywords: Pivot routing, Pivot shifting, Energy efficient routing, Wireless sensor networks.

1 Introduction

Wireless sensor networks have been receiving a great deal of attention to be an infrastructure of ubiquitous computing environments. Many of routing protocols in wireless sensor networks have been introduced in the technical literature. In query based routing protocols, sink node floods query to networks and nodes search and construct routing path to delivery sensing data in response to the query of sink node. Finally, sink node sends query messages to source nodes after selecting path in the multiple paths and source nodes can start data delivery [1, 2, 3].

We propose an efficient routing scheme which extends the network lifetime, called *pivot routing*. In pivot routing, since multiple data paths can be established when a query is flooded, control message overheads can be reduced. These benefits can prolong the network lifetime.

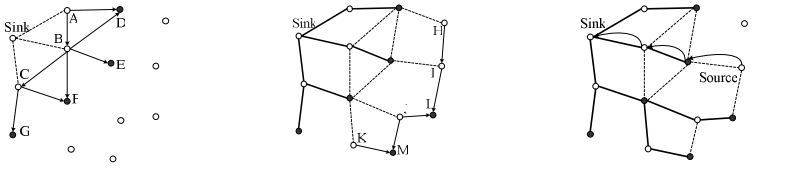
2 Pivot Routing in Wireless Sensor Networks

We propose a pivot routing in this section. In our scheme, since data paths are created between source nodes and sink while queries are propagated in the whole sensor networks, control overhead and total delay can be reduced as compared with the previous routing protocols.

* Correspondent author.

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 574–577, 2007.
© Springer-Verlag Berlin Heidelberg 2007

We use the term *pivot node* to denote the sensors which construct the data paths. Once queries are propagated to sensor networks, selected pivot nodes construct the data paths toward the sink then sources just transmit the sensing information to a closest pivot node. We also have to consider the criteria to choose the pivot nodes. Several metrics can be candidates for selection, such as the hop counts from sink and the previous pivot node, the residual power, etc. In this paper, we only use the hop counts to select pivot nodes. We assume that the sensor nodes are not aware of their location, and already know the list of adjacent sensor nodes by HELLO exchange.



(a) Pivot selection (b) Query complete (c) Next forwarding/query drop

Fig. 1. The construction of data paths when pivot selection hop is 2

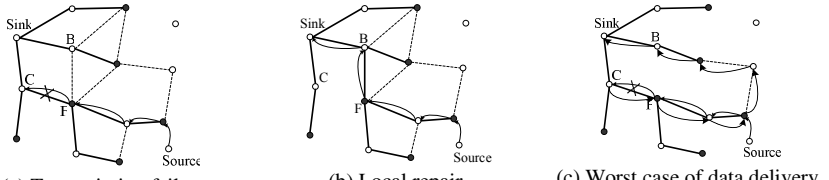
Initially, a sink sends queries for gathering the sensing information. A query message includes some mandatory information; *{query ID, pivot selection hop, prior pivot node ID, traverse node list}*. When a sensor node receives a query message, sensor nodes which being at distance of pivot selection hop are chosen as pivot nodes. Pivot selection hop (>1) is assigned by a sink as a network parameter and included in query message. Whenever a query message is forwarded at a sensor, a sensor appends its own node id to the end of traverse node list in the query regardless of pivot selection. A pivot node maintains the node list table, named *pivot table*. Pivot table is composed of the traverse node list. Pivot nodes construct implicitly downstream path of the data to sink using primary traverse node list in the pivot table. A pivot node will forward the data to a previous pivot node with reference to pivot table. In order to transmit sensing data, previous pivot node id is stored at the all sensors which receive queries. Sensors between a sink and a pivot node would store the sink id.

In pivot routing, once a sensor is assigned as a pivot node, it always acts as a pivot node until it required that a pivot node is changed due to the specific condition, such as insufficient residual power or sensor failure. In order to prevent infinite flooding, each sensor node does not forward query which is already received.

Fig. 1 shows an example about the construction of data paths. Fig. 1(a) shows that a sink initially propagates queries to the sensor field. Even though B and C satisfy the pivot selection condition, these nodes can not be chosen as pivot node because these nodes are directly connected with sink node. The multiple paths to sink are very useful to overcome the transmission failure. When the query forwarding is completed, the data paths to the sink are accomplished, shown as fig. 1(b, c). In fig. 1(c), a source sends the sensing information immediately through pre-determined paths.

In pivot routing, the backup entries in the pivot table can be used to recover the link failure. Assume that link failure is occurred between node F and C, as shown fig. 2. F then looks up its own pivot table whether backup paths exist. F has a new data path to sink via node B. Since a link failure is detected at a pivot node F, F finds

easily the detour data path. It is different with link failure which is occurred at an intermediate node. In this instance, previous pivot id in the data is useful. An intermediate node sends *link_failure* message to the upstream pivot node. Upstream pivot node then selects an alternative path after looking up its pivot table. If no alternative path is found, a *link_failure* message is sent again to the upstream pivot node. Because pivot routing can maintain the multiple paths, it has little possibility that the pivot node closest to the source receives *link_failure*. In worst case, source node sends data to sink again as fig. 2(c), but such case is not likely to occur, because sensor nodes are distributed with high density.



(a) Transmission failure (b) Local repair (c) Worst case of data delivery

Fig. 2. Recovery from transmission failure

3 Performance Evaluation

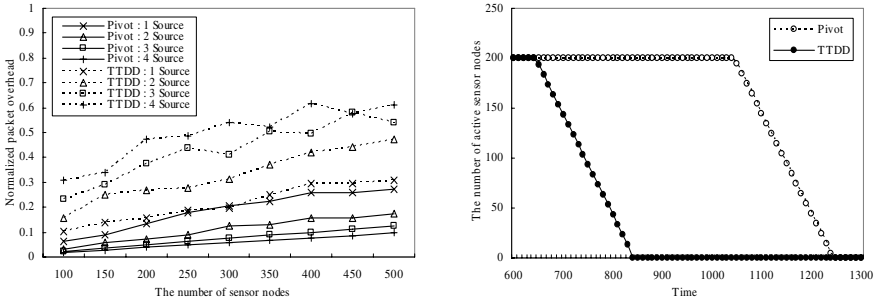
We evaluate pivot routing scheme through simulation. Simulation results are compared with TTDD in stationary sink situation. The size of sensor networks is $2000\times2000(m^2)$. The number of sensor nodes is increased from 100 to 500 nodes random distribution manner. Source nodes are randomly chosen, and the radius of radio transmission range is given 300(m). A source generates a data packet per second. Each simulation runs for 200 seconds, each result is averaged over several repetitions. Sink broadcasts a query only once after the sensor network initialized. In our simulations, we assume that there is no unidirectional links and each nodes know its own neighbors using HELLO exchange.

We use two performance metrics to validate a proposed pivot routing. The *normalized packet overhead* is the ratio of the number of control packet to the total number of packets in whole sensor networks. This metric is directly concerned with overall network lifetime. The *network lifetime* means that how many nodes are alive.

Fig. 3 shows the simulation results. Normalized packet overhead indicates how many packets are traveling in sensor networks except data packets.

In pivot routing, a sink node forwards only one query packet. Although the number of source node increases, control message overhead decreases more and more, because the number of query message does not increases. Because the number of DAM message increases in proportion to the increment of the source nodes in TTDD, control message overhead rapidly increases in the whole sensing networks.

Fig. 3(b) shows that difference of control packets influence on the overall network lifetime because of the energy consumption for exchanging the control message with each other sensor nodes.



(a) Normalized packet overhead (b) Network lifetime

Fig. 3. Simulation results

Because the number of query message from sinks is same even though different pivot selection hop, the control message overhead keeps same pattern. Data packets almost traverse the same path, nothing but the hop distance among pivot nodes is just longer. Although pivot selection hop is varied, it does not affect on the number of control message as well as network lifetime.

4 Conclusions and Future Works

In this paper, we propose pivot routing algorithm to reduce the control message overhead and prolong the network lifetime. In our scheme, pivot nodes are determined when queries go through the whole networks just one time. Pivot nodes play the role of designated router and multiple paths are created between pivot nodes. Simulation results show that pivot routing provides lower control overhead and prolongs the network lifetime than TTDD.

We are going to consider the sink mobility and use other parameters to determine the pivot nodes (residual power, neighbor density and so on) in the future.

Reference

1. J. N. Al-Karaki and A. E. Kamal, “Routing techniques in wireless sensor networks: a survey,” *IEEE Wireless Communications*, vol. 11, pp. 6–28, 2004.
2. K. Akkaya and M. Younis, “A survey on routing protocols for wireless sensor networks,” *Ad Hoc Networks*, Vol.3, Issue 3, pp. 325-349, May 2005.
3. C. Intanagonwiwat, R. Govindan, D. Estrin, J. Heidemann and F. Silva, “Directed diffusion for wireless sensor networking,” *IEEE/ACM Transactions on Networking*, Vol.11, Issue 1, pp.2-16, Feb.2003.
4. H. Luo, F. Ye, J. Cheng, S. Lu and Lixia Zhang, “TTDD: Two-Tier Data Dissemination in Large-Scale Wireless Sensor Networks,” *Wireless Networks*, Vol.11, pp.161-175, 2005.

Encoding-Based Tamper-Resistant Algorithm for Mobile Device Security

Seok Min Yoon, Seung Wook Lee, Hong Moon Wang, and Jong Tae Kim

School of Information and Communication Engineering,
Sungkyunkwan University, Korea
jtkim@skku.ac.kr

Abstract. Due to the advance in development technology for mobile systems, the attacks on the embedded systems become sophiscated. Especially, tampering with the infomation of mobile devices by software or hardware attack may lead to a serious problem like leaking personal information. In order to protect the tampering attack, encryption technique for embedded information is proposed. However, when the specification of encryption algorithm is known, the system is easily tampered. In this paper, we propose a novel tamper-resistant algorithm by encoding program instructions, which can detect the tampering attack. In this algorithm, when a malicious attacker try to tamper with a part of the system, the logical interdepency of program instrutions make it impossible unless he gets the access authorization of whole system.

Keywords: Tamper-resistant, Security, Mobile Device.

1 Introduction

While technological advances that have improved the development of embedded systems bring the conveniences in human life, it can lead to serious problems: leaking out personal information which stored in a handheld mobile embedded system such as PDAs and smart phones or operating unattempted actions by malicious attacker[1]. The attacks on the embedded systems can be classified in two categories: software attack which exploits implementation flaws by attacking vulnerability of software and hardware attack which exploit the internal information by physical modification or external devices[1]. The hardware attacks which are frequently used on the embedded system are bus tapping and data alteration between processor and external devices. The bus tapping and data alteration, mixed with software vulnerability, may give more ways of breaking a system to a malcious attacker[2][3]. Since the information which saved in a system can be exposed directly by the bus tapping, encryption of internal information techniques are proposed to protect the system[4]. However, Huang proved that the security techniques are neutralized by data which is acquired by tapping system bus exposed externally with simple hardware attachment in case the secret key used in encryption process is known by attacker[5].

In this paper, we propose a tamper-resistant algorithm based on the encoding process of program instructions. The mobile device with the encoding algorithm achieves high security, because it detects both software and hardware attacks effectively by

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 578–581, 2007.
© Springer-Verlag Berlin Heidelberg 2007

checking the interdependencies of instructions[6]. Especially, it can prevent abnormal changes of control routine and execution of malicious program code.

2 Tampering Detection Algorithm

A properly working program means that it works as the way intended by developers, that is, the program which does not work as the way intended by developers is not working properly. It means that every program have an execution sequence of instructions which is intended by developers. An instruction executing currently by a processor has interdependencies with a previously executed instruction, and also with a next instruction. If there are abnormal instructions which are not intended by developers in the sequence of execution, interdependencies of instructions are broken. We can, therefore, detect the software and hardware attacks on our system effectively if it is possible to check the integrity of current instruction by examining interdependencies among previous instruction, current instruction and next instruction.

We derive the interdependencies of the instructions by encode current instruction with previous instruction and next encoded code as shown in figure 1. The encoding and decoding operation of current instruction are defined as below.

$$C_t = E_k(I_t \text{ xor } I_{t-1} \text{ xor } C_{t+1}) \quad (1)$$

$$I_t = D_k(C_t) \text{ xor } I_{t-1} \text{ xor } C_{t+1} \quad (2)$$

E is an encoding function and D is a decoding function which use cryptography algorithm(ex. DES, AES) with secret key k. The previous instruction and next encoded code are necessary in decoding process, so a processor needs two registers for them.

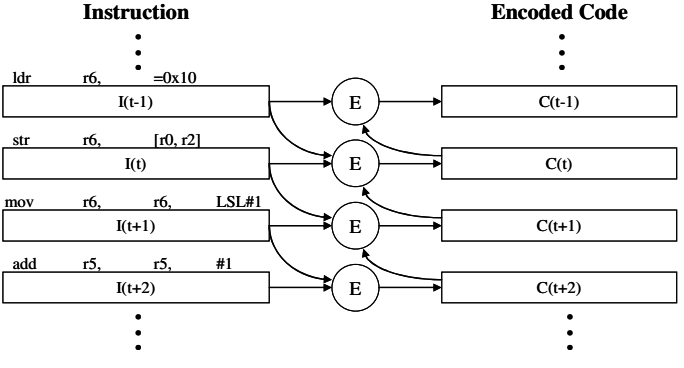


Fig. 1. Encoding process of instructions

3 Handling of Broken Interdependency

Basic concept of our algorithm is continuous checking process with the interdependencies of instructions. However, we cannot conclude that every point of broken interdependency is not a point having security problems. In ordinary program, the

interdependency can be broken by the nature of program itself. The interdependency with a previous instruction cannot be assured at the starting point of a program and the interdependency with a next instruction cannot be assured at the end point of a program. We solve this problem by using magic numbers. Figure 2 shows a simplified program containing m number of instructions. By setting M_1 as a magic number that only the developers of security process know, the encoding process can be started at the starting point of the program. Similarly, by setting M_2 as a parity of whole program, encoding process can be done at the end point.

Instruction	Encoded Code
M_1	
I_1	$C_1=E_k(I_1 \text{ xor } M_1 \text{ xor } C_2)$
I_2	$C_2=E_k(I_2 \text{ xor } I_1 \text{ xor } C_3)$
I_3	$C_3=E_k(I_3 \text{ xor } I_2 \text{ xor } C_4)$
I_4	$C_4=E_k(I_4 \text{ xor } I_3 \text{ xor } C_5)$
…	…
I_m	$C_m=E_k(I_m \text{ xor } I_{m-1} \text{ xor } M_2)$
M_2	

Fig. 2. Simplified program

```
0x2000000  mov      r5,     #0
0x2000004  ldr      r6,     =0x10
0x2000008  nop
0x200000c  led_loop:
0x200000c  str      r6,     [r0, r2]
0x2000010  mov      r6,     r6,     LSL#1
0x2000014  add      r5,     r5,     #1
0x2000018  cmp      r5,     #loopnum
0x200001c  bne      led_loop
0x2000020  ldr      r0,     =0x001f
```

Fig. 3. Insertion of NOP operation

Branch instruction can make a broken interdependency. To prevent this problem, we insert a NOP(No OPeration) instruction at the branch address shown in figure 3. Whenever a branch instruction is executed, NOP instruction is executed. In general, as the NOP instruction does not have operands, we can save parity as operands for additional comparison process.

4 Security Analysis

The encoding and decoding functions use cryptography algorithms. Generally, the key length is very important, because the safety of cryptography algorithms relies on the length of secret key. We assume that the system has 32 bits instruction length, and encoding and decoding operation use 32 bits secret key and the length of encoded code is 32 bits. The 32 bits secret key is fragile on safety because of small key space of 2^{32}. In order to make up for the weak point, we rearrange the bit composition of encoding input. This has the same effect as increasing the number of cases. Then, the attacker should know the encoding algorithm, the secret key and the bit composition of data used during encoding process. If the attacker know the encoding algorithm, the probability, which the attacker successes on decode, is $3\times32!\times2^{32}\approx2^{151}$ because the attacker should know the secret key, 2^{32}, and the bit composition of each input, $32!\times32!\times32!$. In other words, the encoding algorithm has the same effect as using 151 bits secret key.

The basis of our algorithm is logical clearance of the interdependency, not the computational complexity of encoding algorithm. In order to execute current instruction, the integrity of previous instruction, and next instruction should be assured, and

the interdependency should be guaranteed logically. As we can see in figure 1 and 2, the encoded code I(t) has the interdependencies with previous and next instruction until the end of the program. This means that the interdependencies of whole program are affected by each instruction. In other words, if one instruction has a problem, this problem destroys the interdependencies of whole program.

5 Conclusion

In this paper, we proposed an encoding-based tamper-resistant algorithm by checking the interdependency of program instructions as a countermeasure against the malicious attacks on embedded system. The proposed algorithm prevents the system from executing abnormal instructions, when the system is tampered by software or hardware attacks. The encoding algorithm of instruction has the same effect as an cryptography algorithm using 151 bits length of secret key on the system having 32 bits length of instruction. In addition, the logical clearance of interdependency of program instructions makes the attacker impossible to tamper the system unless the attacker gets the access authorization of whole system.

Acknowledgments. This work was supported by the Korea Research Foundation Grant funded by the Korean Government (MOEHRD, Basic Research Promotion Fund) (KRF-2006-521-D00376).

References

1. S. Ravi, A. Raghunathan, P. Kocher and S. Hattangady, “Security in Embedded Systems: Design Challenges”, ACM Transactions on Embedded Computing Systems: Special Issue on Embedded Systems and Security (Guest Editors: D. Serpanos and H.Lekatsas), 2004.
2. X. Zhuang, T. Zhang, and S. Pande. Hide “An infrastructure for efficiently protecting information leakage on the address bus”, In Proceedings of the ASPLOS-XI, Oct. 2004.
3. G.E. Suh, D. Clarke, B. Gassend, M.v. Dijk, S. Devadas, “aegis: Architecture for Tamper-Evident and Tamper-Resistant Processing”, In Proceedings of the 17th Int. Conference on Supercomputing, Jun. 2003.
4. G.E. Suh, D. Clarke, B. Gassend, M.v. Dijk, S. Devadas, “Efficient Memory Integrity Verification and Encryption for Secure Processors”, Proc. Annual IEEE/ACM Int. Symposium on Microarchitecture (MICRO), pp. 339-350, Dec. 2003.
5. Andrew Huang, “Keeping Secrets in Hardware: The Microsoft XboxTM Case Study”, pp. 213 - 227, May, 2002.
6. S. W. Lee and J. T Kim, “Tampering Detection Technique in Instruction Level using Error Detection Code”, *Lecture Series on Computer and Computational Sciences*, Nov. 2005..

Adaptive Vertical Handoff Management Architecture

Faraz Idris Khan and Eui Nam Huh*

Internet Computing and Security Lab
Department of Computer Engineering, Kyung Hee University, 449-701 Suwon, South Korea
{faraz,johnhuh}@khu.ac.kr

Abstract. The 4G mobile system is a collection of radio networks providing access to IP based services. It ensures seamless roaming and users are always connected to the best network providing QoS (Quality of Service) to the end user. The mechanism of switching to a different network is termed as *vertical handoff*. In this paper we propose a vertical handoff management architecture which ensures efficient utilization of a mobile terminal resource i.e. CPU. This is achieved by employing feedback mechanism which allocates CPU resource according to the network load by adapting the CPU scheduler. In the end the simulation results are presented which are in accord with our idea.

Keywords: feedback scheduling, CPU scheduler adaptation, quality of service, resource management, vertical handoff.

1 Introduction

The interest in 4G networks is increasing at a rapid pace as wireless networks and mobile communications are growing at an astonishing rate. It promises to provide a broader range, lower access costs, the convenience of using a single "all in one" device. This is achieved by overlaying two or more networks with differing characteristics. One of the advantages of such architecture is that it enables applications to maintain QoS (Quality of Service) by switching to a better network. There are two kinds of vertical handoff situation that might occur, low data rate to high data rate or vice versa. The application consumption rate should be adaptively adjusted proportional to the arrival rate of the packets by changing the CPU allocation. In order to provide undisrupted service to the end user there are various vertical handoff management architecture proposed in literature [1] which mostly considers content adaptation after vertical handoff. There is relatively no architecture proposed which consider the impact of vertical handoff on CPU usage. Section 2 discusses briefly the system architecture with our proposed Adaptive CPU scheduling module which is discussed in section 2.1. Section 3 discusses the simulation results.

2 System Architecture

The architecture is designed with a middleware based approach. The device, application, user, current network profile are stored in a context repository (CR). The decision

* Corresponding author.

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 582–585, 2007.
© Springer-Verlag Berlin Heidelberg 2007

engine (DE) monitors the context cache for triggering the handoff upon detection of context changes i.e. which can be degradation in QoS due to changes in network characteristics, disconnection from the current network etc. Network Resource Monitoring Agent (NRMA) monitors the network resources i.e. packet arrival rate. The interaction among the modules is shown in figure 1. The CPU scheduler adaptation module is discussed in detail in subsection 2.1.

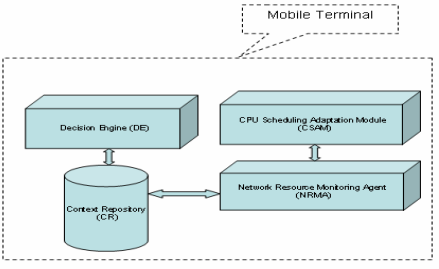


Fig. 1. System Architecture designed using a middleware based approach

2.1 CPU Scheduler Adaptation Module

A mobile terminal can either move from high data rate to low data rate or low data rate to high data rate. In case of first scenario the application consumes the data from the buffer at a faster rate than the arrival rate of the packets. This means that the application is allocated more CPU than needed. In the later case the buffer will fill up at a rate faster than the application consumption rate. The monitoring is done after an interval t which we call an *epoch*. An optimal point *q* is defined in a buffer which is achieved by dynamically changing the CPU allocation CPU_{alloc} proportional to the arrival rate of the packets.

We define two parameters *buffer fill level* and *proportional change* to implement CPU scheduler adaptation module.

Buffer fill level: The buffer fill level q(t) at time t is given by the formula

$$q(t) = \frac{Current\ buffer\ size\ at\ time\ t}{Total\ buffer\ size} \tag{1}$$

Proportional change: Let us represent the arrival rate of the packets in the current epoch as $\lambda_{current}$ and the arrival rate of the packets in the previous epoch as $\lambda_{previous}$. Then proportional change p(t) at time t is given by the formula

$$p(t) = \frac{\lambda_{current}}{\lambda_{previous}} \tag{2}$$

The algorithm for CPU scheduler adaptation module is given in Figure 2.

$\lambda_{previous} = updateArrivalRate()$ //Get initial arrival rate of the packets

for ever do

$\lambda_{current} = updateArrivalRate()$ //Get current arrival rate of the packets

$p(t) = \frac{\lambda_{current}}{\lambda_{previous}}$ //Calculate proportional change

If (p(t) < q or p(t) > q) //Check current buffer level

$CPU_{alloc} = CPU_{alloc} * p(t)$ //Proportionally allocate CPU resource

End if

$\lambda_{previous} = \lambda_{current}$

End for

Fig. 2. CPU Scheduling Adaptation Logic

3 Simulation

The simulation considers the following vertical handoff scenarios

- Switching from low data rate network to high data rate network
- Switching from high data rate network to low data rate network

We simulated our idea in Discrete Event Simulation using C++. For our simulation we have considered the optimal buffer size of 100 packets. Table 1 shows the parameters that we have chosen to simulate the vertical handoff situation.

3.1 Low Data Rate Network to High Data Rate Network

When a mobile terminal switches to a network of high data rate the queue length will increase with the passage of time due to increase in arrival rate of the packets. From Figure 3 it is evident that the buffer size remains optimal during vertical handoff scenario and interestingly the average waiting time experienced by the packet also converges as shown in Figure 3.

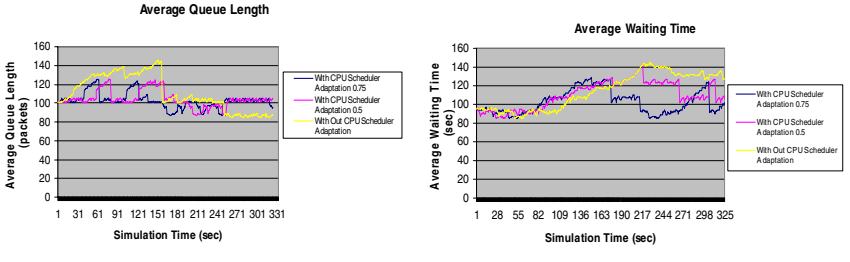


Fig. 3. Left side shows average queue length and right side shows average waiting time (Vertical handoff from low data rate to high data rate network)

3.2 High Data Rate Network to Low Data Rate Network

In case of moving to a network with low data rate the queue size will decrease which is shown in figure 4. The standard deviation from the optimal point with out CPU scheduling adaptation is 40 and with CPU scheduler adaptation the standard deviation is 15.3 and 18 for a proportional change of 0.75 and 0.50 respectively. The graph on the right hand side of Figure 4 shows the average waiting time of the packets in the queue in this scenerio.

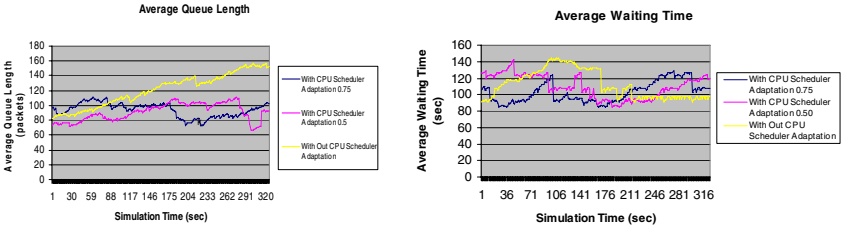


Fig. 4. Left side shows average queue length and right side shows average waiting time (Vertical handoff from high data rate to low data rate network)

Acknowledgements. This research was supported by MIC (Ministry of Information and Communication), Korea, under ITRC (Information Technology Research Center) support program supervised by the IITA (Institute of Information Technology Advancement). (IITA-2006-C1090-0603-0040).

Reference

1. Helal, S., Lee, C., Zhang, Y.G., Richard III, G.G.: An Architecture for Wireless LAN/WAN Integration. Wireless Communication and Networking Conference Vol. 3. IEEE, Chicago (2000) 1035-1041

Performance Evaluation of the Route Optimization Scheme in Mobile IPv6

In-Hye Shin[1,*], Gyung-Leen Park[1,**], Junghoon Lee[1], Jun Hwang[2], and Taikyeong T. Jeong[3]

[1] Department of Computer Science and Statistics, Cheju National University, Korea
{ihshin76,glpark,jhlee}@cheju.ac.kr
[2] College of Information and Media, Seoul Women's University, Korea
hjun@swu.ac.kr
[3] Department of Communications Engineering, Myongji University, Korea
ttjeong@mju.ac.kr

Abstract. The paper presents the performance comparison between the triangle routing and the route optimization. The paper presents not only the analytical models to compare the performance but also the threshold value to decide whether the mobile node had better use the route optimization or not. The model provides the approximate guideline when a network administrator is to implement mobile IPv6 with the route optimization under the given network environments.

1 Introduction

The Internet Engineering Task Force (IETF) has proposed "route optimization" in Mobile IP [1] to solve this "triangle routing" problem [1] that the packets from correspondent node (CN) must be forwarded to the mobile node (MN) via a home agent (HA). The route optimization enables the packets to be delivered directly between the MN and the CN. How effective is the route optimization? And will we always need the route optimization capability when we are to implement Mobile IPv6 under some conditions? There is no concrete answer for such questions so far [3]. Most of works have not presented the exact quantitative results considering various factors such as the number of hops between end-to-end host, the network bandwidth, the size of data packets, and the failure rate of the binding update. Therefore, we propose analytical models to evaluate the performance of the triangle routing scheme and that of the route optimization scheme and show the quantitative results of performance comparison.

2 The Proposed Model

Using the triangle routing as shown in Figure 1 (solid line), the packet destined to the MN is intercepted and tunneled by the HA. The tunneled packet will be forwarded to

* This research was supported by the MIC (Ministry of Information and Communication), Korea, under the ITRC support program supervised by the IITA (IITA-2006-C1090-0603-0040).
** The corresponding author.

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 586–589, 2007.
© Springer-Verlag Berlin Heidelberg 2007

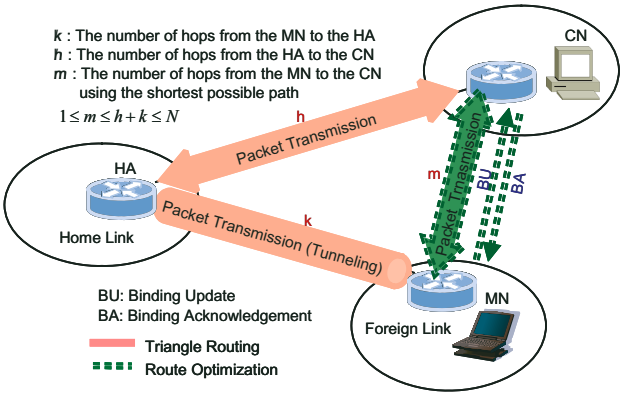


Fig. 1. The Triangle Routing and the Route Optimization in Mobile IPv6

the MN. Using the route optimization, the MN and the CN can communicate directly after using the BU/BA messages as shown in Figure 1 (dotted line).

Destination Option Header, to include Home Address Option [2], should be added to the packet routed from the MN to the CN, while Type 2 Routing Header should be used for the packet routed from the CN to the MN reversely, to carry its home address [4]. Table 1 depicts the notations used in the proposed analytical models.

Table 1. The Notations Used in the Proposed Model

Notation	Description	Value
BW_{wl}	The Average Bandwidth in Wireless Network (Mbps)	10
BW_{wd}	The Average Bandwidth in Wired Network (Mbps) = $BW_{wl} \times 10$	100
p	The Success Probability of the Binding Update Message	0.99
pkt	The Size of Receiving and Sending Data (byte)	
$PMTU$	The Size of Path MTU (byte)[5]	1500
IPH	The Size of IP Basic Header (byte) [5]	40
$MobH$	The Size of Mobility Header (byte) [2]	
$AuthH$	The Size of Authentication Header (byte) [6]	20
$FragH$	The Size of Fragment Header (byte) [5]	8
$DestOpH$	The Size of Destination Option Header (byte) [5]	20
$RoutH$	The Size of Type 2 Routing Header (byte) [2]	24
$IPTun$	The Additional Size for IP-IP Tunneling (byte) [4]	40
BU_M	The Size of the Binding Update Message (byte) = 40(IPH)+12($MobH$ with MH Type=5)+20($DestOpH$)	72
BA_M	The Size of the Binding Acknowledgement Message (byte) = 40(IPH)+12($MobH$ with MH Type=6)	52

Equation (1) shows the total costs for sending a BU message and a BA message. The retransmission for BU message failure has exponential back-off time and continues until an MN receives a BA or until a maximum (256 seconds) is reached [3]. Consequently, the cost of the routing optimization procedure is obtained by Equation (2).

$$T_{Msg} = \left\{ \frac{BU_M + BA_M}{BW_{wl}} + (m-1) \times \frac{BU_M + BA_M}{BW_{wd}} \right\} \times \frac{8(bit)}{1000 \times 1024^2 (bps)} \tag{1}$$

$$T_{ROpro} = p \times T_{Msg} + \sum_{i=2}^{10} p(1-p)^{(i-1)} (T_{Msg} + T_w), \qquad 1 \le T_w = 2^{(i-2)} \le 256 \tag{2}$$

Equation (3) creates temporary variables to keep the final formula simple.

$$\begin{aligned}
&g1 = IPH + AuthH + FragH, \quad g2 = IPH + AuthH + FragH + DestOpH \\
&g3 = IPH + AuthH + FragH + RoutH, \quad g4 = IPH + AuthH + FragH + IPTun \\
&r1 = PMTU - g1, \quad r2 = PMTU - g, \quad r3 = PMTU - g3, \quad r4 = PMTU - g4 \\
&q1 = \left\lfloor \frac{pkt}{r1} \right\rfloor \times PMTU + \left(pkt - \left\lfloor \frac{pkt}{r1} \right\rfloor \times r1 \right) + g1, \; q2 = \left\lfloor \frac{pkt}{r2} \right\rfloor \times PMTU + \left(pkt - \left\lfloor \frac{pkt}{r2} \right\rfloor \times r2 \right) + g2 \\
&q3 = \left\lfloor \frac{pkt}{r3} \right\rfloor \times PMTU + \left(pkt - \left\lfloor \frac{pkt}{r3} \right\rfloor \times r3 \right) + g3, \; q4 = \left\lfloor \frac{pkt}{r4} \right\rfloor \times PMTU + \left(pkt - \left\lfloor \frac{pkt}{r4} \right\rfloor \times r4 \right) + g4
\end{aligned} \tag{3}$$

Equation (4) shows the total delay required when an MN sends packets to a CN and receives ones from the CN. Equation (5) is obtained by adding the delay required when an MN sends the tunneled packets to the HA to the delay required when the HA sends the packets to the CN.

$$T_{MN:CN} + T_{CN:MN} = \left[\left\{ \frac{q2}{BW_{wl}} + (m-1) \frac{q2}{BW_{wd}} \right\} + \left\{ \frac{q3}{BW_{wl}} + (m-1) \frac{q3}{BW_{wd}} \right\} \right] \frac{8(bit)}{1000 \times 1024^2 (bps)} \tag{4}$$

$$T_{MN:HA:CN} = T_{CN:HA:MN} = \left\{ \frac{q4}{BW_{wl}} + \frac{h \times q1 + (k-1) \times q4}{BW_{wd}} \right\} \frac{8(bit)}{1000 \times 1024^2 (bps)} \tag{5}$$

The total delay (ms) for the packet transmission using the routing optimization and the triangle routing, are obtained by Equation (6) and (7), respectively.

$$\begin{aligned}
T_{RO} &= T_{ROPro} + T_{MN:CN} + T_{CN:MN} \\
&= p \times T_{Msg} + \sum_{i=2}^{10} p(1-p)^{i-1} (T_{Msg} + T_w) + \left\{ \frac{q2+q3}{BW_{wl}} + (m-1) \frac{q2+q3}{BW_{wd}} \right\} \frac{8(bit)}{1000 \times 1024^2 (bps)}
\end{aligned} \tag{6}$$

$$T_{TR} = +T_{MN:HA:CN} + T_{CN:HA:MN} = 2 \times T_{MN:HA:CN} = \left\{ \frac{q4}{BW_{wl}} + \frac{h \times q1 + (k-1) \times q4}{BW_{wd}} \right\} \frac{16(bit)}{1000 \times 1024^2 (bps)} \tag{7}$$

3 The Results of Performance Evaluation

We compare the performance of the routing optimization (RO) scheme and the triangle routing (TR) scheme in terms of the total transmission delay for the messages using the corresponding routing path. The difference in the numbers of hops between the MN and the CN with respect to two routing methods is relatively small. For example, m is 5 or 8 when h+k = 10, because the HA can be switched due to load balancing.

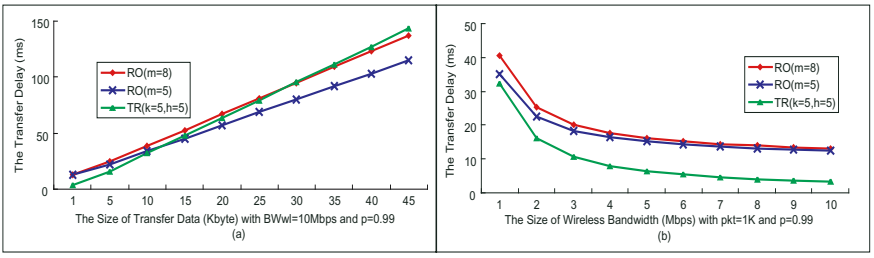


Fig. 2. (a) The Delay Comparison according to the Size of Transfer Data (b) The Delay Comparison according to the Bandwidth in Wireless Networks

Figure 2-(a) shows that as the size of the data increases, the delay of TR and that of RO do. In Figure 2-(b), the larger the size of the bandwidth is, the smaller the total delay is. Figure 2-(a) provides the threshold values when p=0.99 and wireless bandwidth is 10Mbps. Figure 2-(b) shows that the delay of TR is smaller than that of RO with the short term packet (1K), when p=0.99 and the bandwidth $BW_{wd} \leq$ 10Mbps.

4 Conclusion

The paper develops the analytical models to compare the performance of the triangle routing scheme and the route optimization scheme, considering several important factors such as the packet type (long or short term) between the MN and the CN, the number of hops between them, the network bandwidth, and the failure rate of the binding update procedure. It presents the threshold value to decide whether the MN should use the route optimization or not under various situations. The model also provides the approximate guideline when a network administrator is to implement mobile IPv6 with the route optimization capability.

References

1. Perkins, C., Johonson, D.B.: Route Optimization in Mobile IP. IETF Internet Draft draft-ietf-mobileip-optim-11.txt (2001)
2. Johnson, D., Perkins, C., Arkko, J.: Mobility Support in IPv6. IETF RFC 3775 (2004)
3. Soliman, H.: MobileIPv6, Addison-Wesley (2004)
4. Perkins, C.: IP Encapsulation within IP. IETF RFC 2003 (1996)
5. Deering, S., Hinden, R.: Internet Protocol, Version 6 Specification. IETF RFC 2460 (1998)
6. Kent, S., Atkinson,R.: IP Authentication Header. IETF RFC 2402 (1998)

An ID-Based Random Key Pre-distribution Scheme for Wireless Sensor Networks*

Tran Thanh Dai and Choong Seon Hong**

Networking Lab, Department of Computer Engineering, Kyung Hee University, Korea
daitt@networking.khu.ac.kr, cshong@khu.ac.kr

Abstract. When wireless senor networks (WSNs) are deployed in hostile areas, they indeed need to be secured by security mechanisms. To do this, cryptographic keys must be agreed on by communicating nodes. Unluckily, due to resource constraints, the key agreement problem in wireless sensor networks becomes quite intricate. In this paper, we propose a new ID-based random key pre-distribution scheme that is comparable to Du et al.'s scheme [2] in terms of network resiliency and memory usage. On the other hand, our later analysis shows that our scheme outperforms Du et al.'s scheme in terms of computational and communication overhead.

Keywords: ID-based, random key pre-distribution, key agreement, security, wireless sensor networks.

1 Introduction

There has been a trend that WSNs have been getting mature together with wider and wider applications and deployments. In such trend, providing security services based on solving the key agreement problem becomes one of the major concerns. Unfortunately, due to resource constraints of WSNs, such problem becomes quite intricate. Motivated by such challenge, in this paper, we propose a highly resilient, robust, resource-efficient, and ID-based random key pre-distribution scheme. Our scheme as analyzed later is much like Du et al.'s scheme [2] (*Du's scheme*) regarding network resiliency with the same memory cost. Moreover, our scheme significantly improves resource usage relating to computational and communication overhead compared to Du's scheme. The rest of this paper is organized as follows: section 2 mentions the related work; section 3 describes our ID-based random key pre-distribution scheme; section 4 analyzes the resiliency of our scheme against node capture attack and presents the performance analysis in terms of memory usage, communication overhead, and computational overhead; section 5 concludes the paper and states our future work.

2 Related Work

Matsumoto and Imai proposed an efficient scheme (IM scheme) for the key agreement problem between two entities [1]. Fig. 1 illustrates how a pairwise

* This work was supported by MIC and ITRC Project.
** Corresponding author.

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 590–593, 2007.
© Springer-Verlag Berlin Heidelberg 2007

key $K_{ij} = K_{ji}$ is generated where N - maximum number of deployable nodes; $m = \log_2(N)$ - number of bits used to represent an effective ID of each entity; l - number of $(m \times m)$ symmetric matrices M_ωs over finite field $GF(2)$; y_i $(i = \overline{1, N})$ - m-dimensional vector, effective ID of node S_i; y_i $(i = \overline{1, N})$ - m-dimensional vector, effective ID of node S_i.

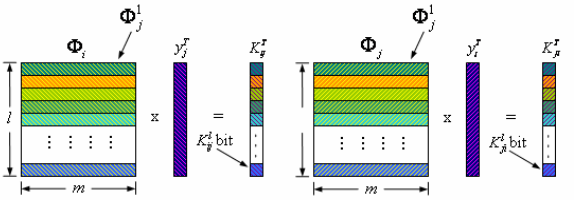


Fig. 1. Pairwise key generating in MI scheme

3 ID-Based Multiple Space Key Pre-distribution Scheme

3.1 Keying Material Pre-distribution Phase

During this phase, we have to pre-distribute keying material to each node such that after deployment neighboring nodes can derive a pairwise key between them using this material. This phase is performed as follows. First, a central server generates λ key spaces. Each key space Ω_i consists of l $(m \times m)$ symmetric matrices $M_{i\omega}$s as defined in IM scheme. Then, we randomly choose μ distinct key spaces from key spaces for each node. For each space chosen by node S_j, we first compute keying material Φ_{ji} and then store it at this node. Therefore, each node S_j has distinct values of Φ_{ji}s. Using MI scheme; two nodes can derive a pairwise key if they have both chosen a common key space.

3.2 Pairwise Key Establishment Phase

After deployment, each node needs to discover whether it shares any key space with its neighbors. Suppose that nodes S_i and S_j are neighbors, then they instantly broadcast a message containing the following information: each node's effective ID and the indices of the key spaces it carries. If they figure out that they have an identical index of a key space (or identical key space Ω_s), they can easily compute their pairwise key using MI scheme. Conversely, there is the case that two nodes who even are neighbors could not establish a pairwise key. To tackle this problem, the method presented in [2] could be utilized. Due to the limited size of paper, we do not discuss in detail here.

4 Security and Performance Analysis

Our security evaluation is conducted by finding the answer to two questions: (i) Given that b nodes are captured, what is the probability that at least one key space is broken? (ii) Given that b nodes are captured, what fraction of the additional communications (communications among un-captured nodes) also becomes compromised? These two questions are already answered in [2]. The answer to the first question is as follows:

$$P(\text{at least one space is broken}|C_b) = \lambda \sum_{k=m}^{b} \binom{b}{k} \left(\frac{\mu}{\lambda}\right)^k \left(1-\frac{\mu}{\lambda}\right)^{b-k}. \quad (1)$$

The answer to the second question is the following equality:

$$P(s \text{ is broken}|C_b) = \sum_{k=m}^{b} \binom{b}{k} \left(\frac{\mu}{\lambda}\right)^k \left(1-\frac{\mu}{\lambda}\right)^{b-k}. \quad (2)$$

where s denotes an additional secure communication link.

To analyze our scheme performance, we evaluate its memory usage, communication overhead, and computational overhead using Du's scheme as a benchmark. For each key space, according to MI scheme, each node S_i has to spend $m \times l$ bits on storing the value of Φ_i. Thus the total memory usage (KB) for each node with μ chosen key spaces is: $\frac{m \times l \times \mu}{8 \times 1024}$. This value is exactly identical to the memory consumption of Du's scheme. Concerning the communication overheads of our scheme and Du's scheme, we draw a self-explanatory comparison as shown in fig. 2.

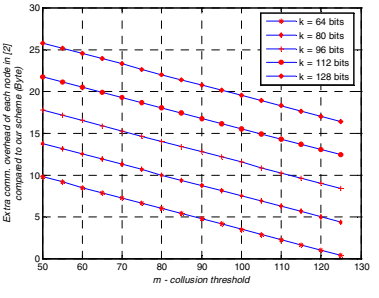


Fig. 2. Extra communication overhead of each node in [2] compared to our scheme

Regarding computational overhead, to compute a pairwise key, each node of our scheme needs to perform a multiplication of a $(l \times m)$ matrix and an $(m \times 1)$ effective ID. Therefore, each node needs $l \times m$ single-precision multiplications while each node in [2] needs to do $2 \times (m-1) \times l^2$ single-precision multiplications. It follows that the computational overhead of our scheme is far less than that in [2]. The numbers in fig. 3 reinforce our argument.

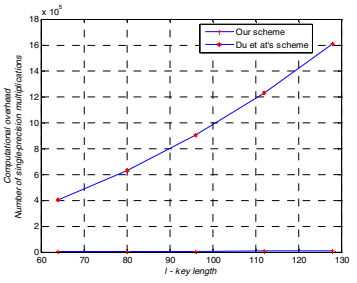


Fig. 3. Computational overhead in each node with various key lengths

5 Conclusions and Future Work

This paper proposes a new key pre-distribution scheme for WSNs that can be considered as a refinement of two types of schemes: ID-based key pre-distribution scheme and random key pre-distribution scheme. As a result, our scheme possesses a number of attractive properties. First, our scheme is scalable and flexible in terms of network size. Second, our scheme substantially improves network resiliency against node capture attack compared to schemes [3], [4] and are comparable to Du's scheme. The performance of our scheme is also investigated to show its efficiency. Accordingly, our scheme is the same as Du's scheme in terms of memory usage. Moreover, our scheme is more efficient than Du's scheme concerning communication overhead. Finally, computational overhead of our scheme is argued to be far less than that of Du's scheme. However, our scheme is still vulnerable to node replication attack and key-swapping collusion attack. Therefore, in our future work, we would like to explore additional mechanisms to efficiently and radically thwart these attacks.

References

1. Matsumoto, T., Imai, H.: On the Key Predistribution System: A Practical Solution to the Key Distribution Problem. CRYPTO'87, LNCS Vol. 293, 8(1987)185-193
2. Du, W., Deng, J., Han, Y.S., Varshney, P.K., Katz, J., Khalili, A.: A pairwise key predistribution scheme for wireless sensor networks. ACM Trans. Info. Sys. Sec., Vol. 8, No. 2, 5(2005)228-258
3. Eschenauer, L., Gligor, V.D.: A key-management scheme for distributed sensor networks. Proc. of the 9th ACM Conference on Computer and Communications Security, 11(2002) 41-47
4. Chan, H., Perrig, A., Song, D.: Random key predistribution schemes for sensor networks. Proc. IEEE Symposium on Security and Privacy, 5(2003)197-213

An Adaptive Mobile System to Solve Large-Scale Problems in Wireless Networks*

Jehwan Oh and Eunseok Lee**

School of Information and Communication Engineering, Sungkyunkwan University
300 Chunchun jangahn Suwon, 440-746, Korea
{hide7674,eslee}@selab.skku.ac.kr

Abstract. In this paper, we propose an adaptive mobile system that uses mobile grid computing to achieve infeasible work in mobile device. Recently, grid computing has gained significance for its ability to achieve certain goals by sharing the idle resources of computing devices, and thus overcoming the various constraints of the mobile computing environment. In the mobile environment, it is very important for mobile devices to recognize correctly their resource state. The proposed system includes a Grid Inference Engine, which decides on the resource state by using fuzzy logic. In this paper, a prototype is implemented to evaluate the proposed system. The effectiveness of the system is confirmed through experiment.

1 Introduction

In the development of wireless network technology, the diffusion of mobile devices is spreading, which is further increasing their availability. The mobile devices are continually developing in terms of technology and capability. However, mobile devices do not always satisfy user requests effectively, because they are limited by small display, low speed CPU, and low capacity memory. Therefore, applications executed on mobile devices should be very lightweight, and large-scale applications that require complicated operations are difficult or impossible to execute. Users' desire to be offered applications with service quality comparable to desktop computers on mobile devices is increasing, but programmers can only develop just applications that provide restrictive services to user. To effectively solve these issues, grid computing[1] was introduced, which uses the idle resources of many computers connected in a network, and therefore, yields results similar to traditional wired networks. Grid computing produces very high efficiency super-computing resources, which can easily be used to solve work requiring high efficiency, and large-scale calculation ability, while operating in a cooperative environment.

* This work was supported in parts by Ubiquitous Autonomic Computing and Network Project, 21th Century Frontier R&D Program, MIC, Korea, ITRC IITA-2006-(C1090-0603-0046), Grant No. R01-2006-000-10954-0, Basic Research Program of the Korea Science & Engineering Foundation, and the Post-BK21 Project.

** Corresponding author.

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 594–597, 2007.
© Springer-Verlag Berlin Heidelberg 2007

In existing mobile grid research[2][3][4][5], mobile device nodes use the resources of wired grid networks. In the proposed system, mobile devices can offer their own resources as well as use the resources of wired grid networks. It is very important for mobile devices to recognize correctly their own resource state. Therefore, we propose the Grid Inference Engine (GIE) for this purpose. The GIE decides whether the module can provide resources, and if possible, it decides on the resources to provide. Also, it decides whether the resource of Grid computing is required, according to resource state and requested jobs. If it is required, it decides the number of jobs to request. We used fuzzy logic to make this decision.

In this paper, a prototype is implemented. The effectiveness of the system is confirmed through experiment.

2 Proposed System

We consider the technological and functional constraints of mobile devices. We propose a method to monitor mobile resources and to decide on the resources to provide. This system creates new possibilities in mobile devices. The proposed system is based on mobile grid computing, which is an extension of traditional grid computing.

2.1 System Architecture

The overall architecture of the proposed system is composed of two components: the Client Module, embedded in the client device, and the Server Module, which operates on the server side.

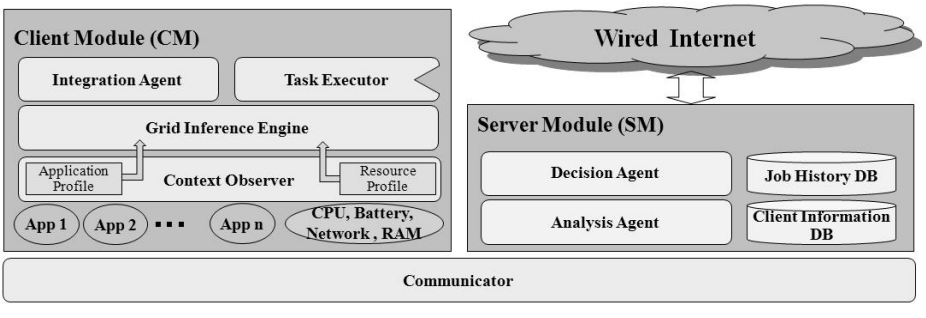


Fig.1. Components of the Proposed System

- *Context Observer:* gathers dynamically changing context information such as CPU state, battery life and creates Application profile and Resource profile
- *Grid Inference Engine:* decides the resource state of the device
- *Integration Agent*: integrates the received job results from other devices, and delivers the integrated results to the application
- *Task Executor*: executes the jobs required for an external peer
- *Communicator*: performs the interaction between modules

2.2 Grid Inference Engine

We propose the GIE to decide the resource state of a mobile device. The GIE uses fuzzy logic to make this decision. The GIE decides whether the module can provide resources. If possible, the module decides how much resources can be supported. Also, it decides whether a Grid is required according to resource state and requested jobs. If it is required, it decides the number of jobs to request.

System Profile: The proposed system recognizes changes in resources. Information regarding resources may be divided into static resources and dynamic resources. Static resources represent information that does not change, such as CPU type, RAM size, storage size, and so on. Dynamic resources represent information that changes in response to varying environments, such as CPU load, free RAM, remaining battery, usable storage, and so on. The proposed system describes static and dynamic resource information using XML, and maintains a list of varying resources.

Application Profile: The application is programmed in Java. Java is an object-oriented language, meaning that data is encapsulated and applications consist of a collection of objects. The application is designed so that distributed classes execute independently of each other. The application profile describes the required information for class execution.

In the proposed system, the level of the Membership function is divided into five levels: MH(Max High), H(High), N(Normal), L(Low) and ML(Max Low). The system developer or user has to input directly the level of membership. We generate the 125 rules for fuzzy inference. Fuzzy logic has four levels, which are, fuzzy matching, inference, combining fuzzy conclusion, and defuzzification [6].

3 System Evaluation

The effectiveness of the proposed system is achieved by applying it to an 'Application for Emergency-situation' using mobile devices. The Application for Emergency-situation has functions such as identifying a patient, searching medical records, enabling video-conferencing, etc. This application consists of objects that execute independently.

The capability and utility of the system were validated using mobile devices developed for the Personal Java runtime environment on iPAQ Pocket PCs. Two Pocket PCs were used for performance evaluation. The Pocket PC used was an iPAQ h5450, equipped with a 400 MHz processor, 64 MB of RAM and 11-Mbps IEEE 802.11b wireless card. The experiment assumes same bandwidth values for communication.

The experiment was conducted to test the system's execution time. Pocket PC A used the proposed system, and Pocket PC B used the raw resource state. For evaluation, each Pocket PC executes the 'Application of Emergency' application, and is compared in terms of execution time. We continually executed the same service to create overhead on Pocket PC A and B. The execution time is the processing time taken, from receiving an initial user task request, to completing the task.

Initially the result of the execution time of Pocket PC A is similar to the result of the execution time of Pocket PC B. However, without the GIE, the application on

Pocket PC B crashed when the service was executed 10 times. The application on Pocket PC A crashed when the service was executed 15 times. The GIE secured stable execution time operating on mobile devices. This confirms the efficiency of the proposed system in terms of deciding on the most effective resource state.

4 Conclusion

This paper presented an adaptive mobile system using mobile grid computing to overcome the constrained performance of mobile devices, and proposed the Grid Inference Engine which monitors resources and decides on the resource state.

A prototype was implemented in order to evaluate the proposed system in terms of deciding on the most efficient resource state. The effectiveness of this system is confirmed through experiment.

References

1. Foster, I., Kesselman, C., Nick, J.M., Tuecke, S.: The Physiology of the Grid: An Open Grid Services Architecture for Distributed Systems Integration.OPEN Grid Service Grid Services Architecture for Distributed Systems Integration. Open Grid Service Infrastructure WG, Global Grid Forum, 6(2002)
2. Chan, A.T.S., Chuang, S.N.: MobiPADS: A Reflective Middleware for Context-Aware Mobile Computing. IEEE Transaction on Software Engineering. vol.29, no.12(2003)1072-1085
3. Gu, X.H., Nahrstedt, K., Messer, A., Greenberg, I., Milojicic, D.: Adaptive offloading for pervasive computing. Pervasive Computing, IEEE vol.3, no.3, (2004)66-73
4. Hwang, J., Aravamudham, P.: Middleware Service for P2P Computing in Wireless Grid Networks. IEEE Internet Computing vol.8, no.4, (2004)40-46
5. Phan, T., Huang, L., Dulan, C.: Challenge: Integrating Mobile Wireless Devices into the Computational Grid. Proceedings of the 15th Annual ACM Symposium on Principles of Distributed Computing, ACM Press, MOBICOM'02, 9(2002)271-278
6. Lee, K.H.: First Course on Fuzzy Theory and Applications. Advances in Soft Computing. Springer(2005)
7. Lum, W.Y., Lau, F.C.M.: User-Centric Content Negotiation for Effective Adaptation Service in Mobile Computing. IEEE Transaction on Software Engineering. vol.29, no. 12(2003)1100-1111

Answer Extracting Based on Passage Retrieval in Chinese Question Answering System*

Zhengtao Yu[1,2], Lu Han[1], Cunli Mao[1], Yunwei Li[1], Yanxia Qiu[1], and Xiangyan Meng[1]

[1] The School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, P.R. China 650051
ztyu@bit.edu.cn,hanlu530530@163.com, maocunli13888832386@yahoo.com.cn, li1225yingwei@126.com, ytaq007@soho.com,mengxiangyan1984@yahoo.com.cn
[2] The Institute of Intelligent Information Processing, Computer Technology Application Key Laboratory of Yunnan Province, Kunming,P.R.China 650051

Abstract. Passage retrieval is a base of answer extracting in Chinese question answering system.In order to improve the precision of the passage retrieval which includes answers, this paper analyze first existing passage retrieval methods,and point to their shortcomings, then put forward a new method of passage retrieval, in which factors are considered ,such as frequency of query words and query expansion words in passage, the length of passage, the distribution density and minimum match span of query words and query expansion words in passage ,when calculating passage weight.The new method can realize passage retrieval on the basis of the weight calculation, and obtain optimization answers passage. Final, experiment result of answer extracting based on passage retrieval have shown that the method given has better effect.

Keywords: Chinese Question Answering System;Passage retrieval, Word frequency, Distribution density, Match Span.

1 Introduction

The current information search engines (Google, Yahoo, etc.) can provide users with a large number of WebPages. Users have to find out useful information from these WebPages again. Question Answering System provides users with answers which have been treated. For example, user's question Q1: 什 出 的?(When was Zedong Mao born?)The demanded answer is a sentence which includes "1895", or a direct answer string "1895", but not a lot of WebPages associated with the question. The distinction between Question Answering System and conventional search engines is that the form of providing answers is different. Search engine only positions to the WebPages and the answer is relatively rough. But Question Answering System provides users with the passage, phrase and words, which the answer is relatively brief.

* This paper is supported by National Nature Science Fundation (60663004),Ministry of Education PH.D. Foundation (No.20050007023) and Kunming University of Science and Technology PH.D. Fundation (2006-12).

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 598–605, 2007.
© Springer-Verlag Berlin Heidelberg 2007

The key to Question Answering System is to extract the answer which is more briefer and more relevant to the questions from the recalled Webpages of information retrieval. In TREC8,TREC9 and TREC10, for English Factoid questions, it is aimed to extract the passage which contains 50 or 250 bytes string as the answer[1]. Since the TREC11, it was aimed to extract the correct answer directly but not including the other information which has nothing to do with the question[2]. Because of the passage is the base of answer extraction, if the answer is not included in the passage, the extraction of the answer will be fail. Conversely, if the answer is included in the passage and in the front list of candidate passage, the accuracy and efficiency of answer extraction will be improved. So, in order to improve the accuracy of the answer extracting ,this paper studies answer extracting based on passage retrieval.

2 Related Works

Different from text retrieval, the object of passage retrieval is passage rather than text. The calculation of passage weight is the key of the search ranking. At present, there are many methods of calculation passage weight aiming at English Question Answering System: MITRE used the word overlap algorithm which is presented by Light[3]. It believes that the passage weight is determined by the number of the same words between passage and query words. Then the number of the same words between the passage and query words will be calculated as similarity metrics; MultiTex is a density-based passage retrieval algorithm[4]. This method considers that a short sentence which includes more query words should have a higher weight. And the passage weight is determined by the weight of the words and the length of the passage; SiteQ's passage retrieval algorithm receives the weight of a passage composed with n sentences by calculating every sentence weight, and the sentence weight is determined by the weight of matching query words and the distance of the query words in the sentence[5]; Alliance's passage retrieval algorithm receives it's weight by calculating the cosine value between query words vector and the passage vector. It mainly considers the value *tfidf* of words that appear in both the query words and the passage[6].

The researches on passage retrieval algorithm about Chinese Question Answering System are very less. Yongkui Zhang etc present a model of N-gram model and a modified space vector model[7], the first one mainly considers the relationship of N-gram. It distributes the passage weight by the appearing probability of N-grim. The second one adopts Okapi's method of weight distributing[8]. It mainly considers the frequency of the query words appear in the passage and the length of the passage. Shifu Zheng presents a method to calculate the passage weight [9]. It mainly considers the query word in the passage, the synonyms of the query word and the query words not in the passage etc. Huan Cui etc present a method of calculating the similarity between question and passage[10]. It mainly considers the number of the same keywords between the question and the passage, the order of the keywords appear in passage, the distance of the keywords appear in passage and the length of question and passage.

From the methods to calculate the passage weight above, there are many factors affect the passage weight. But there are no mature methods now. Currently, Chinese passage weight calculation mainly considers the frequency feature of the query words

which appear in the passage, less considers the query words semantic information and the location of the query words in the passage. The word frequency of query words and query expansion words in passage, the length of passage, the distribution density and minimum match span of query words and query expansion words in passage are considered in this paper. Then a method to calculate the passage weight is presented, and the candidate passages set is retrieved on the basis of weight calculation.

3 Answer Passage Retrieval Based on Weight Calculation

3.1 Passage Retrieval Based on Vector Space Model

Whether a passage associates with the question can be judged by calculating the similarity between the passage and question. In information retrieval, the VSM is commonly used in calculating the similarity.The passage P which contains the answer is expressed as a vector $P = \{p_1, p_2 \cdots p_i \cdots p_n\}$, n is the number of all words, p_i is the frequency of No. i word which appears in the passage. Similarly, the question Q is expressed as a vector $Q = \{q_1, q_2 \cdots q_i \cdots q_n\}$, then calculate the cosine value of the two vectors as the similarity of the question and passage. The similarity calculate expressions based on VSM is given as follows[10]

$$Sim(Q, P) = \frac{\sum_{i=1}^{n} q_i \times p_i}{\sqrt{\sum_{i=1}^{n} (q_i)^2 \sum_{i=1}^{n} (p_i)^2}} \tag{1}$$

The method based on VSM is a statistic method,in which frequency information (occurring, not occurring, frequency occurring) of words in question and passage can be considered. Therefore, only the more words are included in the passage, then the more times relative words occur repeatedly, and then the good effect of this statistic method can be embodied. But at passage retrieval, because the number of words in the passage is less, so it can not enough to embody the effect of the method.

3.2 The Passage Weight Calculation Method

In passage retrieval, the question must be segmented first, then query word will be extracted. But the words occurring in passage always was not the query words itself but synonymy or related word of query words. Therefore, besides the query words are considered,the impact of query expansion words must be considered when calculating the passage weight. So in this paper, a relatively simple linear expansion strategy is used, firstly expanded query words with related words according to question type, then expanded the noun of query words with synonymy in HowNet[11], then query expansion words is formed. The query expansion words may contain many words, but not all the words have equivalent function for passage retrieval. Therefore, the weight of each query expansion words will be computed by using the method of *tfidf*[11]. So the first m query expansion words will be extracted as query expansion words set to

passage retrieval. The impact of these query expansion words to passage retrieval will be considered in passage retrieval.

The frequency feature of query words and query expansion words is a crucial factor in the passage. Vector space model reflects the words frequency characteristics, but the affect of length of the passage to weight is not considered. So in this paper, the VSM is modified according to the length of the passage. The specific modified methods are given in calculation methods as follows.

The distribution density of query words and query expansion words in the passage have great affect to passage retrieval. If the distribution of query words and query expansion words in the passage more intensive, the possibility of the passage which contains relevant answers is greater, and the passage weight is greater,so the distribution density of query words and query expansion words in the passage will be considered in this paper.

The author had discussed that query words and query expansion words has a great impact on the text retrieval [11]. Certainly, the distribution of the query words and query expansion words has a different of influence to passage retrieval. So the minimum match span methods are presented in this paper, (the specific method refer to the literature [11] definitions 1 and definitions 2) to calculate the minimum match span of the query words and query expansion words in the passage. Then the influence of the distribution of query words and query expansion words in passage retrieval will be considered.

All these factors are considered, a method of Chinese passage weight calculation is presented.The weight both of the word frequency of query words and query expansion words in passage and the length of passage weight ($W_{overlap}$), the weight of distribution density of query words and query expansion words in passage ($W_{density}$), the weight of minimum match span of the query words and query expansion words in passage (W_{mms}) are considered . The passage weight calculation (P_i) is given as following:

(1) The weight both of the word frequency of query words and query expansion words in passage and the length of passage ($W_{overlap}$)

$$W_{overlap}=\sum_{i=1}^{|q^{'}|}\frac{3\times tf(t_i)\times idf(t_i)}{0.5+1.5\times\frac{Pas_Len}{Pas_len_avg}+tf(t_i)}+\sum_{j=1}^{|q_a|}\frac{3\times tf(t_j)\times idf(t_j)}{0.5+1.5\times\frac{Pas_Len}{Pas_len_avg}+tf(t_j)} \quad (2)$$

$q^{'}$ is the set of query words, q_a is the set of query expansion words, $|q^{'}|$ is the number of query words in query words set, $|q_a|$ is the number of query expansion words in the query expansion words set, *Pas_Len* is the number of the words in the passage , *Pas_len_avg* is the average length of passage in the candidate passage set, *tf* is the words frequency of question query words and query expansion words, *idf* is inverted document frequency of question query words and query expansion words.

(2) The weight of the distribution density of query words and query expansion words in passage ($W_{density}$)

$$W_{density} = W_{density}^{q} + W_{density}^{q_a} = W_{density}^{q^{'}+q_a} \tag{3}$$

$$W_{density} = \frac{\sum_{j=1}^{|q|-1} tf(t_j) \times idf(t_j) + tf(t_{j+1}) idf(t_{j+1})}{\alpha \times dist(j, j=1)^2} \times |matched_q| \tag{4}$$

q is composed by query words $q^{'}$ and query expansion words q_a

$dist(j, j+1)$ is distance between word j and word $j+1$ in passage

$matched_q$ is a set of query words q matched in a passage

α is constant.

(3) The weight of minimum match span of the query words and query expansion words in passage (W_{mms})

$$W_{mms} = \left(\frac{\left|q^{'} \cap p\right| + ne(q_a \cap p)}{1 + \max(mms) - \min(mms)} \right)^{\mu} \times \left(\frac{\left|q^{'} \cap p + \mathrm{ne}(q_a \cap p)\right|}{\left|q^{'}\right| + \mathrm{ne}(q_a)} \right)^{\nu} \tag{5}$$

mms is the minimum match span based on query words and query expansion words, $\left|q^{'} \cap p\right|$ is the number of query words that appear in the passage, $q_a \cap p$ is the set of query expansion words that occur in the passage , the $ne(\cdots)$ function returns the number of set $q_a \cap p$. If $\left|q_a \cap p\right| > 0$, $ne(q_a \cap p)$ returns $\left|q_a \cap p\right|$ which is the number of query expansion words appears in passage. Otherwise, $ne(q_a \cap p)$ returns 0. It shows that query expansion words can not be found in passage. μ ν are constant numbers which always $\mu = 1/8$ $\nu = 1$.

All the above factors are considered, then the formula is given as following:

$$Weight(p_i) = \lambda_1 W_{overlap} + \lambda_2 W_{density} + \lambda_3 W_{mms} \tag{6}$$

$$\lambda_1 + \lambda_2 + \lambda_3 = 1 \tag{7}$$

3.3 Extracting Passage as Answer of Question

Given question and candidate passage set, according to formula (6), similarity between question and each candidate passage can be computed, but it is not sure for correct answers to be contained in passage with maximum similarity. Therefore, we have to get rid of some passage not according with semantic logic with the help of semantic analysis[12].During the course of question parsing, question type of each question can be recognized[13].Therefore,it is required that the candidate passage extracted must contain the corresponding answer type, such as question asking “人名

(HUM-PERSON)" is required that named entity of "人名(HUM-PERSON)" must be contained in candidate passage.The recognition of answer type is very complicated. We recognize the answer type according to whether the named entity or related words of corresponding question type are included in passage or not. Recognition algorithm of named entity refers to LPT: Language technology platform API which is developed by Information Retrieval Laboratory, Harbin Institute of Technology[14]. This Recognition algorithm has very good name entity recognition ratio for 7 kinds of question types such as "人名(HUM-PERSON) ", " 名(LOC-LOC)" , " 份(TIME-YEAR)" and so on. It is regarded that the passage is the candidate passage of the question if the named entity of corresponding question type in passage is recognized. But parts of question types have not corresponding named entity, so it can not be recognized by named entity. Therefore, a related words list for all question types is constructed. The number of the same words in passage and list of related words for corresponding question type is counted, and it is regarded that this passage is the corresponding answer type when the number exceeds a constant value.

Through judging the answer type of every passage, if the passage answer type does not according with corresponding question type, the passage is excluded outside the corresponding answer candidate passage set .Because it is possible that a passage may have several answer types, it is regarded the passage as the candidate passage as long as corresponding answer type appears in the answer. After get rid of the passage which does not correspond answer type, all candidate passage will be sorted according to passage weight, and then the first 5 passage will be extracted as answers.

4 Experiment and Evaluation

The experiment mainly focuses on evaluating the performance on answer extracting with passage retrieval. At present, there is no standard test data set of question and passage in Chinese question answering system, therefore, the experiment is only restricted in passage retrieval to factoid questions of 5 fine-grained question types including "人名(HUM-PERSON)", " 市(LOC-CITY)", " 份(TIME-YEAR)" ," NUM-WEIGHT)", " OBJ-FOOD)". Select 60 questions for 5 types respectively to practice. Analyze those 60 questions, then get the query words and submit them to Baidu search engine, retrieve the first 80 recalled abstract as the resource of the passage retrieval.

The evaluating method of the passage retrieval refers to TREC8, and takes MRR (Mean Reciprocal Rank) as evaluation criterion. The formula (8) of MRR is following:

$$MRR = \frac{1}{n}\sum_{1}^{n}\frac{1}{r_i} \tag{8}$$

n is the number of the testing question, r_i is the position of the passage which include the first right answer to the i question.

Takes VSM as the baseline in the experiment, considers the weight both of the word frequency of query words and query expansion words in passage and the length of passage ($W_{overlap}$), the weight of distributing density of query words and query

expansion words in passage ($W_{density}$) and the weight of minimal match span of query words and query expansion words in passage (W_{mms}) respectively. Then, we design four experiment methods. Method 1:only consider $W_{overlap}$ ($\lambda_1 = 1, \lambda_2 = 0, \lambda_3 = 0$; Method 2: consider $W_{overlap}$ and $W_{density}$ ($\lambda_1 = 0.6, \lambda_2 = 0.4, \lambda_3 = 0$; Method 3: consider $W_{overlap}$, $W_{density}$ and W_{mms} ($\lambda_1 = 0.5, \lambda_2 = 0.3, \lambda_3 = 0.2$; In experiment, the first 2 query expansion words are selected to calculate the weight (m=2). Table 1 shows different experiment results with different passage retrieval methods.

From the experiment data of table1, takes the different methods to passage retrieval, takes VSM as the baseline, considers both the word frequency of query words and query expansion words in passage and the length of passage, the distributing density of query words and query expansion words in passage, the minimal match span of query words and query expansion words in passage etc. Compare the effect of the passage retrieval. When only considers VSM, the MRR is 0.36.When considers both the word frequency of query words and query expansion words in passage and the length of passage, the MRR is 0.41. When more considers distributing density of query words and query expansion words in passage, the value of MRR is improved. It is 0.45. Specially, when considers the four factors together, the value of MRR is improved largely. It has good effects that MRR value is 0.50.

Table 1. Passage Retrieval Experiment Result

Question type	number	Recall rate	MRR			
			VSM	Method 1	Method 2	Method 3
(HUM-PERSON)	60	82.2%	0.45	0.49	0.53	0.59
城 (LOC-CITY)	60	77.8%	0.26	0.31	0.35	0.41
年 (TIME-YEAR)	60	80.6%	0.31	0.35	0.40	0.46
量(NUM-WEIGHT)	60	78.6%	0.41	0.45	0.48	0.52
(OBJ-FOOD)	60	81.3%	0.38	0.43	0.47	0.50
Average	60	80.1%	0.36	0.41	0.45	0.50

5 Conclusions

Aiming at characters of the Chinese question answer system, the method of calculating the passage weight which is presented in the paper not only considers the number of the same words between the query words and the expanding words in question and passage, the length of the passage, the distributing density of the query words and the expanding words, the minimal distributing distance etc. The experiment results of candidate passage retrieval show that the method has good effect. Next, the relationship among the several factors affected the weight of the passage retrieval and the answer extraction arithmetic will be studied.

Acknowledgements. Thanks for information retrieval laboratory, Harbin Institute of Technology provided the LTP: Language Technology Platform API with us to word segmentation, passage split, Name entity recognition.

References

1. Voorhees, E., Tice, D.: The TREC-8 Question Answering Track Evaluation. In Proceeding of the Eighth Text REtrieval Conference. NIST Special Publication,Gaithersburg, (2000)83-105
2. Voorhees, E.: Overview of the TREC 2002 Question Answering Track. In Proceeding of the 11th Text REtrieval Conference, NIST Special Publication,Gaithersburg, (2002) 115-123
3. Light, M., Mann, G.S., Riloff, E.: Analyses for Elucidating Current Question Answering Technology. Journal of Natural Language Engineering, Special Issue on Question Answering, Fall–Winter (2001)
4. Clarke, C., Cormack, Tudhope, G.E.: Relevanceranking for One to Three Term Queries. Information Processing and Management, 36(2000)291-311
5. Lee, G. , Seo, J., Lee, S.: SiteQ: Engineering High Performance QA System Using Lexico-Semantic Pattern Matching and Shallow NLP. In Proceedings of the Tenth Text REtrieval Conference, NIST Special Publication, Gaithersburg,(2001)
6. Llopis, F., Vicedo, J. L.: IR-n: A Passage Retrieval System at CLEF-2001. In Proceedings of the Second Workshop of the Cross-Language Evaluation Forum, (2001)
7. Zhang, Y.K, Zhao, Z.B. L.: Internet-based Chinese Question-Answering System. Computer Engineering. 29(15)(2003)84-86.
8. Robertson, S.E., Walker, S., Jones, S.: Okapi at TREC4. In Proceedings of the Forth Text REtrieval Conference, NIST Special Publication , Gaithersburg, (1996)73-79
9. Zheng, S.F., Liu, T., Qin B.: Overview of Question Answering. Journal of Chinese Information Processing, 16(6)(2002) 46~52
10. Cui, H., Cai, D.F., Miao, X.L.: Research on Web-based Chinese Question Answering Systerm and Answering Extraction. Journal of Chinese Information Processing. 18(3)(2004)24-31
11. Yu, Z.T., Fan X.Z., Song L.R.: The Research on Query Expansion for Chinese Question Answering System. Lecture Notes in Artificial Intelligence, Vol. 3613. Springer-Verlag, Berlin Heidelberg New York, (2005) 571-579
12. Kim, S., Baek, D., Kim, S.B.: Question Answering Considering Semantic Categories and Co-occurrence Density TREC-9. In Proceeding of the 11th Text Retrieval Conference, NIST Special Publication ,Gaithersburg, (2000)317-325
13. Yu, Z.T.,Fan, X.Z.,Guo, J.Y.:Chinese Question Classification Based on Support Vector Machine. 33(9)(2005)25-29
14. Lang, J., Liu T., Zhang, H.P., Li, S.: LTP:Language Technology Platform. In Proceeding of the Third Student Workshop on Computation Language, Shengyan, (2006) 64-68

Performance Evaluation of Fully Adaptive Routing for the Torus Interconnect Networks

F. Safaei[1,3], A. Khonsari[2,1], M. Fathy[3], and M. Ould-Khaoua[4]

[1] IPM School of Computer Science, Tehran, Iran
[2] Dept. of ECE, Univ. of Tehran, Tehran, Iran
[3] Dept. of Computer Eng., Iran Univ. of Science and Technology, Tehran, Iran
[4] Dept. of Computing Science, Univ. of Glasgow, UK
{safaei,ak}@ipm.ir,{f_safaei,mahfathy}@iust.ac.ir,
mohamed@dcs.gla.ac.uk

Abstract. Adaptive routing algorithms have been frequently suggested as a means of improving communication performance in parallel computer networks. These algorithms, unlike deterministic routing, can utilize network state information to exploit the presence of multiple paths. Before such schemes can be successfully incorporated in networks, it is necessary to have a clear understanding of the factors which affect their performance potential. This paper proposes a new analytical model to obtain message latency in wormhole-switched 2-D torus interconnect networks. The analysis focuses on a fully adaptive routing which has been shown to be one of the most effective in torus networks. The validity of the model is demonstrated by comparing analytical results with those obtained through simulation experiments.

Keywords: adaptive routing algorithm, parallel computer network, torus Interconnect Networks.

1 Introduction

Large-scale massively parallel computers, Multiprocessors System-on-Chip (MPSoCs), and peer-to-peer communication networks provide high performance computing that allows users to deal with large and heavy computational tasks. Most of the contemporary massively parallel computers use wormhole switching (also widely known as wormhole routing [1, 2]) mechanism to support their interprocess communication. In the wormhole switching, a message is divided into a sequence of fixed-size units of data, called *flits*. If a communication channel transmits the first flit of a message, it must transmit all the remaining flits of the same message before transmitting flits of another message. The main drawback of wormhole switching is that blocked messages remain in the network, therefore they waste channel bandwidth and block other messages.

The routing algorithm specifies how a message selects a path to cross from the source to destination, and has great impact on network performance. Many practical systems have used deterministic routing [1, 2] with virtual channels to ensure deadlock avoidance. This is achieved by forcing messages to visit the virtual channels in a

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 606–613, 2007.
© Springer-Verlag Berlin Heidelberg 2007

strict order. This form of routing has the advantage of being simple; however, if any channel along the message path is heavily loaded, the message experiences large delays and if any node or channel along the path is faulty, the message cannot be delivered at all. Alternatively, adaptive routing improves both the performance and, more importantly, fault-tolerance of an interconnect network.

Message passing in massively parallel computers is implemented based on a routing algorithm that determines the path a message follows to reach its destination. Routing algorithms for these systems are generally classified as being either *deterministic* or *adaptive* [1]. In deterministic routing, all messages between a given source/destination pair will follow the same path. One of the main benefits is that in-order arrival of messages is preserved. However, deterministic routing usually makes an inefficient use of the network resources [1, 3]. An alternative, but complementary approach, to improve network performance consists of using adaptive routing. This routing strategy provides alternative paths to route messages, thus avoiding congested regions in the network and increasing throughput.

This paper proposes an analytical approach to investigate the performance behavior of a fully adaptive routing algorithm in wormhole-switched 2-D torus fortified with a routing scheme suggested in [4], as an instance of routing methodology widely used in the literature to achieve high adaptivity.

The rest of this paper is structured as follows. Section 2 briefly describes node structure, routing scheme, on which the present study is based. Section 3 gives an overview of the assumptions and describes the proposed analytical model. Section 4 compares the delays predicted by analytical model with those obtained through simulation experiments. Finally, Section 5 summarizes our findings and concludes the paper.

2 Node Structure in the Torus Networks

A k-ary 2-cube (2-D torus) is a direct network with $N= k^2$ nodes; k is called the radix. Links (channels) in the torus can be either uni- or bi-directional. Each node can be identified by a 2-digit radix k address (a_1, a_2). Nodes, with address (a_1, a_2), (b_1, b_2) are connected if and only if $a_1= (a_2 +1) \bmod k$ or $b_1= (b_2 +1) \bmod k$. In order to allow processors to concentrate on computational tasks and permit the overlapping of communication with computation, a *router*, is used for handling message communication among processors, and is usually associated with each processor. Consequently, each node consists of a Processing Element (PE) and router.

2.1 Adaptive Wormhole Routing in the Torus Networks

Recently, several fully adaptive routing algorithms on tori have been evaluated in [5] of which the one using Negative Hop-based (NHop) deadlock free routing augmented with a new idea called *bonus cards* (Nbc) has been shown to have the best performance. In [5], the Nbc routing scheme has been used in the context of Duato's methodology [6] and a routing algorithm, named *Duato-Nbc* with high performance and minimum virtual channel requirements was resulted. Investigations showed that Duato-Nbc has a better performance compared to other algorithms reported in the literature and the other algorithms proposed in [4].

3 The Analytical Model

In this section, we derive an analytical model for Duato-Nbc routing algorithm [5]. The most important performance measure in our model is the average message latency.

3.1 Model Assumptions

The model makes assumptions, which are commonly used in the literature [5, 7-11], and are listed below.

- Messages destinations are uniformly distributed across the network nodes.
- Nodes generate traffic independently of each other, following a Poisson process with an average rate of λ_g messages per cycle.
- The message length is fixed at M flits, each of which requires one cycle transmission time between two adjacent routers.
- The local queue at the injection channel in the source node has infinite capacity. Messages at the destination node are transferred to the local PE one at a time through the ejection channel.
- V virtual channels per physical channel are used. These virtual channels are used according to the Duato-Nbc routing scheme, which are divided into V_1 and V_2 classes ($V = V_1 + V_2$). This number of virtual channels yields the optimal performance compared to the other schemes proposed in [4].

In the following section, we present the mathematical model that approximates the behavior of 2-D torus communication system using Duato-Nbc routing algorithm.

3.2 Derivation of the Model

The average message latency is composed of the average network latency, $\overline{S}$, which is the time to cross the network and the average waiting time seen by the message in the source node, $\overline{W}_s$, before entering the network. However, to capture the effects of virtual channels multiplexing, the mean message latency has to be scaled by a factor, say $\overline{V}$, representing the average degree of virtual channels multiplexing, that takes place at a given physical channel. Therefore, the average message latency can be approximated as [9]

$$Average\ Message\ Latency = (\overline{S} + \overline{W}_s)\overline{V} \tag{1}$$

In what follows, we will describe the calculation of $\overline{S}$, $\overline{W}_s$, and $\overline{V}$.

3.2.1 Calculation of the Mean Network Latency

Under uniform traffic pattern, the average number of channels that a message visits along a given dimension and across the network, $\overline{k}$, $\overline{d}$ respectively, are given by Agarwal [10]

$$\overline{k} \approx k/4, \quad \overline{d} = 2\overline{k} \tag{2}$$

Fully adaptive routing allows a message to use any available channel that brings it closer to its destination resulting in an evenly distributed traffic rate on all network channels. The mean arrival rate, λ_c, on a given channel is determined as follows. A PE generates, on average, λ_g messages in a cycle, which are evenly distributed among the 4 output channels. Since each message travels, on average, $\bar{d}$ hops to cross the network, we can write the rate on a channel, λ_c, as

$$\lambda_c = \lambda_g \bar{d} / 4 \tag{3}$$

Since the torus topology is symmetric, averaging the network latencies seen by the messages generated by only one node for all other nodes gives the average message latency in the network. Let $s = (s_x, s_y)$ be the source node and $d = (d_x, d_y)$ denote a destination node such that $d \in G - \{s\}$ where G is the set of all nodes in the network. We define the set $H = \{h_x, h_y\}$, where h_x and h_y denote the number of hops that the message makes along X and Y dimensions, respectively, i.e. $(s_x + h_x) \bmod k = d_x$, and $(s_y + h_y) \bmod k = d_y$.

$$h_x = \left\| s_x - d_x \right\|, \quad h_y = \left\| s_y - d_y \right\| \tag{4}$$

where $\|x - y\|$ denotes the distance between a source node x and a destination node y. The network latency, S_H, seen by a message crossing the network from node s to node d consists of two parts: the first part is the actual message transmission time, and the second part is the blocking time in the network. Therefore, S_H can be written as

$$S_H = M + \left\| H \right\| + \sum_{h=1}^{\|H\|} B_h \tag{5}$$

where M is the message length, $\|H\|$ is the distance (in terms of hops made by the message) between the source and the destination nodes, and B_h is the blocking time seen by a message in its h^{th} hop. The terms $\|H\|$ and B_h are given by

$$\left\| H \right\| = h_x + h_y \tag{6}$$

$$B_h = P_{block_h} \bar{W}_c \tag{7}$$

where P_{block_h} is the probability that a message is blocked at its h^{th} hop channel, and $\bar{W}_c$ is the average waiting time to acquire a virtual channel in the event of blocking. Let us now calculate the probability P_{block_h}. To do so, let φ_h be the number of dimensions, or output channels, that a message still has to visit when crossing the h^{th} hop channel. The calculation of φ_h has been derived in [9]. We recollect briefly here the main equations for the calculation of φ_h. The number of channels, φ_h, that the message can select when crossing the h^{th} hop channel, $(1 \le h \le \bar{d})$ is given by

$$\varphi_h = \sum_{t=0}^{1} (2 - t)\, \psi_h^t \tag{8}$$

where ψ_h^t is the probability that the message has entirely crossed t ($0 \le t \le 1$) dimensions along on its h-hop path. The probability that there remains only one dimension to cross a message h-hops away from its destination, P_{φ_h}, is given by

$$P_{\varphi_h} = \begin{cases} \dfrac{2}{\bar{d} - h + 1} & \bar{k} \le h < \bar{d} - 1 \\ 0 & 0 \le h < \bar{k} \end{cases} \tag{9}$$

Thus, the probability that the message has entirely crossed t dimensions along on its h-hop path can be obtained by

$$\psi_h^t = \begin{cases} 1 - P_{\varphi_h} & t = 0 \\ P_{\varphi_h} & t = 1 \end{cases} \tag{10}$$

A message is blocked at a given channel when all the adaptive virtual channels and also $V_2 - \lceil \Delta/2 \rceil + 1$ virtual channels of deadlock-free class (which are used for Nbc algorithm) are busy, where Δ is the number of remaining hops to reach the destination. When blocking occurs a message has to wait for all V_1 fully adaptive virtual channels, and $V_2 - \lceil \Delta/2 \rceil + 1$ deadlock-free virtual channels. In order to calculate P_{block_h}, we need to categorize messages into three classes based on their previous and next hop.

- **Class 1:** This class contains messages which have used a virtual channel of fully adaptive class in their previous hop (with probability $\mathrm{P}_{\mathrm{block}_1}$).
- **Class 2:** This class contains messages which have used a virtual channel of deadlock-free class in their previous hop and their next hop is negative (with probability $\mathrm{P}_{\mathrm{block}_2}$).
- **Class 3:** This class contains messages which have used a virtual channel of deadlock-free class in their previous hop and their next hop is positive (with probability $\mathrm{P}_{\mathrm{block}_3}$).

Since the number of messages in Class 2 is identical to the number of messages in Class 3, we can write the probability of message blocking as

$$P_{block_h} = \left(P_{block1} + \left(P_{block2} + P_{block3} \right) / 2 \right)^{\varphi_h} \tag{11}$$

In what follows, we compute $\mathrm{P}_{\mathrm{block}_1}$, $\mathrm{P}_{\mathrm{block}_2}$, and $\mathrm{P}_{\mathrm{block}_3}$.

When a message used a virtual channel of fully adaptive class in its previous hop, it can use any of V_1 virtual channels of this class, and also $V_2 - \lceil \Delta/2 \rceil + 1$ virtual channels of deadlock-free class. The blocking occurs, when all $V_1 + V_2 - \lceil \Delta/2 \rceil + 1$ virtual channels at a given physical channel are busy. We can therefore write

$$\mathrm{P}_{\mathrm{block}_1} = \sum_{v=V_1+V_2-\lceil \Delta/2 \rceil+1}^{V} P_{v_0} P_v \binom{\lceil \Delta/2 \rceil - 1}{v - V_1 - V_2 + \lceil \Delta/2 \rceil - 1} \Big/ \binom{V}{v} \tag{12}$$

where P_v, ($0 \le v \le V$), and P_{v_0} denote the probability that v virtual channels at a physical channel are busy (which is calculated later in Section 3.2.3), and the probability

that the message uses a virtual channel of Class *a* in its previous hop, respectively. P_{v_0} is given by

$$P_{v_0} = V_1 \Big/ \left(V_1 + V_2 - \lceil \Delta/2 \rceil + 1\right) \tag{13}$$

Moreover, a message belonging to the Classes 2 and 3 which has already used virtual channel l at its previous hop, can employ any of V_1 virtual channels of fully adaptive class, and also $V_2 - \lceil \Delta/2 \rceil - l + 1$ and $V_2 - \lceil \Delta/2 \rceil - l + 2$ virtual channels of deadlock-free class, respectively. Therefore, the $\mathrm{P}_{\mathrm{block}_2}$ and $\mathrm{P}_{\mathrm{block}_3}$ are obtained by

$$\mathrm{P}_{\mathrm{block}_2} = \sum_{l=1}^{V_2 - \lceil \Delta/2 \rceil} \sum_{v=V_1+\lceil \Delta/2 \rceil + l}^{V} P_{v_l} P_v \binom{V_2 - \lceil \Delta/2 \rceil - l}{v - V_1 - \lceil \Delta/2 \rceil - l} \Big/ \binom{V}{v} \tag{14}$$

$$\mathrm{P}_{\mathrm{block}_3} = \sum_{l=1}^{V_2 - \lceil \Delta/2 \rceil + 1} \sum_{v=V_1+\lceil \Delta/2 \rceil + l + 1}^{V} P_{v_l} P_v \binom{V_2 - \lceil \Delta/2 \rceil - l - 1}{v - V_1 - \lceil \Delta/2 \rceil - l - 1} \Big/ \binom{V}{v} \tag{15}$$

where P_{v_l} indicates the probability that a message has used virtual channel l at its previous hop and is given by

$$P_{v_l} = 1 \Big/ \left(V_1 + V_2 - \lceil \Delta/2 \rceil + 1\right) \qquad 1 \le l \le V_2 - \lceil \Delta/2 \rceil + 1 \tag{16}$$

The average network latency, $\bar{S}$, is obtained by averaging S_H, the average network latency of H-hops messages, over the $(N-1)$ possible destination nodes in the network. Therefore, $\bar{S}$ can be determined as

$$\bar{S} = \frac{1}{N-1} \sum_{d \in G - \{s\}} S_H \tag{17}$$

To determine the average waiting time to acquire a virtual channel, a physical channel may be treated as an M/G/1 queue [8]. Since the minimum service time at a channel is equal to the message length, M, following a suggestion proposed in [11], the variance of the service time distribution can be approximated by $(\bar{S} - M)^2$; where $\bar{S}$ is the average service time at a given channel and can be calculated as the mean of S_H of all source and destination nodes that have at least one path between each other that traverse the channel. Hence, the average waiting time becomes

$$\bar{W}_c = \lambda_c \bar{S}^2 \left(1 + (\bar{S} - M)^2 / \bar{S}^2\right) \Big/ \left(2(1 - \lambda_c \bar{S})\right) \tag{18}$$

3.2.2 Calculation of the Average Waiting Time in the Source Node

A message originating from a given source node sees a network latency of $\bar{S}$. Since a message in the source node can enter the network through any of the V virtual channels, the average arrival rate to the queue is λ_g/V. Applying the Pollaczek-Khinchine (P-K) mean value formula [8] yields the average waiting time experienced by a message at the source node as [8]

$$\bar{W}_s = \left(\lambda_g / V\right)\bar{S}^2\left(1+(\bar{S}-M)^2/\bar{S}^2\right)\Big/\left(2\left(1-\left(\lambda_g/V\right)\bar{S}\right)\right) \tag{19}$$

3.2.3 Calculation of Average Multiplexing Degree of Virtual Channels

The probability, P_v ($0 \le v \le V$), that v virtual channels at a given physical channel are busy can be determined using a Markovian model (details of the model can be found in [7, 9]). In the steady state, the model yields the following probabilities [7].

$$Q_v = \begin{cases} 1 & v=0 \\ Q_{v-1}\lambda_c\bar{S} & 0<v<V \\ Q_{v-1}\lambda_c/(1/\bar{S}-\lambda_c) & v=V \end{cases} \qquad P_v = \begin{cases} \left(\sum_{v=0}^{V} Q_v\right)^{-1} & v=0 \\ P_{v-1}\lambda_c\bar{S} & 0<v<V \\ P_{v-1}\lambda_c/(1/\bar{S}-\lambda_c) & v=V \end{cases} \tag{20}$$

When multiple virtual channels are used per physical channel they share the bandwidth in a time multiplexed manner. The average degree of virtual channel multiplexing, that takes place at a given physical channel, can be estimated by [7]

$$\bar{V} = \sum_{v=1}^{V} v^2 P_v \Big/ \sum_{v=1}^{V} v P_v \tag{21}$$

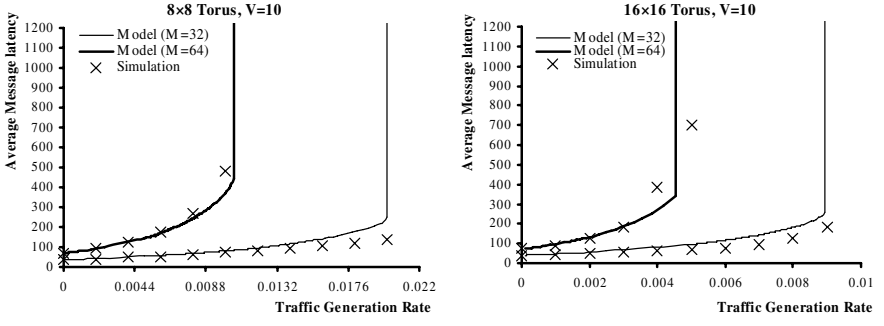


Fig. 1. Average message latency calculated by the analytical model against those obtained through simulation for 8×8 and 16×16 torus networks using Duato-Nbc routing with message length M=32, 64 flits, and V=10 virtual channels per physical channel

4 Simulation Results

To further understand and evaluate the performance issues of the routing algorithm, we have developed an event-driven simulator at flit-level. The average message latency is defined as the average amount of time from the generation of a message until the last data flit reaches the local PE at the destination. We ran each simulation for 300,000 flit times and sufficient warm-up times (through discarding the information obtained during the first 10,000 flit times) is provided to allow the network to reach the steady state. Numerous experiments are conducted for different sizes of the network, virtual channels, and message lengths to assess the accuracy of the analytical model. Fig. 1 depicts latency results predicted by the analytical model plotted against

those provided by the simulator for 2-D 8×8 and 16×16 torus networks, respectively; with V=10 virtual channels per physical channel, and different message lengths, M=32, 64 flits. The x-axis in this figure represents the traffic rate injected by a given node in a cycle (λ_g) while the y-axis shows the average message latency (in flit cycles). The figure indicates that the analytical model predicts the mean message latency with a good degree of accuracy in all regions. However, some discrepancies around the saturation point are apparent. This is a result of the approximations made when constructing the analytical model, e.g. the approximation used to estimate the variance of the service time distribution at a channel. This approximation greatly simplifies the model by avoiding the computation of the exact distribution of the message service time at a given channel.

5 Conclusions

In this paper, we proposed a mathematical performance model to predict the average message latency in wormhole-switched 2-D tori using the fully adaptive scheme proposed in [5]. Simulation experiments have revealed that the message latency results predicted by the analytical model are in good agreement with those obtained through simulation under different working conditions. We showed that the proposed model manages to achieve a good degree of accuracy while maintaining simplicity, making it a practical evaluation tool that can be used by the researchers in the field to gain insight into the performance behavior of fully adaptive routings in wormhole-switched torus networks. Our next objective is to develop an analytical modeling approach to investigate the performance behavior of this routing scheme in the presence of failures.

References

1. Duato, J., Yalamanchili, S., Ni, L.M: Interconnection networks: An engineering approach, Morgan Kaufmann Publishers, New York (2003).
2. Ni, L. M., McKinley, P. K.: A Survey Of Wormhole Routing Techniques in Direct Networks, IEEE Computer Society Press, 26 (2) (1993) 62-76.
3. Dally, W. J., Aoki, H.: Deadlock-free adaptive routing in multicomputer networks using virtual channels, IEEE TPDS, 4 (4) (1993) 66-74.
4. Boppana, R.V., Chalasani, S.: A Framework for Designing Deadlock-Free Wormhole Routing Algorithms, IEEE TPDS, 7 (2) (1996) 169-183.
5. Safaei, F., et al.: Performance Comparison Of Routing Algorithms In Wormhole-Switched Fault-Tolerant Interconnect Networks, International Conference on Network and Parallel Computing (NPC06), Japan, October 2006.
6. Duato, J.: A New Theory of Deadlock-Free Adaptive Routing in Wormhole Routing Networks, IEEE TPDS, 4 (12) (1993) 320-1331.
7. Dally, W. J.: Virtual Channel Flow Control, IEEE TPDS, 3 (2) (1992) 194–205.
8. Kleinrock, L.: Queuing Systems, Vol. 1, John Wiley, New York (1975).
9. Ould-Khaoua, M.: A Performance Model of Duato's Adaptive Routing Algorithm in k-ary n-cubes, IEEE TC, 48 (12) (1999) 1-8.
10. Agarwal, A.: Limits on Interconnection Network Performance, IEEE TPDS, 2 (4) (1991) 398-412.
11. Draper, J.T., Ghosh, J.: A Comprehensive Analytical Model for Wormhole Routing in Multicomputer Systems, JPDC, 32 (2) (1994) 202-214.

A Study on Phonemic Analysis for the Recognition of Korean Speech

Jeong Young Song[1], Min Wook Kil[2], and Il Seok Ko[3]

[1] Department of Computer Eng., Paichai University, 14 Yeonja 1Gil, DaeJeon, South Korea
jysong@pcu.ac.kr
[2] Dept. of Medical Inform., Mun Kyung College, HoGyeMyun, Mun Kyung, South Korea
mwkil@mkc.ac.kr
[3] Dept. of Computer and Multimedia., Dongguk University, 707 Seokjang-dong, Gyeongju, South Korea
isko@dongguk.edu

Abstract. This paper is a study related to phonemic analysis of Korean speech. For speech recognition, the best method is that speech should be recognized as phonemic unit. However, to segment speech into phonemic units is not well-applied because of various changes of phonation. Therefore, I in this study try to arrange a phonemic system of Korean speech and segment speech into phonemic units on the base of the phonemic system of Hunminjeongeum, the Korean script. I also look for the features of each phoneme by observing the changes of frequency domain, mel bandwidth and mel capstrum of phoneme and aim to form the phonemic segmentation system of Korean speech.

Keywords: phonemic analysis, pattern recognition.

1 Introduction

This paper is a study on phonemic segmentation of speech for recognition of Korean speech. A phoneme is the smallest unit to consist speech. Hence, speech recognition by phonemes is much more effective and efficient than recognition of syllable or word unit. It also can offer a basis of embodiment of the effective speech recognition system.

Therefore, I in this paper analyze Korean speech on phonemic system of Hunminjeongeum, the Korean script and also do a comparative analysis between features of these and frequency domain, Mel frequency band, Mel cepstrum. After that, I find out the rules that can separate Korean speech into phonemic unit by extracting phonemic features of Korean speech.

This paper analyzes the phonemic system of hunminjeongeum in chapter2 and comparative-analyzes speech into frequency domain, Mel frequency band, and Mel cepstrum according to the phonemic system of Hunminjeongeum. Chapter 4 shows the conclusions.

2 The System of Hunminjeongeum

Hunminjeongeum classifies speech into an initial sound, a medial vowel and a final consonant.

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 614–620, 2007.
© Springer-Verlag Berlin Heidelberg 2007

2.1 The System of Initial Sound

Hunminjeongeum is the phonemic letters for phonetic representation. Initial sounds are classified into entire-purity, second-purity, entire-impurity and no-purity-no-impurity like Table1. They are also consisted of the sounds from molar, tongue, lip, tooth, throat, anti-tongue, and half-tooth in according to the location of speech pronounced[1]. In modern phonetics, entire-purity is classified into common-sound, second-purity into an aspirate, entire-impurity into a fortis and no-purity-no-impurity into sympathetic sound[2].

Table 1. System of Initial Sound

Position / Manner	Back teeth	Tongue	Lips	Teeth	Throat	Semi tongue
Lax	ㄱ[g]	ㄷ[d]	ㅂ[b]	ㅅ[s] ㅈ[j]		
Aspirate	ㅋ[k]	ㅌ[t]	ㅍ[p]	ㅊ[ch]	ㅎ[h]	
Strong	ㄲ[kk]	ㄸ[tt]	ㅃ[pp]	ㅆ[ss] ㅉ[jj]		
Sonorant	ㅇ[ng]	ㄴ[n]	ㅁ[m]			ㄹ[l]

2.2 The System of Medial Vowel

A medial vowel is consisted of fundamental tone, conjunction, and wall in according to the shape of lips when they tune. There are intermediate vowels like [Ø],[r],[j],[i] which are developed in between initial sounds and medial vowels. And when [ ] and [∂] are united with these intermediate vowels, next things happen; [Ø]+[∂]→ , [r]+[∂]→ , [j]+[∂]→ , [i]+[∂]→[ㅣ]. that's how fundamental tones, [, , ㅣ], are established as Table2.

Table 2. System of Medial Vowel

fundamental tone	conjunction[w]	wall[ɥ]
	ㅗㅗ	ㅏ
	ㅜ	
ㅣ	/	/

If I indicate the thing that makes the shape of lips round as [w] and the thing that makes the shape of lips open as [ɥ] when they tune, two phonemes which are

[w]+[ ]→[ㅗ], [ɥ]+[ ]→[ㅏ] will be established. The following shows that another two phonemes which are established in [ ]; [w]+[ ]→[ㅜ][jwə], [ɥ]+[ ]→[ ][jɥə]. Then again, if [ㅗ, ㅏ, ㅜ,] are united with intermediate vowel [i], there will be phonemes created; [, , ,][1].

2.3 The System of Final Consonant

A final consonant is that consists the rhyme of letter from receiving an initial sound and a medial vowel. There are 8 of them; , , , ,ㅂ, , , . These are not much different from 7 of ending sounds of syllables treated as phonemics in modern phonology. Only the difference is that " " and " " can be treated the same in the final consonant.

3 Phonemic Segmentation

3.1 Conditions

This experiment did a sampling of speech with 16kHz, 16bit, and Mono by using a table microphone, found out the frequency through Fourier transform. And after transforming this into Mel frequency band, Mel cepstrum was gotten

3.2 Phonemic Segmentation

Speech was divided into initial sound, medial vowel and final consonant as Table3 and each of pictures shows spectrum per frequency(F/kHz), time of pronunciation(Second), Mel Frequency band(MF/melf), Mel Cepstrum(MC) of the range of 8kHz through Fourier transform. It finds out the sum of frequencies which are transformed into Mel frequency band and establishes the sound intervals. Again it was separated as a voiceless sound when high frequency was strong and it was separated as a voiced sound when low frequency was strong.

Table 3. Sections of Speech

sections	contents
initial sound	entire-purity, second-purity, entire-impurity, no-purity-no-impurity
medial vowel	fundamental tone, conjunction, wall
final consonant	, , , , , ,( ),

3.2.1 Segmentation of Initial Sound

An initial sound is divided up into entire-purity, second-purity, entire-impurity and no-purity-no-impurity like Figure 1 to Figure 4

Entire-purity of initial sound forms high frequency and has little spectrum of Mel frequency band, so it can be easily distinguished.

Second-purity(, , ㅛ, ,) of initial sound is an aspirated sound and forms comparatively longer time of high frequency. Especially " " has very strong frequency and little Mel spectrum, so it is easy to be distinguished

Entire-impurity(, , , ㅆ,) of initial sound is a fortis that except "ㅆ" has a very short time to pronounce. However, it rapidly changes the spectrum of vowel coming next.

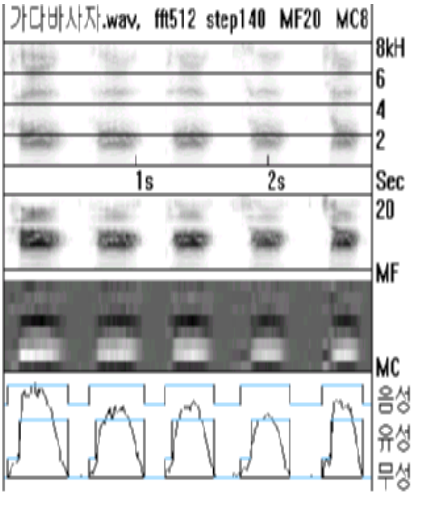


Fig. 1. Entire-Purity(, , , ,)

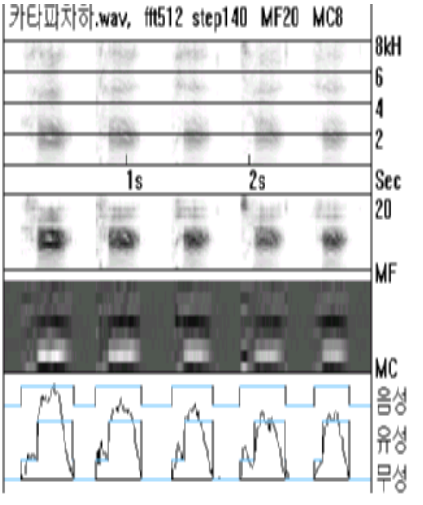


Fig. 2. Second-Purity , ,ㅛ,ㅊ,

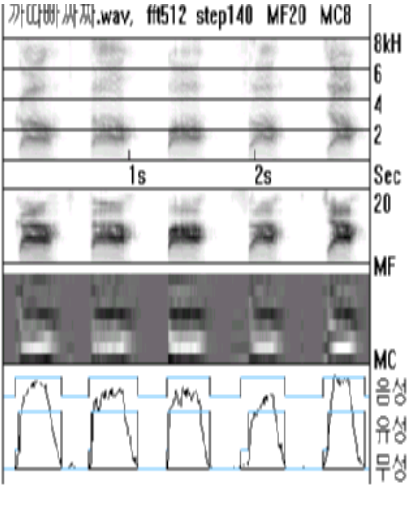


Fig. 3. Entire-Impurity (, , , ,ㅉ)

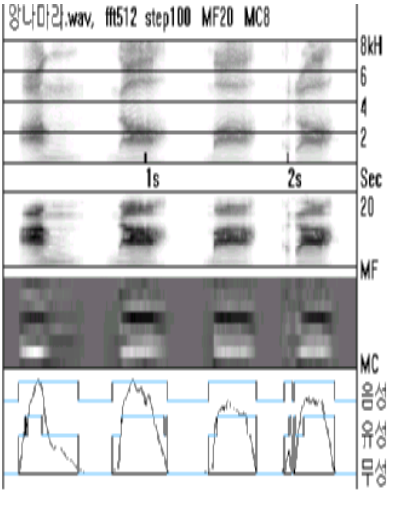


Fig. 4. No-Purity-no-impurity (, , ,)

No-purity-no-impurity(, , ,) of initial sound is a sympathetic sound. Here, " " cannot have a sound of initial sound, but a sound of final consonant. It has an echo after vowel. You can see that " , , " has a change of spectrum of vowel.

3.2.2 Segmentation of Medial Vowel

A medial vowel is divided into fundamental tone, conjunction in Figure 5 and wall in Figure 6. In Figure 5, "ㅗ,ㅜ", the conjunction of vowels, are the shapes added central formant to the spectrum of " ,ㅣ", fundamental tones, and again the spectrum gets changes when " / " start, then it gets stable again the forms of "ㅗ,ㅜ".

In Figure 6, "ㅏ, " among wall sounds keep steady spectrum, " / " gets change from the beginning of "ㅏ, " and again gets stable as forms of "ㅏ, ".

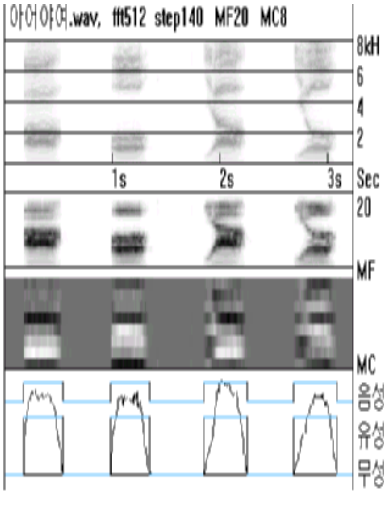


Fig. 5. Fundamental Tone and Conjunction **(, ,ㅗ,ㅜ,ㅛ/)**

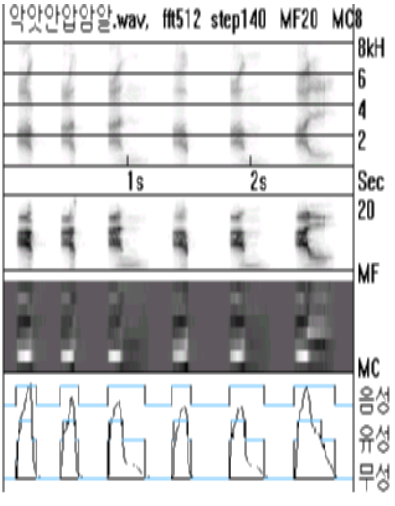


Fig. 6. Wall , ,

3.2.3 Segmentation of Final Consonant

Among 8 of final consonants, " , , , ,ㅂ, ,( ), ", if I treat " " and " " as the same sound and exclude " " shown in no-purity-no-impurity, then it will be as Figure 7.

Some of final consonant, " , , ", are easy to be distinguished because they form echo, but the other, " , , ", hardly show the form and happen to change the spectrum of vowels as shown in Figure 7.

3.3 Speech Experiment

Figure 8 is the pronunciation of " 녕 하 " about for a second. In Figure 8, vowels, (ㅏ,ㅕ,ㅏ, , ,ㅏ), are distinguished easily like frequency and Mel

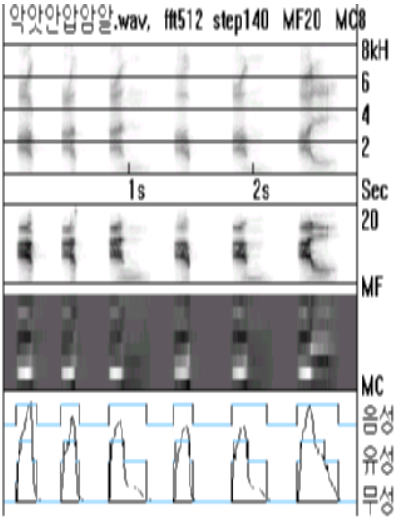


Fig. 7. Final Consonants , , , , , ,

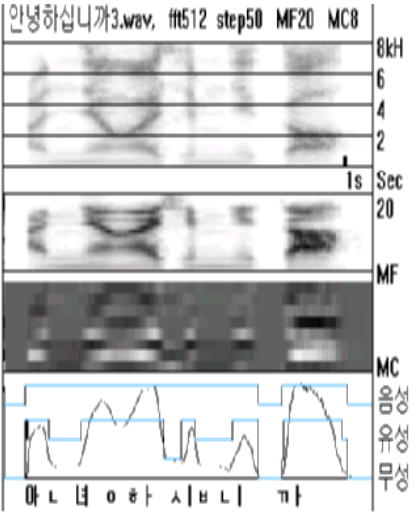


Fig. 8. An nyung ha sip ni kka

frequency band spectrum. " " should be sensed by change of spectrum and "ㅎ" is hardly pronounced and also gets change of spectrum like in Figure 4. " " is showing a certain high frequency as like in Figure 1 and " " also gets change of spectrum as like in Figure 3.

4 Conclusion

I in this paper analyze the features of Korean speech in according to the phonemic system of Hunminjeongeum, the Korean script and use these features to divide speech and feature of each phoneme by using the sum of Mel cepstrum through frequency domain, Mel frequency band and DCT(Discrete Cosine Transform) transform and spectrum transformed into Mel frequency band.

Phonemes of initial sound forms high frequency or changes the spectrum and medial vowels keep the steady spectrum after changing from the fundamental tones or change the spectrum. Vowels especially can be distinguished because they keep a certain form with spectrum. Sympathetic sounds of final consonants can be distinguished because they form echo but the others can be distinguished by sensing change of spectrum.

Therefore, if there is a further study on phoneme of speech that is weakened or omitted when the sound is prolonged, it will bring more accurate phonemic segmentation.

References

1. Younghuan O, " Management of voice data," Heungryung Science Publishers, (1998).
2. Juchae Bae, "Synopsis of Hangul phonology," Shingu Culture Publishers, (1997).

3. Lawrence Rabiner, "Fundamentals of Speech Recognition," Prentice Hall, (1993).
4. Jinsoo Han, "Management of Voice Signal," Osung Media, (2000).
5. Peter B. Denes, Elliot N Pinson, "The Speech Chain," Hanshin Culture Books, (1999).
6. Changyun Yu, "Notes on Hunminjungeum," Hyoungsul Publishers, (1998).
7. John Clark, Collin Yallop, "Phonetics & phonology" Hanshin Culture Books, (1998).
8. Byunggon Yang, "Theory & Practicals of Voice Analysis Using Praat," Mansu Publishers, (2003).
9. Sou-Kil Lee, Jeong-Young Song, "A Study on the Phonemic Analysis for Korean Speech," Proceeding of the Electronics, Information and Systems Conference Electronics, Information and Systems Society, I.E.E. of Japan, (2003)

Media Synchronization Framework for SVC Video Transport over IP Networks

Kwang-deok Seo[1], Jin-won Lee[1], Soon-heung Jung[2], and Jae-gon Kim[2]

[1] Computer and Telecommunications Engineering Division, Yonsei Univ., Gangwon, Korea
Kdseo@yonsei.ac.kr
[2] Broadcasting Media Research Group, ETRI, Daejeon, Korea
{zeroone,jgkim}@etri.re.kr

Abstract. This paper proposes an efficient media synchronization framework for SVC video transport over IP networks. To support synchronization between SVC video and audio signals transported over IP networks, RTP/RTCP protocol suite is usually employed. To provide a framework for media synchronization, we suggest an efficient RTP packetization mode and propose a computationally simple RTCP packet processing method. With the suggested RTP packetization mode, layer synchronization among scalable layers of the SVC video can be effectively achieved. Also by adopting the computationally simple RTCP packet processing, we do not need to process every RTCP SR packet for inter-media synchronization between video and audio signals.

Keywords: media synchronization, SVC, RTP, RTCP.

1 Introduction

SVC (Scalable Video Coding) known as scalable extension of H.264/MPEG-4 AVC is currently standardized in the JVT (Joint Video Team) of the ISO/IEC MPEG and the ITU-T Video Coding Experts Group [4]. SVC aims at achieving both high compression performance and adaptation for video delivery over heterogeneous networks. SVC is based on H.264/MPEG-4 AVC and provides three scalability modes including temporal, spatial, and quality scalability. Unlike the conventional scalability modes supported in the MPEG-2, H.263, and MPEG-4, SVC scalability can provide combined scalability mode in which the three scalability modes can be aggregated to a single SVC bitstream. The general concept for combining temporal, spatial, and SNR scalability in an SVC bitstream is illustrated in Fig. 1, which shows an example GOP structure for SVC with two scalable layers. The SVC bitstream contains two spatial layers: QCIF encoded at 15 fps and CIF encoded at 30 fps. Each spatial layer is composed of one base quality layer and one FGS layer for SNR scalability. The dotted arrow in Fig. 1 designates inter-layer prediction to remove redundancy between spatial layers. In temporal dimension, each picture belongs to one temporal layer indicated by the number in the middle of each picture.

To transport SVC video encapsulated in NAL (network abstraction layer) units over Internet Protocol (IP) in real-time, RTP (real-time transport protocol) and RTCP (RTP control protocol) are usually employed. RTP carries the payload with some

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 621–628, 2007.
© Springer-Verlag Berlin Heidelberg 2007

additional information like sequence number and RTP timestamp. RTCP serves for controlling quality of the transmitted data. RTP timestamp begins at a random number and its rate of increment is proportional to its sampling rate [5]. Thus, they do not directly give information on absolute time reference. To synchronize audio and video data, we need to utilize RTCP Sender Report (SR) packet to find out the absolute time information corresponding to each RTP timestamp carried by each RTP packet.

In this paper, we suggest an efficient RTP packetization mode for layer synchronization among scalable layers of SVC video and propose an efficient inter-media synchronization between SVC video and audio. In the proposed inter-media synchronization method, we do not need to process every RTCP SR packet for synchronization. Moreover, it does not require any floating-point operations or any divisions at all. The proposed method will be compared with conventional method [1].

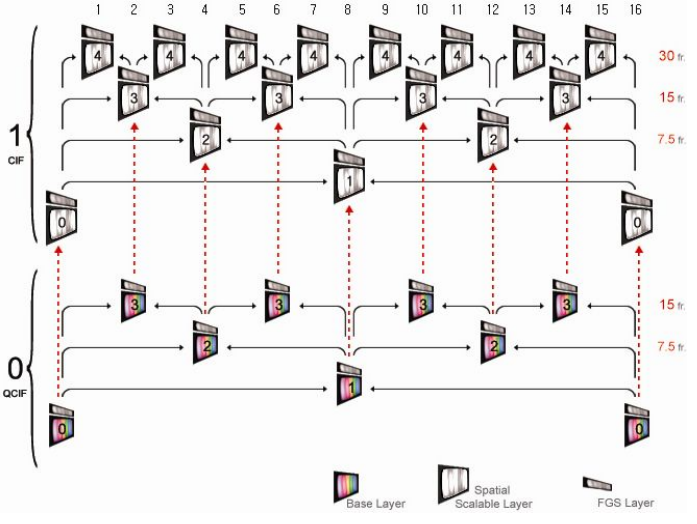


Fig. 1. Combined scalability in an SVC bitstream

2 Suggested RTP Packetization Mode for Layer Synchronization

The concepts of VCL (video coding layer) and NAL in SVC are inherited from H.264/MPEG-4 AVC. While the VCL creates a coded representation of the source content, the NAL encapsulates the data generated by the VCL and provides header information in a way that enables simple and effective customization of the use of the VCL for a broad variety of systems. For this purpose, the SVC NAL unit header contains the spatial, temporal, and quality coordinates of the NAL unit payload in the scalability cube which is used for identification and scaling operations of the NAL units. The NAL unit header is designed to co-serve as the payload header of an RTP payload format. For more details on the syntax and semantics of the SVC NAL unit header, you are referred to [3] and [4].

The Audio/Video Transport (AVT) Working Group of the IETF started in November 2005 to draft the RTP payload format for SVC and the signaling for layered

coding structures [3]. As SVC is a backward compatible extension of H.264, the same should be the case for its RTP packetization. In particular, it is possible to transport the base layer utilizing the same packetization scheme of RFC 3984 [2]. Thus, RFC 3984-aware legacy devices are still capable of utilizing an SVC base layer in an RTP transport environment.

An RTP stream carrying only one layer would carry NAL units belonging to that layer only. An RTP stream carrying a complete scalable video bitstream would carry NAL units of a base layer and one or more enhancement layers. In the former case, however, the system administrator of the server should open a separate UDP port for each RTP session to carry a single layer. Thus, the server should open as many ports as required to transport all the layers. System administrators would like to avoid opening too many UDP ports in their firewalls, because of the security risk and the administrative effort. Moreover, for mass deployment to end terminals, it is desirable to reduce the number of UDP ports in a firewall to the absolute minimum–ideally to a single one. In this respect, the latter approach is much preferred to the former one.

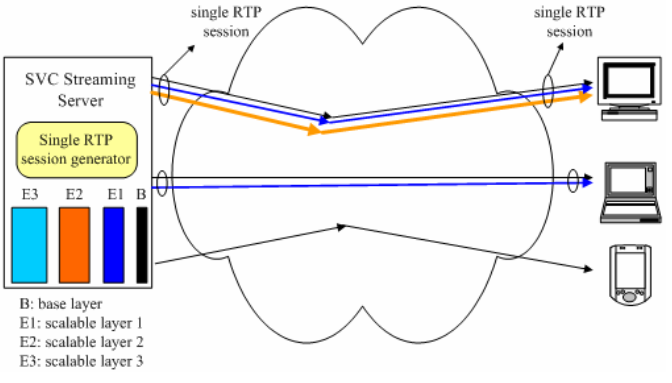


Fig. 2. SVC streaming scenario based on single RTP session

This line of thought leads to the service scenario as depicted in Fig. 2, where the server opens only a single RTP session to carry one or more layers. For each terminal, the server composes a bitstream tailored to the terminal's needs by aggregating NAL units of appropriate layers. 'Single RTP session generator' is used to aggregate the extracted contents from potentially more than one scalable layer into a single RTP stream carrying one or more layers.

In order to support the service scenario shown in Fig. 2, it is necessary to support encapsulating NAL units from multiple SVC layers into a single RTP packet in the payload format. The IETF specification on RTP payload format for SVC contains mechanisms such as STAP (single-time aggregation packet) and MTAP (multi-time aggregation packet) to aggregate more than one NAL unit into a single RTP packet, and another mechanism called FU (fragmentation unit) to split overly large NAL unit into multiple RTP packets [3]. Two fundamentally different packetization modes of operation are supported in [3]: non-interleaved mode and interleaved mode. Table 1 summarizes the allowed packet types for each packetization mode. In non-interleaved

mode, the NAL units should be aggregated in decoding order by adopting STAP-A, whereas in interleaved mode, NAL units belonging to multiple pictures can be aggregated out of decoding order by adopting STAP-B and MTAP. Non-interleaved mode is intended to avoid excessive RTP/UDP/IP header overhead that would result when encapsulating small NAL units in single NAL unit packets, whereas the interleaved mode provides an error resilience tool against burst errors. STAP-A aggregates NAL units with identical NALU-time, whereas MTAP aggregates NAL units with differing NALU-time. Here, NALU-time is defined as the value that the RTP timestamp would have if that NAL unit would be transported in its own RTP packet. In Fig. 1, pictures belonging to different spatial layer but having the same picture number (or display time) must have the same NALU-time. Thus, by adopting STAP-A, it is far more feasible to provide synchronization between pictures belonging to different spatial layers but with identical NALU-time. Therefore, Non-interleaved mode is more suitable for systems that require very low end-to-end latency and timely synchronization among NAL units from multiple SVC layers aggregated in a RTP packet. Furthermore, as shown in Table 1, only non-interleaved mode supports 'single NAL unit' type that can contain only a single NAL unit in the RTP payload. As a result, non-interleave mode can be suggested as a mandatory packetization mode for fast and real-time streaming requiring timely synchronization among SVC layers, and interleaved mode can be considered as an optional mode for error resilience by providing interleaving function against burst packet loss. In this paper, we employ non-interleaved mode to provide layer synchronization among different scalable layers.

Table 1. Allowed packet types for RTP packetization modes of SVC

NAL Unit Type	Packet Type	Single NAL Unit Mode	Non-interleaved Mode	Interleaved Mode
0	undefined	ignore	ignore	ignore
1~23	single NAL unit	yes	yes	no
24	STAP-A	no	yes	no
25	STAP-B	no	no	yes
26	MTAP16	no	no	yes
27	MTAP24	no	no	yes
28	FU-A	no	yes	yes
29	FU-B	no	no	yes
30~31	undefined	ignore	ignore	ignore

3 Conventional Inter-media Synchronization Method

The next problem to resolve for synchronization issue is providing inter-media synchronization between audio and SVC video. According to RFC 3550, it is stated that separate audio and video streams should not be carried in a single RTP session and demultiplexed based on the payload type or SSRC fields [5]. However, we cannot directly use RTP timestamp to synchronize data carried by different RTP sessions for

the following two reasons. Firstly, RTP timestamp should be initialized to random offsets at session startup to minimize the risk of breaking encryption. Secondly, RTP timestamp increases in proportion to the sampling rate of media. Usually the sampling rates of audio and video data are quite different. Thus, the rates of increase in RTP timestamp for audio and SVC video sessions are not the same. To circumvent these problems, RTCP SR packets carrying both the RTP and the NTP timestamp are generally employed.

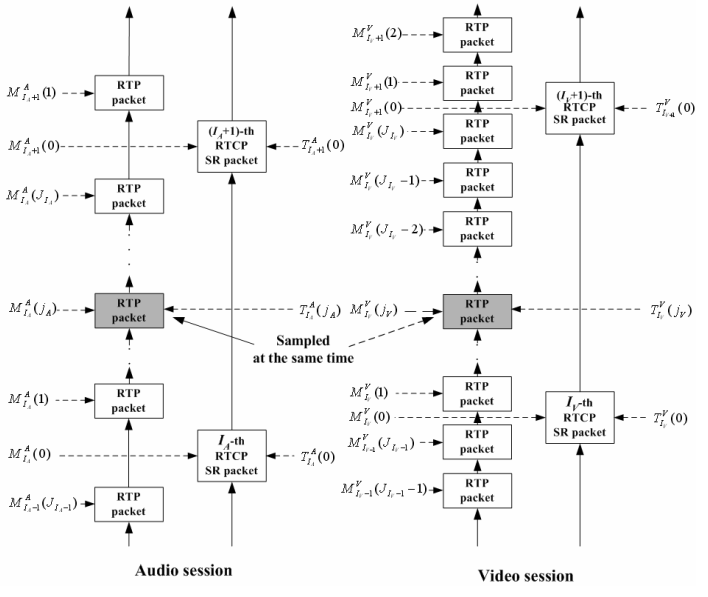


Fig. 3. RTP/RTCP streams for audio and SVC video sessions

Fig. 3 shows RTP and RTCP streams for audio and SVC video sessions. In Fig. 3, each RTP packet of SVC video session aggregates NAL units that all share the same NALU-time. The RTP timestamp of each RTP packet must be set to the NALU-time of all the NAL units to be aggregated. An aggregation packet can carry as many NAL units as necessary. However, the total amount of data in an aggregation packet obviously must fit into an IP packet, and the size should be chosen so that the resulting IP packet is bound by the MTU size of the transport channel.

The superscripts *A* and *V* are used to denote audio and SVC video sessions, respectively. For the derivation of the relationship between the RTP timestamp of a specific RTP packet and the absolute time reference, let us consider the shaded RTP packet in the audio session shown in Fig. 3. For this RTP packet, $T^A_{I_A}(j_A)$ is the absolute time reference for RTP timestamp of the j_A-th RTP packet after the I_A-th RTCP packet has been received. NTP (Network Time Protocol) tells us how to set the absolute time information. As a special case, when $j_A = 0$, $T^A_{I_A}(0)$ is the NTP timestamp contained in the I_A-th RTCP packet. Similarly, $M^A_{I_A}(j_A)$ is the RTP timestamp contained in the j_A-th

RTP packet after the I_A-th RTCP packet, and $M_{I_A}^A(0)$ is the RTP timestamp for the I_A-th RTCP packet.

Let us assume that the shaded RTP packet in the SVC video session is sampled at the same time with the shaded RTP packet in the audio session. If the absolute time reference of this SVC RTP packet is represented by $T_{I_V}^V(j_V)$, it is required that $T_{I_A}^A(j_A) = T_{I_V}^V(j_V)$ for perfect synchronization. However, the transmission rates of RTP packets are normally not the same for different sessions. Moreover, RTCP packets for each session may be transmitted at different time. Thus, even if $T_{I_A}^A(j_A) = T_{I_V}^V(j_V)$, I_A and j_A of the audio session may not be equal to I_V and j_V of the SVC video session, respectively. Based on this fact, we can compute $T_{I_A}^A(j_A)$ of a RTP timestamp by using $M_{I_A}^A(j_A)$. $M_{I_A}^A(0)$ and $T_{I_A}^A(0)$ values are also used in the computation, which can be obtained by the I_A-th RTCP packet. In the method proposed in [1], the absolute time reference $T_{I_A}^A(j_A)$ is obtained by

$$T_{I_A}^A(j_A) = T_{I_A}^A(1) + \sum_{k=2}^{j_A} \frac{\Delta M_{I_A}^A(k)}{R^A}, \tag{1}$$

where R^A is the sampling rate of audio data. $T_{I_A}^A(1)$ is obtained by

$$T_{I_A}^A(1) = T_{I_A}^A(0) + \frac{M_{I_A}^A(1) - M_{I_A}^A(0)}{R^A}. \tag{2}$$

In (1), $\Delta M_{I_A}^A(k)$ is the difference between the RTP timestamps of two adjacent RTP packets and is given by

$$\Delta M_{I_A}^A(k) = M_{I_A}^A(k) - M_{I_A}^A(k-1). \tag{3}$$

Computation of (1) and (3) continues until a new RTCP packet is received. After receiving the (I_A+1)-th RTCP packet, (1) and (2) are computed again using $T_{I_A+1}^A(0)$ and $M_{I_A+1}^A(0)$ carried by this RTCP packet. When computing $T_{I_A}^A(j_A)$ by (1) in Bertoglio's method, they do not compute the term $\sum_{k=2}^{j_A} \frac{\Delta M_{I_A}^A(k)}{R^A}$ directly. Instead, since they already know the value of $T_{I_A}^A(j_A - 1)$, they compute the value by

$$T_{I_A}^A(j_A) = T_{I_A}^A(j_A - 1) + \frac{\Delta M_{I_A}^A(j_A)}{R^A}. \tag{4}$$

The same procedure (1)-(4) can be applied to the SVC video session to obtain $T_{I_V}^V(j_V)$.

When processing j_A–th RTP packet for the audio session and j_V–th RTP packet for the SVC video session, Bertoglio's decision rule based on $T_{I_A}^A(j_A)$ and $T_{I_V}^V(j_V)$ for synchronization is as follows:

$$\begin{aligned}
&T_{I_V}^V(j_V) - T_{I_A}^A(j_A) > \eta_+ && \text{: SVC video is ahead of audio,} \\
&\eta_+ \geq T_{I_V}^V(j_V) - T_{I_A}^A(j_A) \geq -\eta_- && \text{: audio and SVC video are in synch,} \\
&T_{I_V}^V(j_V) - T_{I_A}^A(j_A) < -\eta_- && \text{: audio is ahead of SVC video,}
\end{aligned} \tag{5}$$

where η_+ and η_- are thresholds used for boundaries of the in-sync region. To apply this decision rule, it is evident that we need to inspect every RTCP packet for the computation of (1) through (5).

4 Proposed Inter-media Synchronization Method

Now we will derive the proposed scheme from the conventional method described by (1)-(5). In this study, we exploit the fact that after a connection has been setup, the codec type and the sampling rate are usually sustained during the connection.

By canceling out each term in the computation of $\sum_{k=2}^{j_A} \frac{\Delta M_{I_A}^A(k)}{R^A}$ in (1) and by using (2), we can simplify (1) into the following form

$$\begin{aligned} T_{I_A}^A(j_A) &= T_{I_A}^A(1) + \frac{M_{I_A}^A(j_A) - M_{I_A}^A(1)}{R^A} \\ &= T_{I_A}^A(0) + \frac{M_{I_A}^A(j_A) - M_{I_A}^A(0)}{R^A}. \end{aligned} \tag{6}$$

Assuming that R^A is kept as a constant, we can obtain (7) from the NTP and RTP timestamps carried by the *0*-th and the I_A-th RTCP packet as follows

$$R^A = \frac{M_{I_A}^A(0) - M_0^A(0)}{T_{I_A}^A(0) - T_0^A(0)}. \tag{7}$$

Rearranging (7) for $T_{I_A}^A(0)$ and substituting it into (6) yields

$$T_{I_A}^A(j_A) = T_0^A(0) + \frac{M_{I_A}^A(j_A) - M_0^A(0)}{R^A}. \tag{8}$$

Similarly, we can apply this procedure to obtain the relation (9) for the RTP stream of SVC video session.

$$T_{I_V}^V(j_V) = T_0^V(0) + \frac{M_{I_V}^V(j_V) - M_0^V(0)}{R^V}. \tag{9}$$

By subtracting (9) from (8) and applying this result to (5), we can derive the following compact decision rule (10) after some arithmetic,

$$\begin{aligned} &d > \eta_0 + R^A R^V \eta_+ && \text{: SVC video is ahead of audio,} \\ &\eta_0 + R^A R^V \eta_+ \geq d \geq \eta_0 - R^A R^V \eta_- && \text{: audio and SVC video are in sync,} \\ &d > \eta_0 - R^A R^V \eta_- && \text{: audio is ahead of SVC video,} \end{aligned} \tag{10}$$

where the variable *d* and the threshold constant η_0 are defined by

$$d = R^A M_{I_V}^V(j_V) - R^V M_{I_A}^A(j_A), \tag{11}$$

$$\eta_0 = R^A R^V (T_0^V(0) - T_0^A(0)) + R^V M_0^A(0) - R^A M_0^V(0). \tag{12}$$

Note that we only need to compute *d* by (11) when examining the synchronization of each pair of RTP packet by (10). The reason is that $\eta_0 + R^A R^V \eta_+$ and $\eta_0 - R^A R^V \eta_-$ in (10) are needed to be computed just once after receiving the first RTCP packet. Since all of the R^V, R^A, $M_{I_V}^V(j_V)$, and $M_{I_A}^A(j_A)$ values in (11) are fixed-point numbers themselves, there is no need to utilize floating-point operations at all. Obviously, this is a

clear advantage for embedded processors, which usually do not have floating point units. Moreover, (11) does not require any division operations like the case of (1)-(4). For ARM processors, avoiding division is a great advantage, since they do not have any hardware divider. In [1], they reported that truncation round-off errors may accumulate, since their method is computed by repeated divisions and summations. However, it is obvious that there are no possibilities of error accumulation in the computation of (11). Note that only two fixed-point multiplications and one subtraction are needed for the computation of d in (11).

5 Conclusions

In this paper, we addressed the problem of synchronization for SVC video transport over IP networks. The synchronization issue includes layer synchronization among scalable layers of SVC video and inter-media synchronization between audio and SVC video. It has first been discussed non-interleaved mode of RTP packetization is suitable to provide layer synchronization among scalable layers of SVC video. Then, we proposed a computationally simple RTCP packet processing for inter-media synchronization. The advantageous aspects of the proposed method can be summarized in three ways. First, the decision rule is far simpler than the conventional method. Second, it does not require RTCP SR packet processing for synchronization except first RTCP packet. Finally, the proposed method does not suffer from the accumulation of round-off errors that are inherent in the conventional method.

Acknowledgements. This work was supported in part by the Gangwon-Alberta Research Collaboration Fund.

References

1. Bertoglio L., and Migliorati P.: Intermedia synchronization for video conference over IP. Signal Processing: Image Communication, Vol. 15, No. 1, (1999) 149-164.
2. Wenger S., Hannuksela M., Westerlund M., Singer D.: RTP payload format for H.264 video. RFC 3984, IETF, (Feb. 2005).
3. Wenger S.: RTP payload format for SVC video. IETF Internet Draft: draft-wenger-avt-rtp-svc-03.txt, (Oct. 2006).
4. Reichel J., Schwarz H., and Wien M.: Scalable video coding- Working Draft 3. JVT-P201, Poznan, Poland . (July 2005,)
5. Schulzrinne H., Casner S., Frederick R., and Jacobson V.: Real-time transport protocol. RFC 3550, IETF, (July 2003).

Developing Value Framework of Ubiquitous Computing

Jungwoo Lee, Younghee Lee, and Jaesung Park

Graduate School of Information, Yonsei University
Sudaemun Gu, Shinchon Dong 134, Seoul, Korea
{jlee,rarayes}@yonsei.ac.kr, tomypink@hotmail.com

Abstract. In this study, a value framework for ubiquitous computing is developed and presented. Using 'value focused thinking' approach suggested by Keeney, twenty-two potential users of ubiquitous computing were interviewed and 435 statements describing values expected from ubiquitous technology were obtained from these interviews. Subsequent purification and redundancy removal reduces these 435 statements into 166 objectives that these users have in their mind when thinking about ubiquitous computing. These 166 objectives were factorized into 35 objectives through clustering by three experts. These 35 objectives were classified into a means-ends network diagram by analyzing reciprocal relationships among them in a focus group activity of another three experts. Resulting means-ends network reveals a value framework inherent in users' perception about ubiquitous computing technologies. This framework will be useful as a reference in developing new business models for ubiquitous computing as well as developing useful technologies themselves.

Keywords: Ubiquitous Computing, user's values, value-focused thinking, value framework, means-ends network.

1 Introduction

Since Mark Weiser termed 'ubiquitous' as a new paradigm for computing in 1988, efforts have been concerted to develop and advance information technologies towards 'connecting, invisible calm and silent, and real' ubiquitous computing [1]. Recently, efforts have been concerted to develop ubiquitous technologies such as RFID and to standardize interfaces and communications such as EPC Networks, in order to expedite the adoption of ubiquitous computing in the real world. Ubiquitous computing is now considered to be the technology critical not only for the support of businesses, but also as an enabler of strategic positioning of businesses [2, 3].

However, despite good progresses made in the ubiquitous technology itself and high public expectations generated from popular movies and dramas, practitioners are still struggling over what are the appropriate business models which may provide incredible value as a new ubiquitous businesses [4]. In this regard, questions still remained unanswered about real and practical values provided by ubiquitous computing. What are the real and practical offerings of ubiquitous computing to users and businesses?

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 629–635, 2007.
© Springer-Verlag Berlin Heidelberg 2007

In this context, this research is purported to elicit values expected from the users' side about how ubiquitous computing technology would benefit their life and work, and present these values as a value framework so that further technological and business development can be assessed against this value framework.

2 Research Method

This study is methodologically based on 'value focused thinking (VFT)' approach [5] This VFT approach consists of four steps designed to elicit and frame values: (1) conduct interviews and construct a list of what they want in the decision context, (2) convert these statements into a common format of objectives (an object and a preference), and (3) purify and cluster these objectives into a more general and simpler set of fundamental objectives and means objectives, and (4) establish means-ends network of these objectives so that fundamental objective values can be identified with instrumental values to leading to these fundamental values.

In the context of this research, this four-step method is applied in order to assess values attached, by users, to ubiquitous computing. Twenty two users were interviewed. Twelve were graduate students while ten were practitioners in the field. At the beginning of each interview, three scenarios of ubiquitous environment were explained in detail: a news service on a personal digital assistant, an intelligent refrigerator informing the owner about the status of food in it, and an automatic gate keeper recognizing authorized persons automatically. Subjects were already familiar with these scenarios through public media. Table 1 summarizes details of the research methodology.

Table 1. Research methodology

Question	What are the values of ubiquitous computing to users?
Strategy	Case study.
Paradigm	Interpretive (value focused thinking).
Data collection	Dialogue; in-depth interviews.
References	Keeney [5-7], Torkzadeh [8]
Informants	Users of ubiquitous computing
Type of results	Wish list, means-fundamental objectives network, in-depth predictive description.

Construct a wish list from each interview. Users' values were found out by a dialogue with triggering questions. As, in many cases, users' values are hidden under the surface, Keeney recommends several stimulation techniques to surface these latent values. We chose a combination of two techniques: wish list and probing. First, each interviewee was asked to express what benefits and services that they expect from ubiquitous computing technologies, as many as possible. Whenever subjects are having problem articulating what they want, interviewer posed probing questions mostly by adding why and how to the statements that they were making. Twenty two

interviews generated 298 initial wish statements and 137 additional statements triggered by probing questions, totaling 435 statements.

Convert statements into objectives. These statements are converted into objectives, using an action verb plus an object so that directions of wanted changes and target of change are more clearly and directly stated. Some statements were compound sentences, which contain more than one objective and these statements were divided into simpler ones. Some statements were ambiguous and in which case, interviewee were recalled and asked triggering questions again through phone calls. In this clarification process, two researchers were involved in reviewing each item on the list independently. Through this review process, a list of 166 objectives was finalized.

Purify and cluster these objectives into a simpler set. In using ubiquitous computing, users wanted to achieve these 166 objectives. However, these objectives are not properly articulating values yet, and also duplication can be found. The objectives were then categorized in order to surface the latent meanings and more general values attached to ubiquitous computing. Categorization by three expert focus group activities produced 35 value statements as shown in table 2.

Table 2. Means Objectives and End Objectives

Ends Objectives	
(*Overall*) Maximize the value of individual	
Secure privacy & security	Minimize cost
Increase safety	Maximize personal freedom
Increase personalization /individualization	Increase utility
	Increase portability
Maximize ease of use	
Means Objectives	
Perform tasks without concerning environment	Learn my preferences
	Increase anonymity
Prevent missing information	Make the environment in the way I want
Receive preferable information	
Receive information appropriate for the context	Authenticate users
	Provide multi-purpose functions
Lower the device price	Minimize efforts on my side
Lower the usage fee	Support where I want
Make devices more robust	Support when I want
Make devices easy to use	Provide choice for personal preference
Make devices with minimum options	
Allow free entry and exit from ubiquitous environment	Support personal privacy
	Provide educational impact
Increase the variety of alternatives	Provide accurate service for me
Help personal management	Save time
Increase planning ability	Support decision making

Establish means-ends network. As a next step of framing values out of the list of objectives, 35 objectives were classified into two categories by analyzing their precedence relations: means objectives and fundamental objectives. The criterion of classification is whether an objective is an intermediate one, i.e. is it a means to achieve another objective or is it a fundamental one in assessing ubiquitous computing. The means objectives, a total of 27 along with 8 fundamental objectives are presented in table 2, respectively. Finally, all objectives are organized in a means-ends relationship diagram using directional relationships. This value framework is presented in Figure 1.

3 Discussion

Figure 1 can be interpreted as a framework of values users attach to ubiquitous computing. Objectives on the left side represents means objective which are considered instrumental to achieve ends objectives presented on the right side of the diagram. Among nine ends objectives, "maximize value of individual" seems to be the final fundamental value to which other ends objectives are instrumental to achieve.

From the close inspection of the framework, two points emerges. First, it seems that these objectives can be easily classified using 'five any' typology, which is very popular in ubiquitous computing literature – anybody, any service, any place, any time, and any device [9]. As shown on the left side of figure 1, first three objectives such as 'increase anonymity,' 'help personal management,' and 'authenticate users,' belong to people (anybody) category. As these objectives were collected from users in a bottom-up manner, this fitting confirms that user value system (or the mental model) is familiar with the theoretical model proposed by most ubiquitous computing literature.

Second point is about the progression from left side of the diagram towards the right side: from general to specific or from supply-side view towards demand side view. For example, 'increase anonymity' relates to general notions of anybody while 'secure privacy and security' relates to protection of my rights. It seems that fundamental objectives are mostly concerned with personal specifics while the left-most means objectives refer to more "general" values.

In other words, the perspectives on ubiquitous computing taken by general side are focused on such issues as anybody, anytime, anywhere, any service, and any device and emphases are on: to reach anybody, on the network, any time, anywhere, through any device for any service. However, when you follow the arrow in the framework more toward the fundamental objectives, these values are more concerned with 'to me,' 'right now,' 'here,' 'whatever I need,' and 'on whatever I have.'

These differences in perspective may play an important role in determining requirements for ubiquitous computing business model. For example, 'anybody' requirement can be translated into service universally standardized to different users from the supply-side ("any" side) view, but from the demand-side ("users' side"), it would better be translated as personally customizable service that can be accessed universally. It seems clear that the developers of ubiquitous computing and application need to reconcile these differences in values in order to develop successful technological artifacts which may generate business value which users could buy in.

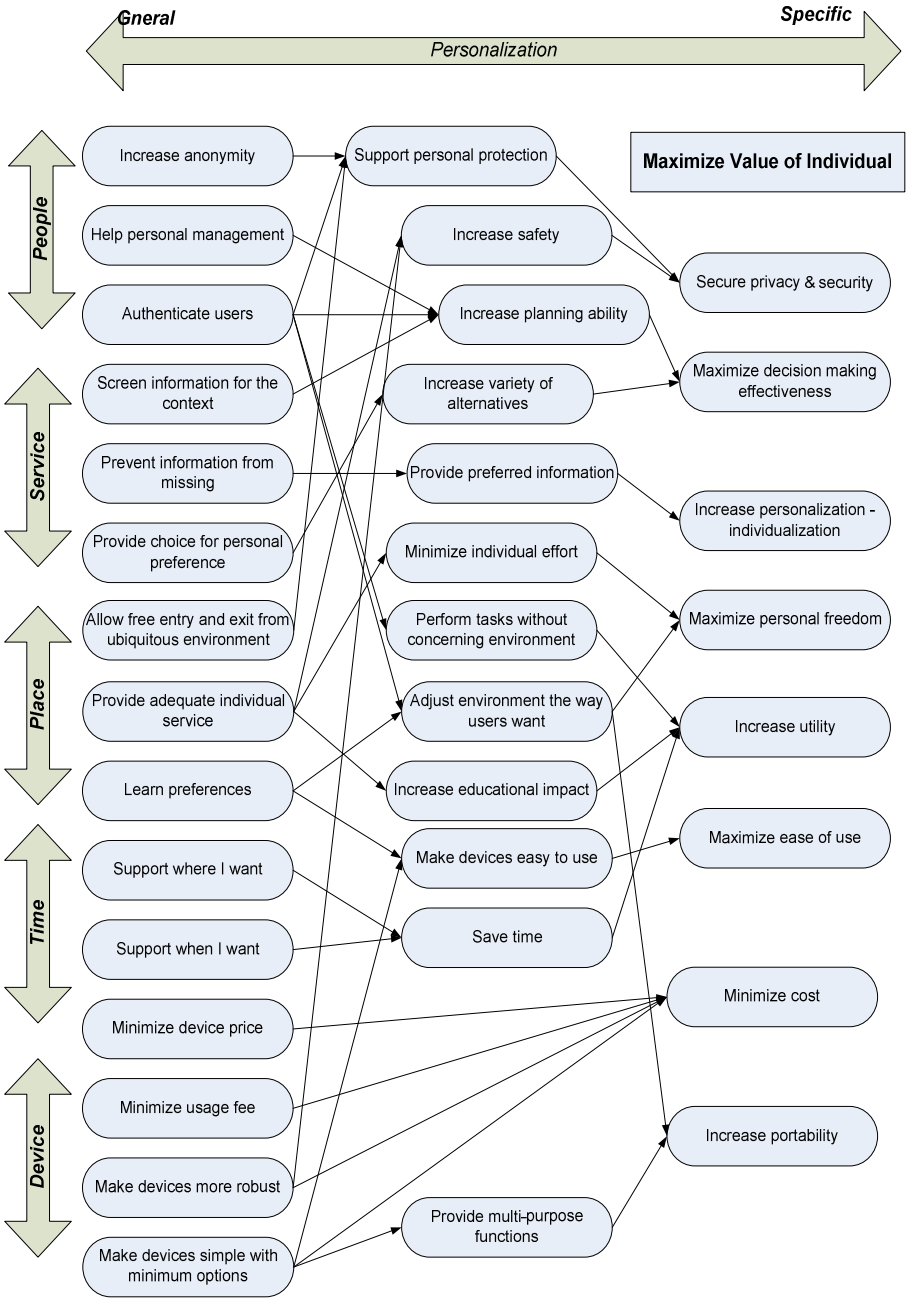


Fig. 1. Means-Ends Network of Values in Ubiquitous Computing

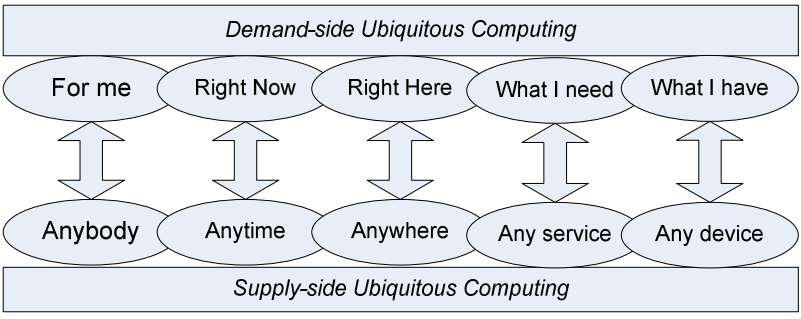


Fig. 2. Reconciliation of Perspectives on Ubiquitous Computing

4 Conclusion

Using the notion of ascribing users' values to ubiquitous computing, we are able to develop a value framework that may provide good referential guidelines for technology development as well as new business development with regards to ubiquitous computing. Research presented in this paper identifies a value framework and various propositions that will be a precursor to measure the value of ubiquitous computing, especially in service delivery to users. This instrument is under development.

The value framework proposed here suggests that the progression from supply-side value to demand-side value is an essential problem to be analyzed and solved for successful business model development for ubiquitous computing as well as further development of ubiquitous technology itself.

This study is still in progress. Contradiction underlying the assumptions and belief may deter killer applications to be developed for future ubiquitous computing. It seems important to articulate these contradictions and reconcile them as ubiquitous computing advances. Contradiction dimensions needs further development with supporting theory as well as practical examples. Case studies and scenario analyses would be the next step in revealing implications of these contradictions and interdisciplinary theoretical development should accompany phenomenological observations and experiments.

References

1. Weiser, M.: The computer of the 21st centry. Scientific American. Vol. 265. No. 3 (1991) 66-75
2. Krishnan, A., S. Jone.: Time Space: activity-based temporal visualization of personal information spaces. Personal Ubiquitous Computing. Vol. 9. No. 1 (2005) 45-65
3. May, A.J., Ross, T., Bayer, S.H., Tarkiainen, M. J.: Pedestrian navigation aids: information requirements and design implications. Personal Ubiquitous Computing. Vol. 7. No. 6 (2003) 331-338
4. Tamminen, S., Oulasvirta, A. Toiskallio, K. Kankainen, A.: Understanding mobile context. Personal Ubiquitous Computing. Vol. 8. No. 2 (2004) 135-143

5. Keeney, R.L.: Value-Focused Thinking: A Path to Creative Decision-making. Cambridge, Massachusetts: Harvard University Press (1992)
6. Keeney, R.L.: Creativity in Decision Making with Value -Focused Thinking. Sloan Management Review. Vol. 35. No. 4 (1994) 33-41
7. Keeney, R.L.: The Value of Internet Commerce to the Customer. Management Science. Vol. 45. No. 4 (1999) 533-542
8. Torkzadeh, R., Dhillon, G.: Measuring factors that influence the success of Internet commerce. Information Systems Research. Vol. 13. No. 2 (2002) 187-204
9. Weiser, M., Brown, J.S.: Designing calm technology. PowerGrid, Vol. 1.01 (1996)

An Enhanced Positioning Scheme Based on Optimal Diversity for Mobile Nodes in Ubiquitous Networks*

Seokyong Yang and Sekchin Chang

Dept. of Electrical and Computer Engineering, University of Seoul, Seoul, Korea
{syuun,schang213}@uos.ac.kr

Abstract. A lot of schemes have been proposed for the realization of u-city. Especially, most of the schemes are based on ubiquitous networks. For various u-city services to be available, mobile nodes can be employed in ubiquitous networks. However, an accurate positioning scheme is required for the practical usage of the mobile nodes in the ubiquitous networks. In this paper, an optimal diversity technique is proposed to improve the positioning performance. Simulation results indicate that using the optimal diversity scheme, the positioning performance can considerably be enhanced for mobile nodes in ubiquitous networks.

1 Introduction

A lot of attention has recently been paid to the establishment of the services such as intelligent disaster prevention, intelligent building management, intelligent health care, intelligent traffic control, and so on in metropolitan cities [1]. The services can be converged more efficiently and can be accessed more easily via ubiquitous networks [2]. The city which can offer the services based on ubiquitous networks is generally called u-city. Usually, the wireless sensor node is considered a realistic basis for the implementation of ubiquitous networks [3]. When the wireless sensor node is equipped with mobility, the node acts as a mobile node in ubiquitous networks. For the practical realization of the mobile node, the sensor node algorithm can be implemented in a cellular modem [4]. Using the implementation approach, the design limitation such as low cost and low power [3, 5] can be overcome since the node is able to utilize the resources of the cellular phone. Moreover, once the wireless sensor node is implemented in the cellular modem, the u-city services can easily be offered to the cellular phone holder in ubiquitous networks. However, an efficient positioning is an inevitable requirement for the practical use of the mobile node in ubiquitous networks. As proposed in [4], it is assumed in this paper that the positioning approach consists of location detection and location tracking. In this paper, an optimal diversity scheme is proposed to improve the positioning performance. The optimal scheme is based on multiple antennas. Especially, a novel structure and new

* This work was supported by Smart (Ubiquitous) City Consortium under Seoul R&BD Program.

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 636–643, 2007.
© Springer-Verlag Berlin Heidelberg 2007

message format are presented to efficiently implement the multiple antennas and to effectively acquire the optimal diversity, respectively. Simulation results indicate that the proposed diversity scheme can significantly improve the positioning performance for mobile nodes in ubiquitous networks.

2 The Efficient Structure for Mobile Nodes with Multiple Antennas

For the realistic implementation of the mobile node, the sensor node module can be included in a cellular modem [4]. In addition, the mobile node employs multiple antennas to achieve the optimal diversity. Fig. 1 illustrates the structure of the cellular modem to include the mobile node with multiple antennas [6]. Since most of the recent cellular modems are based on the system-on-chip

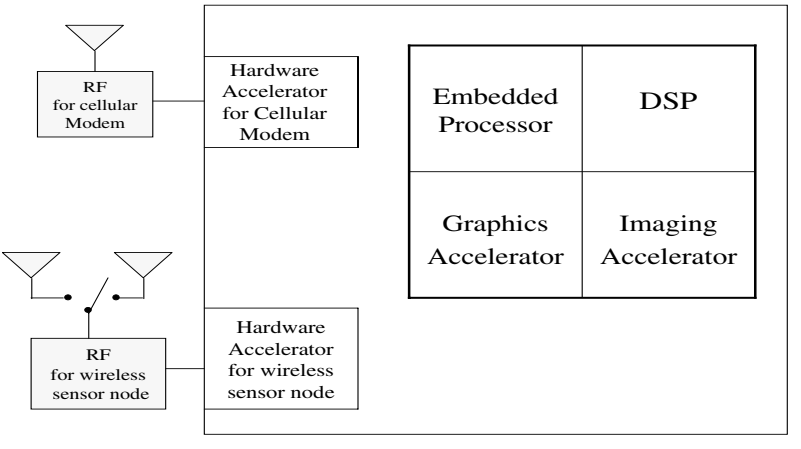


Fig. 1. The modem structure for the mobile node with diversity

(SoC) technology in the design, the cellular modem mainly consists of embedded processor, embedded DSP, hardware accelerators, graphics and imaging accelerators as shown in Fig. 1. In the view of such SoC structure, the hardware accelerator can be considered a kind of peripheral module. This indicates that the SoC structure relatively easily allows the addition of the hardware accelerators with low complexity. Therefore, the sensor node algorithm can be added in the cellular modem as a simple hardware accelerator since the algorithm exhibits relatively low complexity [7]. As illustrated in Fig. 1, the multiple antennas can be utilized to acquire the optimal diversity gain for the positioning of the mobile node. As indicated in [6], the modem structure utilizes only one RF module for two antennas in Fig. 1, which significantly alleviate the hardware overhead for the multiple antennas. Therefore, an efficient antenna switching and corresponding message format are also required for the optimal diversity combining.

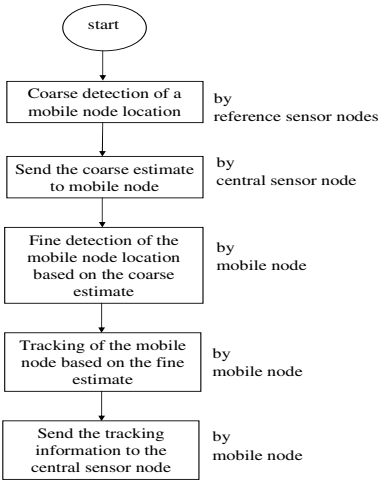


Fig. 2. The efficient positioning scheme for mobile nodes

3 The Enhanced Positioning Scheme Based on Optimal Diversity for Mobile Nodes

The mobile node can afford various u-city services to the phone user in ubiquitous networks. However, the location of the mobile node should accurately be detected and tracked to fully utilize the functionality of the node because the node can move in ubiquitous networks. Fig. 2 depicts the efficient positioning scheme for mobile nodes [4]. The positioning scheme consists of coarse and fine location detection, and location tracking in Fig. 2. As indicated in the figure, the reference sensor nodes roughly detect the location of the target mobile node. As the coarse detection, the trilateration scheme [8] is usually utilized by the reference sensor nodes which are selected from general wireless sensor nodes. Since the coarse detection usually exhibits simple computation, the general sensor nodes can perform the detection without high power consumption. However, the fine detection and the location tracking require highly intensive computation, which usually leads to high cost and high power consumption. Therefore, the target mobile node should perform the detection and the tracking by itself since the node can fully exploit the resources of the cellular phone. In the fine detection, the mobile node can increase the accuracy of the coarse estimate using an optimization technique such as the steepest descent algorithm [9]. Usually, the kind of algorithm exhibits fairly good performance when the initial value is close to the optimal value. If the coarse estimate can be used as the initial value, the algorithm can generate a fairly accurate estimate. For the use of the coarse estimate as the initial value in the fine estimation, the coarse estimate should be transmitted to the target mobile node by a central sensor node which is selected from the reference nodes. However, the received estimate usually includes the error due to the channel fading, which degrades the performance of

the fine location detection. To overcome the performance degradation, the optimal diversity can be utilized in the mobile node. For the mobile node to achieve the optimal diversity gain on the modem structure of Fig. 1, the central node should transmit the coarse estimate using the proposed message format which is illustrated in Fig. 3. As shown in the figure, the message format includes

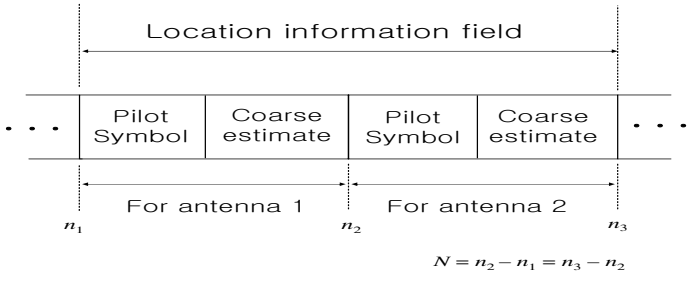


Fig. 3. The message format for the optimal diversity of mobile nodes

location information field which consists of the same two coarse estimation values and two pilot symbols. The first pilot symbol and the second pilot symbol are located at the discrete time-index of n_1 and n_2, respectively in the figure. As depicted in Fig. 3, the total time length for pilot symbol and coarse estimate is N. When the central sensor node transmits the coarse estimates in the message format, the received coarse estimate signal at the k^{th} antenna $r_k(n)$ is expressed as

$$r_k(n) = \alpha_k \cdot s(n) + \eta_k(n) \tag{1}$$

where $s(n)$ denotes the transmitted coarse estimation signal, and α_k and $\eta_k(n)$ indicate the channel parameter and additive white Gaussian noise (AWGN), respectively at the k^{th} antenna. In (1), the flat fading is considered the channel effect. In addition, in (1) $k = 1$ when $n_1 \leq n < n_2$, and $k = 2$ when $n_2 \leq n < n_3$. In other words, the mobile node receives the first coarse estimate at the 1^{st} antenna, and then switches to the 2^{nd} antenna and receives the second coarse estimate at the 2^{nd} antenna. Then, the received coarse estimate signals are combined as follows:

$$r(n) = w_1 \cdot r_1(n) + w_2 \cdot r_2(n) \tag{2}$$

where w_k denotes an optimal weighting coefficient for the k^{th} antenna. For the optimal diversity gain in (2), the weighting coefficients, w_1 and w_2 are determined to maximize the signa-to-noise ratio (SNR) of $r(n)$. Using Cauchy's inequality, the SNR of $r(n)$ can be expressed as

$$\frac{\sigma_s^2 |\sum_{k=1}^{2} w_k \alpha_k|^2}{\sigma_n^2 \sum_{k=1}^{2} |w_k|^2} \leq \frac{\sigma_s^2 \sum_{k=1}^{2} |w_k|^2 \sum_{k=1}^{2} |\alpha_k|^2}{\sigma_n^2 \sum_{k=1}^{2} |w_k|^2} = \frac{\sigma_s^2}{\sigma_n^2} \sum_{k=1}^{2} |\alpha_k|^2 \tag{3}$$

where σ_s^2 and σ_n^2 denote signal power and noise power, respectively. Using the condition for the equality in (3), w_k is determined as

$$w_k = \alpha_k^* \tag{4}$$

where $*$ denotes the complex conjugate. As indicated in (4), the channel parameters need to be estimated for the optimal coefficients. The channel estimation can be performed using the pilot symbols in the message format of Fig. 3. From (2) through (4), the weighing coefficients of (4) maximizes the SNR of the combined signal $r(n)$ of (2), which leads to the optimal diversity combining. Since the signal with the maximal SNR usually causes the best receiver performance, the mobile node can achieve the highest probability for decoding the received coarse estimate signal correctly using the proposed optimal diversity technique. This surely increases the estimation accuracy in the location detection because the detection searches the fine location position based on the decoded coarse estimate value. Therefore, it is concluded that the mobile node can enhance the estimation performance of the location detection using the optimal diversity.

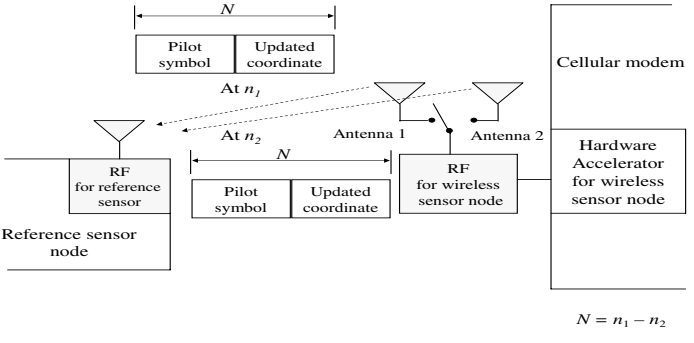


Fig. 4. The transmission of the updated estimate using optimal transmitter diversity in mobile nodes

If the mobile node is in motion, the fine estimate should continuously be updated in the location tracking as follows [4]:

$$x^{(n+1)} = x^{(n)} + \hat{l} \cdot \triangle, y^{(n+1)} = y^{(n)} + \hat{m} \cdot \triangle \tag{5}$$

where $x^{(n)}$ and $y^{(n)}$ are the updated coordinates of the mobile node at time index of n. In (5), $\hat{l}$ and $\hat{m}$ are the integer indices to indicate the motion direction and the motion amount of the mobile node. In addition, $\triangle$ denotes the coordinate incremental for tracking. Since high computation is usually required to determine $\hat{l}$ and $\hat{m}$, the mobile node also performs the location tracking by itself. Therefore, the mobile node should send the updated coordinate back to the central node whenever the update of (5) occurs. For the central node to decode the received

coordinate more correctly, the optimal transmitter diversity can be utilized in the mobile node as shown in Fig. 4. In other words, the mobile node transmits the updated coordinate and the pilot symbol through the antenna 1 at the time index of n_1, and then switches to antenna 2 and transmits the same coordinate and the pilot symbol through antenna 2 at the time index of n_2. In Fig. 4, N denotes the total time length of the updated coordinate data and the pilot symbol, and is defined as $n_2 - n_1$. If the central node combines the received coordinate signals which are also weighted by the optimal coefficients of (4), the probability for correctly decoding the updated coordinate signal can be increased as in the case of the fine detection. Note that the pilot signals are also used for the channel estimation in the central node.

In the optimal diversity scheme, the additional overhead mainly includes simple addition and multiplication as shown in (2), which is minor in the system complexity. In addition, since the channel estimation is usually performed in most wireless systems [10], the additional complexity is not significant in the determination of the weighting coefficients of (4).

4 Simulation Result

Simulation results exhibit the effectiveness of the proposed optimal diversity scheme. For the simulation, the environment of Fig. 5 [4] is considered. This figure illustrates the moving of the mobile node from region 1 to region 2. In addition, 2.4 GHz flat fading and additive white Gaussian noise (AWGN) are assumed as the channel environments for the simulation.

For the performance evaluation of the location detection based on the optimal diversity scheme, the mean squared error (MSE) values between exact and estimated location position are given in Fig. 6. As shown in Fig. 6, the location detection with the optimal diversity achieves the gain of about 1 dB over the location detection without the diversity. For investigating the effects of the optimal transmitter diversity in the location tracking, Fig. 7 exhibits the updated coordinates that the central node decodes after receiving the coordinate signals

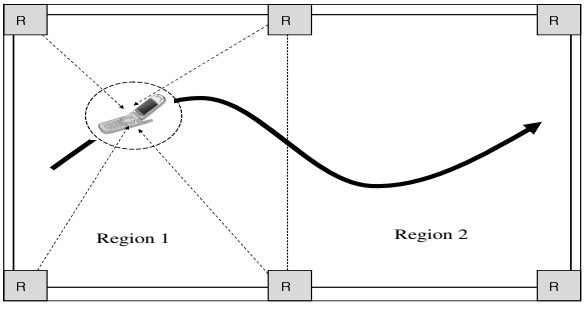


Fig. 5. The mobile node in ubiquitous networks

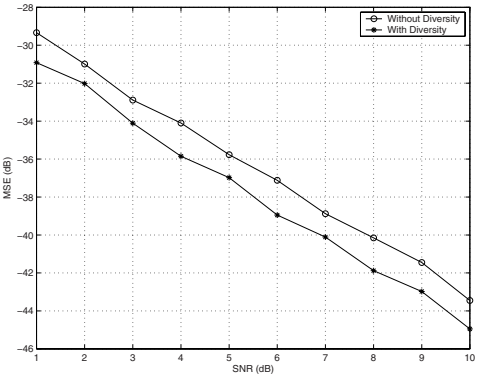


Fig. 6. The MSE performances for the location detection

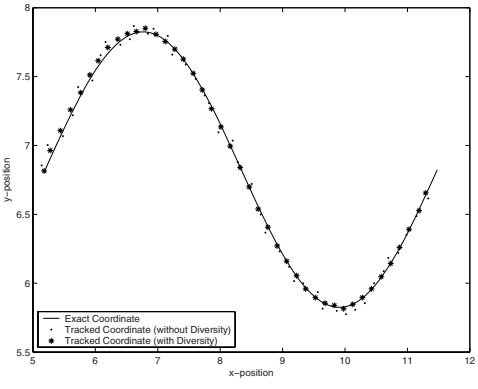


Fig. 7. The updated coordinates that the central node decodes

from the mobile node in motion under the environment of Fig. 5. As illustrated in Fig. 7, the decoded coordinates with the optimal diversity are much closer to the exact coordinates than those without the diversity.

5 Conclusion

In this paper, an optimal diversity scheme is proposed to improve the performances in locating and tracking the mobile node under ubiquitous networks. To acquire the diversity, the mobile node performs the antenna switching between two antennas, which considerably reduces the complexity overhead due to multiple antennas. For the performance enhancement in the location detection, the mobile node performs the optimal diversity combining based on the proposed message format. In addition, for more reliable transmission of the updated coordinates in the tracking, the mobile node utilizes the optimal transmitter diversity.

The simulation results also reveal that the proposed diversity scheme can significantly enhance the detection and the tracking performances. For our further research, we will evaluate our proposed scheme in the real environments.

References

1. I. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, Wireless Sensor networks: A Survey. Journal of Computer Networks **38** (2002) 393–422
2. N. S. Correal, and N. Patwari, Wireless Sensor networks: Challenges and Opportunities. Proc. of Virginia Tech Symp. Wireless Personal Comm. (2001) 1–9
3. E. H. Callaway Jr., Wireless Sensor networks: Architectures and Protocols. Auerbach. (2003)
4. P. Kim and S. Chang, An intelligent positioning scheme for mobile agents in ubiquitous networks for u-city. KES AMSTA 2007
5. C. M. Cordeiro and D. P. Agrawal, Ad Hoc & Sensor Networks: Theory and Applications. World Scientific. (2006)
6. J. Lee and S. Chang, An Intelligent diversity scheme for accurate positioning of mobile agents for u-city. KES AMSTA 2007
7. IEEE Std 802.15.4: Wireless medium access control (MAC) and physical layer (PHY) specifications for Low-Rate Wireless Personal Area Networks (LR-WPANs). (2003)
8. N. Patwari, J. N. Ash, S. Kyperountas, A. O. Hero III, R. L. Moses, and N. S. Correal, Locating the nodes: cooperative localization in wireless sensor networks. IEEE Signal Processing Magazine (2005) 54–69
9. S. Haykin, Adaptive Filter Theory, 4th Edition. Prentice-Hall. (2001)
10. G. L. Stüber, Principles of Mobile Communication, Second Edition. Kluwer Academic Publishers. (2001)

TOMOON: A Novel Approach for Topology-Aware Overlay Multicasting

Xiao Chen, Huagang Shao, and Weinong Wang

Regional Network Center of East China,
Department of Computer Science and Technology,
Shanghai Jiao Tong University, China
{shawn,hgshao,wnwang}@sjtu.edu.cn

Abstract. Most existing overlay multicast approaches avoid considering any network layer support no matter whether it is available or not. This design principle greatly increases the complexity of the routing algorithms and makes the overlay topologies incompatible with the underlying network. To address these issues, we propose TOMOON as a novel overlay multicast approach, which exploits the cooperation between end-hosts and IP multicast routers to construct a topology-aware overlay tree. Through a little modification to PIM-SM, a multicast router is able to receive registration from nearby group members and redirect passing-by join requests to them. Due to the multicast router's support, TOMOON organizes its group members into an overlay multicast tree efficiently, which matches the physical network topology well. Since we don't depend on routers to multicast data packets, the overlay tree can also cover the unicast networks and thus has a good support for the heterogeneous network environment.

1 Introduction

Most of the current overlay multicast solutions (e.g. Narada[3], Nice[2]) share some common drawbacks. First, they can't adapt the overlay to the physical network very well. Thus it is hard to guarantee the cost-efficiency of data transmission in all cases. Second, they force end-hosts to assume too much responsibility in overlay construction, which makes the protocol too complicated to serve as a general platform for other applications. Last, since the recovery from overlay partition can be very slow, it is hard to maintain a stable performance in case of node failures. All these issues result from a common design principle. That is to remain the network layer as simple as possible and refuse to consider any support from the network no matter whether it is available or not.

Our contribution in this paper is to address all these issues by proposing a topology-aware overlay multicast approach, which is supported by IP multicast networks. The key point of our approach is to adopt the overlay network for actual data transmission while exploiting IP multicast routers to optimize the overlay construction. This method has several advantages over traditional overlay multicast schemes. With support of the multicast-enabled network layer, the

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 644–651, 2007.
© Springer-Verlag Berlin Heidelberg 2007

task of overlay topology optimization can be drastically simplified. The complexity of the dynamic membership management is also reduced to a large extent. Meanwhile, end-host based multicast eliminates most of the problems related to IP multicast. Since the multicast routers may be deployed incrementally in our approach, the overlay can easily scale up to support large-sized group communications among heterogenous networks, where not all routers are multicast-capable. Besides the scalability issue and the address collision, overlay multicast also provides a viable solution to the commercial issue, which is difficult in IP multicast. Therefore the two multicast techniques, which seem to contradict at a first glance, are merged to produce an efficient multicast framework in this paper. Hereinafter, we name it by TOMOON, which stands for *Topology-aware Overlay Multicast Over heterOgeneous Networks*.

The rest of the paper is organized as follows: In Sect.2, we define the network model for TOMOON and analyze the mechanism of IP multicast. The detailed protocol is explained in Sect.3, which describes our modification to the PIM-SM[5] protocol and the tree construction algorithm. The performance of TOMOON is evaluated through simulations in Sect.4. Finally we draw conclusions in Sect.5.

2 Network Models

2.1 Physical Network Model

Without loss of generality, we model the underlying network in question by a directed graph G(V, E) with the vertices representing routers or hosts and the edges standing for links. A host is attached to only one router while a router's node degree is more than one. Each edge is assigned a weight to indicate the link's cost or latency. We assume that each node here belongs to an autonomous system (AS) that is either a multicast domain or a unicast domain. All routers of a multicast domain support intra-domain IP multicast. To make the following description easier, we now define some concepts, which will be used throughout the paper.

Definition 1. *A path from node A to node B, denoted by $P(A, B)$, is a sequence of routers comprising a shortest path from A to B according to the underlying unicast routing protocol. $P(A, S)$ is also called a request path of node A if node S is a data source. Since a unique data source is usually presumed, the request path of A is also denoted by $RPath(A)$.*

Definition 2. *Node B is close to the request path of node A if at least one node of $RPath(A)$ is within a certain distance from node B and they are located in the same multicast domain. This distance is also called a capture range, which is determined by the physical network settings around node B.*

An outline of the physical network model is shown in Fig. 1. In this figure, all router nodes are marked with capital letters while the gray nodes represent end hosts. Suppose that node s is the unique source and all other hosts are receivers

of the same group. We use the dotted lines to indicate the request paths of $RPath(a)$ and $RPath(b)$, which originate from a and b respectively and point to node s. If the capture range is just one hop, only c and g are close to $RPath(b)$. Since $e1$ and $e2$ are located in a unicast domain, they are not close to $RPath(b)$ even if they are just one hop away from the router E.

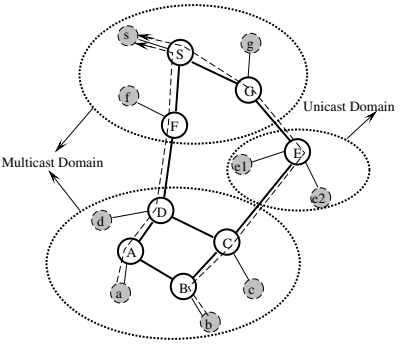


Fig. 1. Physical network model

2.2 Introduction to PIM-SM

Multicast forwarding entry. From a multicast router's perspective, the task of any routing protocol is to build a routing table with forwarding entries for all groups. A forwarding entry for the PIM-SM multicasting is denoted by a quaternary tuple of (S,G,*iif*,*oif*), which applies to all packets originated from source S and destined for group G. The other two fields, namely *iif* and *oif*, refer to input interface list and output interface list respectively. The construction of such a forwarding entry and its usage is explained as follows.

Control Path. For simplicity, we only consider the source specific multicast in PIM-SM. Here PIM-SM routers have to deal with two protocols in order to create the forwarding entries for a (S,G) pair. First, the designated router of the receiver will receive an IGMP[1] source specific membership report from the host, which carries the (S,G) information. Then it turns this membership report into a PIM-SM join request and unicasts it towards the source S. All routers receiving this request should build an entry like (S,G,*iif*,*oif*). The input interface for receiving the IGMP membership report or the PIM-SM join request is added to *oif* while the output interface for sending the join request is added to *iif*. In this way, the multicast forwarding entry is installed on all PIM-SM routers involved in the group session, which form a shortest path tree from the sender to all receivers.

Data Path. Once the shortest path tree is established, the forwarding entry comes into play for packet multicasting. Whenever a router on the distribution

tree receives a multicast packet sent from source S to group G, it will first check whether the input interface of this packet is exactly the one in *iif*. If the packet does come in through the correct interface, a copy of it is sent out through each interface listed in *oif*. The forwarding process is done in a top-down manner from the source's designated router to all leaf routers on the tree.

3 Details of TOMOON

3.1 Multicasting Join Requests

Multicast-capable routers play an important role in our overlay tree construction. Briefly speaking, TOMOON depends on those routers on the new joiner's request path to multicast the join request towards their respective neighbor members so that the new member can find a suitable parent from the tree quickly. While a PIM-SM router is totally qualified to assume such a task from a functional perspective, no existing protocol is designed for this purpose. Therefore we need to introduce some modifications to PIM-SM so that the routers will behave as we expect them to.

Multicast Forwarding Entry. One basic function of the multicast forwarding entry is to determine whether a packet belongs to a specific group so that it can be multicasted accordingly. Unlike the situation in source specific IP multicast, where all data packets of the same group have a common source-destination pair like (S,G), the join requests for the same group session originate from different joiners and thus have different source addresses. Since TOMOON is designed for single source multicast, we will aggregate the join requests of the same session by their destination address/port pair like (S,P). S is in fact the address of the overlay tree's root. The inclusion of the port number P allows for multiple overlay trees rooted at the same node. Therefore the multicast forwarding entry for multicasting the join requests is a ternary tuple like (S, P, *oif*). The field *oif* has a similar meaning as we described previously. Now we are going to explain the installation and application of these entries on routers.

Membership Registration. In order to capture the wanted messages on nearby routers, an in-tree member M sends a membership report toward the tree root with a limited value of TTL. Due to the small TTL, only those routers close to M will receive it. Once a router R has received such a report, which indicates a (S, P) pair, it will initiate a forwarding entry building process, which is described in detail by the pseudo-code routine in Fig. 2.

Forwarding Join Request. The process of multicasting join requests is straightforward once the forwarding entry is calculated as mentioned above. When a join request destined for the session (S, P) arrives at a router where a corresponding forwarding entry is installed, a copy of it is sent through each of the interfaces in *oif* except the input interface of the request. If the *oif* field only has two interfaces and the request is received from one of them, then the

```
Procedure ReceiveReportFor (Address M, Address S, Port P)
   //M-Address of the sender of the report
   //S-Address of the tree root
   //P-Port number identifying the group session
   Var
     Entry e;
     //e-Multicast forwarding entry;
   Begin
     If RoutingTable.includeEntryOf(S, P) Then
        e = RoutingTable.fetchEntry(S, P);
        e.addToOif(InterfaceTo(M));
     ElseIf InterfaceTo(M)!=InterfaceTo(S) Then
              e = new Entry (S, P);
              e.addToOif(InterfaceTo(M));
              e.addToOif(InterfaceTo(S));
              RoutingTable.addEntry(e);
     EndIf
   End.
```

Fig. 2. Algorithm for building forwarding entries

transmission will reduce to a unicast forwarding. This stipulation ensures that all members registered at the routers will receive exactly one copy of the request during the multicasting.

3.2 Tree Construction and Maintaining

Member Joins. Suppose that each new member J knows the address/port pair (S,P) of the data source before its participation. Then a request packet is generated with (S, P) as its destination address/port pair. As we mentioned above, if the join request traverses the capture ranges of other in-tree members, the data source S along with those members close to the request path will be informed of this new joiner. Each of them will respond to J as long as they have sufficient residual fan-out degrees to accept a new child in the overlay tree. Each response indicates a candidate parent. The next step is to choose a best one from them as J's formal parent in the tree.

The specific criteria for choosing a best parent node could vary from one application to another. For simplicity, we order that the candidate closest to the joiner become its parent. If multiple candidates are qualified, the one closest to the data source is selected. This can be done by using tools like *ping* or *traceroute*. Meanwhile, a list of the candidate parents is kept for later use.

Topology Adjustment. As far as tree-based multicasting is concerned, a key reason for topology adjustment is that some in-tree members would like to be connected to newcomers rather than their current parents. This idea is explained in Fig. 3. In this figure, node d is close to the request path of node a. However, node a joins the tree before node d, which leads to a tree structure like

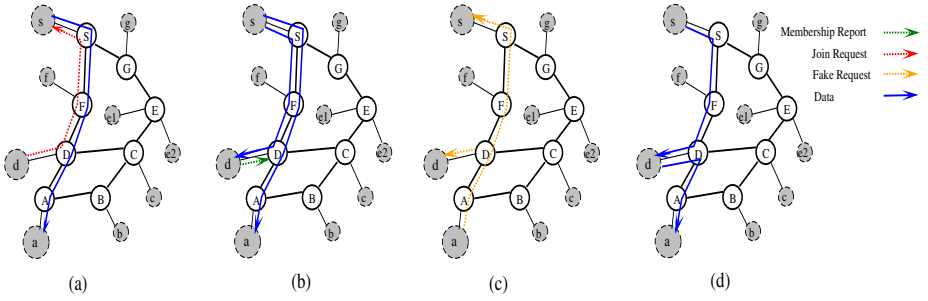


Fig. 3. Topology adjustment

Fig. 3(b). While the tree in Fig. 3(d) is more cost-efficient in terms of bandwidth consumption, the join process of d doesn't provides a mechanism to inform the downstream members of the upstream newcomers.

Our solution is to have downstream in-tree members send a fake join request towards the data source periodically. The fake request has the same destination address/port pair as a true one has. Since the newcomer d is close to the request path of node a, it can capture this fake request as shown in Fig. 3(c). Except the newcomer, other members will not respond to a fake request of a familiar sender. Then the tree could be transformed like Fig. 3(d) for a better shape.

Partition Recovery. If a failure has occurred, the abandoned children follow a join-and-adjust strategy. In other words, the child first tries to make a connection with the best candidate parent in its current list to join the group. Once the overlay link is established, an adjustment process goes on to optimize the overlay topology after the node failure. Due to the candidate parent list, an abandoned child node can find the new parent immediately while still maintaining a good performance of the overlay transmission. Therefore the impact of node failures is reduced to the minimum.

4 Evaluation

In this section, we evaluate the TOMOON protocol using simulations. The main purpose of the simulations here is to investigate how much a TOMOON tree matches the topology of the underlying network. For this purpose, we measure the tree cost in our simulations for a quantified result.

Two benchmark protocols we adopt for comparison with TOMOON are PIM-SM and TAG[4]. As we introduced in Sect.2.2, we consider the former as a lower-bound benchmark. We also choose TAG for comparison because it is similar to TOMOON in its design intention. We adopt the Waxman approach[6] to generate a topology to mimic the Internet for our experiments. The fan-out degree of each node in the graph has an upper bound of 6. The cost and delay of each link is set to the same value, which varies from 1 to 5 for intra-domain links and from 6 to

10 for inter-domain links. To evaluate the performance of our protocol in different network environments, we repeat our simulations in two different scenarios:

1. Physical links have symmetric cost.
2. Physical links have asymmetric cost.

4.1 Simulation Results and Discussions

In the first experiment, we assume that all domains support IP multicast. Whenever a node joins the group in TOMOON, it will broadcast its membership report throughout its domain to capture other join requests passing by. We change the group size from 20 to 100. Fig. 4 shows the tree cost for two scenarios.

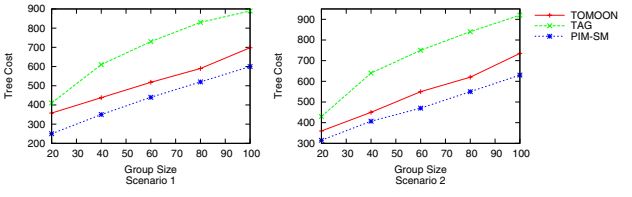


Fig. 4. Group size VS tree cost

From this figure, we can infer that the performance of TOMOON is better than TAG and approaches that of PIM-SM as the membership increases. We give the reasons for this result from two aspects. First, as an increasing number of nodes join the group in TOMOON, it is much easier for a new member to find a parent in its close neighborhood. In other words, the extra cost for accepting a new member into the group decreases when the membership increases. Second, the overlay of TAG is not adapted to the underlying network topology in all situations. Although the parent and its new child share a longest segment on the routes from the source to each of them in TAG, it is still possible that the overlay link between these two nodes is longer than other alternatives. Thus TAG is less aggressive than TOMOON in saving the tree cost.

In the second experiment, we assume that only a part of domains support IP multicast. It is much closer to the situation of the real Internet. We select two

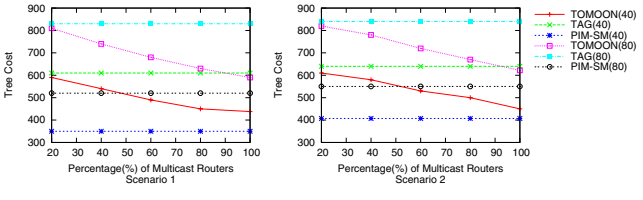


Fig. 5. Percentage of multicast domains VS tree cost

fixed group sizes(40 and 80) and adjust the percentage of multicast domains from 20% to 100% to study TOMOON's performance in different environments. The growing percentage of multicast domains is transparent to other two protocols in their simulations. From Fig. 5, we find that the output of TOMOON is still acceptable even when the multicast domains account for a small percentage in the underlying network.

5 Conclusion

In this paper, we propose TOMOON as a novel overlay multicast approach, which exploits the support from IP multicast routers to construct a topology-aware overlay tree. Compared with IP multicast and other overlay multicast schemes, our approach has the following notable advantages :

1. A TOMOON tree is compatible with the physical network. With the help of multicast routers, inefficient branches in a multicast tree are eliminated.
2. TOMOON makes the membership management much easier in overlay multicast. In other words, TOMOON can deal with member joins and leaves efficiently.
3. TOMOON encourages ISPs to provide better network layer support. Since the multicast routers can capture consumers' requests of various network services, ISPs are willing to build multicast domains for expectable commercial benefits.

References

1. Cain, et. al.: Internet Group Management Protocol, Version 3, RFC 3376, October 2002
2. S. Banerjee, B. Bhattacharjee, and C. Kommareddy: Scalable application layer multicast. *Proc. ACM Sigcomm*, Aug. 2002
3. Y.-H. Chu, S. G. Rao, and H. Zhang: A Case for End System Multicast. *IEEE J. Select. Areas Commun*, VOL. 20, NO. 8, OCTOBER 2002.
4. Kwon, Minseok, and Sonia Fahmy: Path-aware overlay multicast *Computer Networks 2005(47)*, pp. 23–45.
5. Estrin, et. al.: Protocol Independent Multicast-Sparse Mode (PIM-SM): Protocol Specification", RFC 2362, June 1998
6. B.M. Waxman: Routing of multipoint connections.*IEEE J. Select. Areas Commun.*, vol.6, no. 9, Dec. 1988.

Fuzzy-Timing Petri Nets with Choice Probabilities for Response Time Analysis

Jaegeol Yim and Kye-Young Lee

Dept. of Computer and Multimedia Dongguk University at GyeongJu, GyeongBuk, 780-714
Republic of Korea
{yim,lky}@dongguk.ac.kr

Abstract. The system response time is one of the most critical user's requirements. Therefore, a response time analysis must be done at the design stage. This paper proposes a response time analysis method based on 'Fuzzy-Timing Petri net with choice probabilities (FTCP)'. FTCP is a modified version of timed net: A delay time associated with a transition is not a crisp number but a possibility distribution function and a transition of FTCP is also associated with a choice probability. We are proposing a response time analysis procedure for FTCP. It is similar to the minimum cycle time analysis for Timed Nets. The difference is that we take account of choice probabilities of the transitions in the process of generating the T-invariant. The other feature of our analysis method is that it can handle temporal uncertainties. This paper finally demonstrates the usefulness of the method by applying the method on a Web Service System.

1 Introduction

Petri net has been widely used in the field of computer system analysis. Modeling and performance analysis of message-oriented middleware using generalized stochastic Petri net is introduced in [1]. Petri net has also been used in steel making process [2]. Murata introduced Fuzzy-Timing High-Level Petri Nets for time-critical applications [3]. Integrating Fuzzy Timing Petri net and Time Petri net, Zhou et al made up a new temporal Petri net. Using the new temporal Petri net, they built a model of distributed multimedia synchronization and analysis methods [4]. The new temporal Petri net has also been used in modeling and performance analysis of networked virtual environments [5].

A Timed Petri net is a modified Petri net where transitions are associated with delay times. The minimum cycle time analysis method of Timed Petri net has been introduced in [6]. The method has been applied to a Petri net model of a multi-stage production system for performance evaluation of the system. Dasdan and Gupta introduced a more efficient algorithm for minimum cycle time analysis [7]. An algorithm to find the minimum cycle time of a marked graph, a subclass of Petri nets, is introduced in [8]. Given both a timed net and a limit of cycle time, finding the minimum initial marking for which the minimum cycle time is less than or equal to the limit is called the minimum initial marking problem. An answer for this problem is given in [9].

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 652–659, 2007.
© Springer-Verlag Berlin Heidelberg 2007

One of the advantages of Fuzzy Timing Petri net is that it can handle uncertain time whereas one of its disadvantages is the lack of efficient analysis methods. Although reduction methods and simulation methods are introduced in [5] as analysis methods for Fuzzy Timing Petri net, they are essentially time consuming exhaustive methods.

One of the advantages of minimum cycle time method is its efficiency. Once we find S-invariants and T-invariants, we can perform minimum cycle time analysis in a sequence of matrix multiplications and divisions. With the result of minimum cycle time we can find the bottleneck of the system. The weak points of minimum cycle time method include the following two points: 1) It can handle only constant times 2) it does not consider the choice probabilities of events. In the real world, events are associated with not only delay time but also probabilities of choices. For example, considering a read event and a write event in a database, not only delay time for read is different from delay time for write, but also the choice probability of read event is different from the choice probability of write event. However, the existing Timed net only considers delay time.

This paper proposes 'Fuzzy-Timing Petri Nets with Choice Probabilities (FTCP)' where transitions are associated with not only delay time but also probability of choice. The delay time associated with a transition can be a fuzzy time function, or possibility distribution. As an analysis method for FTCP, we propose a modified version of minimum cycle time method. With the method, we can find the response time of the underlying computer system.

2 Related Works

We are assuming that readers are familiar with Petri net terminologies, such as transition, place, arc, weight, initial marking, T-invariant, S-invariant, support, and so on. For introduction to Petri net theory and its applications, please refer to one of the references [10-13]. Among the related works, Fuzzy-Timing Petri nets [3] and minimum cycle time method [10] are the most closely related to this paper.

An example of Fuzzy-Timing High-Level Petri net (FTHN) is shown in Fig. 1 [3]. The FTHN represents a job shop. A job arrival is represented as a token at place P_{in}. Two tokens 'a' and 'b' in place P_{in} implies that two jobs named 'a' and 'b' have been arrived for process. There is a machine which can be in either ready or busy state. A job arrived in P_{in} leaves the job shop via place P_{out} after being processed by the machine. Place P_{busy} is corresponding to the state in which the machine is busy operating on a job. A job in P_{in} can enter P_{busy} if there is a machine in P_{ready}.

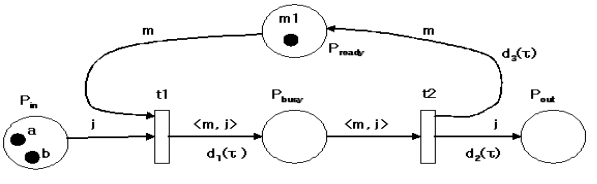


Fig. 1. A FTHN model of a job shop

A token in the FTHN is associated with a timestamp. A timestamp is a possibility distribution function. As examples of timestamps, let

$_{0a}(\tau)$ = (1 3, 3, 5), $_{0b}(\tau)$ = (3, 5, 5, 7) and $_{0}(\tau)$ = (0, 0, 0, 0) be the timestamps associated with tokens 'a' and 'b' at place P_{in}, and the token at place P_{ready}, respectively. An arc from a transition *t* to place *p*, (*t, p*) is associated with a delay time. Let the arcs (t_1, P_{busy}), (t_2, P_{ready}) and (t_2, P_{out}) be associated with $d_1(\tau)$ = (0,0,0,0) and $d_2(\tau) = d_3(\tau)$ = (4,5,7,9), respectively, in Fig. 1. Reference [3] introduces an analysis method with which we can figure out the possibility for the jobs can be finished within certain given times. However, the method is essentially enumerating all the possible firing sequences and too time consuming.

In reference [10], a cycle time is defined as the time to complete a firing sequence from the initial marking back to the same marking after firing each transition at least once on a Timed Net. A mathematical cycle time analysis method is also introduced in [10] as follows, where, y_k is an S-invariant, $A^- = \left[a_{ij}^- \right]_{n \times m}$ and $a_{ij}^- = w(p_j, t_i)$, D is the diagonal matrix of delay d_i, i=1, 2, ..., n, and M_0 is the initial marking.

$$\tau_{\min} = \underset{k}{Max}\{ y_k^T (A^-)^T Dx / y_k^T M_0 \} \tag{1}$$

However, a Timed Net is not equipped with any facility of specifying the choice probabilities for transitions in conflict.

3 Fuzzy-Timing Petri Nets with Choice Probabilities

The main objective of this paper is to propose the Fuzzy-Timing Petri Net with Choice Probabilities (FTCP). We are introducing an example of it, then a formal definition of it, followed by analysis method in this section. An example FTCP is shown in Fig. 2. It is FTCP model of a readers-writers system. In the real world, the frequency of read operation is greater than the frequency of write operation. However, the frequencies are not considered in the timed net. The frequency of read operation, C_1, is labeled on t_1 and the frequency of write operation, C_2, is on t_2 in Fig. 2. The labels for delay times (d_i s) on Fig. 2 are possibility distribution functions.

For the Petri net in Fig. 2, there are two minimal support T-invariants: $x_1 = (1\,0\,1\,0)^T$ and $x_2 = (0\,1\,0\,1)^T$. Let the choice probabilities of C_1 and C_2 be 0.9 and 0.1, respectively. Considering the choice probabilities, we can find a T-invariant x = (9 1 9 1). Furthermore, if we want to find the time required for the k processes to perform either read or write operation once each, we should use x = (0.9k 0.1k 0.9k 0.1k).

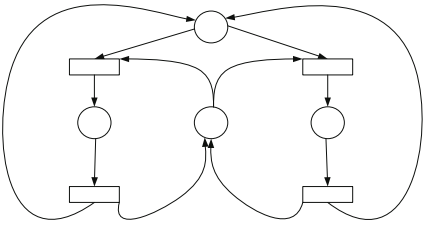


Fig. 2. An example Fuzzy-Timing Petri net with Choice Probabilities

If we use x=(0.9k 0.1k 0.9k 0.1k), then we obtain

$$(A^-)^T Dx = \left[0.9kd_1 + 0.1kd_2 \;\; 0.9kd_3 \;\; 0.9kd_1 + 0.1k^2 d_2 \;\; 0.1kd_4 \right]^T$$

Thus, application of (1) yields the following minimum cycle time:

$$\tau_{\min} = Max\{(0.9kd_1 + 0.1kd_2 + 0.9kd_3 + 0.1kd_4)/k,\ [0.1k^2(d_2 + d_4) + 0.9k(d_1 + d_3)]/k\}$$
$$= 0.1k(d_2 + d_4) + 0.9(d_1 + d_3).$$

An interpretation of the minimum cycle time is that 0.1 k out of k processes will perform 'write operation' spending (d_2+d_4) time each, i.e. 0.1 k processes will spend 0.1k * (d_2 + d_4) time units. On the other hand, 0.9 k processes will perform 'read operation'. Since read operation can be done in parallel, it takes only (d_1+d_3) time units for all k processes to perform read-operation. The minimum cycle time counts only portion of (d_1+d_3), i.e. 0.9(d_1+d_3) time units because only 0.9 k processes perform 'read-operation'. In total, the readers-writers system with k processes needs $0.1k(d_2 + d_4) + 0.9(d_1 + d_3)$ time units in order for k processes to perform either read or write operation once each. It is worth of noting that the time needed for k processes to perform either read or write operation once each is the system response time if k is the average number of processes using the DB concurrently.

As a numerical example, let d_1=d_2=(0, 0.5, 0.5, 1.5), d_3=(1,2,2,4), d_4=(3, 4, 5, 7), and k=10. Then the response time is (0, 0.5, 0.5, 1.5) (3, 4, 5, 7) 0.9((0, 0.5, 0.5, 1.5) (1, 2, 2, 4)) = (3, 4.5, 5.5, 8.5) (0.9, 2.25, 2.25, 4.95) = (3.9, 6.75, 7.75, 13.45). From the result, we can conclude that the system meets the following fuzzy statements: The system mostly responses within 7.8 time units.

For the formal definition of FTCP, we are defining a few terminologies.

[Def. 1] If transitions t_a and t_b are sharing an input place, they are said to be in choice relation, and denoted by 't_a choice t_b' or choice(t_a, t_b).

[Def. 2] If choice(t_a, t_b), then these transitions should be associated with probabilities of choice. The probability of choice for t_a is denoted by Prob(t_a).

For the simplicity we denote the output/input vertices of a vertex v v•/ •v. For example, p_1• = {t_1, t_2} in the case of Fig. 1. Sum of all the choice probabilities of transitions sharing one input place should be 1, i.e. the following holds:

$$\sum_{t_j \in p_i \bullet} \mathrm{Prob}(t_j) = 1$$

A definition of FTCP is given in Table 1.

Since a transition can be associated with a choice probability, a T-invariant can also be associated with a choice probability. In the definition of the choice probability of a T-invariant, we are using the terminology, group, defined in [Def. 3].

Table 1. A definition of FTCP

A FTCP is 7-tuple, FTCP = (P, T, F, W, M_0, DT, Prob,), where: (P, T, F, W, M_0) is a Petri net, DT: T → fuzzy timestamps, is a function determining delay time of a transition, $\mathrm{Prob}: T \to (0,1)$ is a function determining choice probability, where, $\sum_{t_j \in p_i \bullet} \mathrm{Prob}(t_j) = 1.$

[Def. 3] Given a T-invariant x, a group regarding choice(t_1, t_2, ... t_k) is the set of transitions which are in the choice relation and are also members of the support of x.

[Def. 4] Given a group $g_i = \{t_1, t_2, ... t_k\}$, the choice probability of g_i is defined by $Prob(g_i) = Prob(t_1)+Prob(t_2)+...+Prob(t_k)$.

[Def. 5] Given a T-invariant x, the choice probability of x, Prob(x), is defined by the probability obtained by performing the following steps:

Step 1: We find all groups considering all choices.

Step 2: For all the groups found in Step 1, we find choice probabilities of them.

Step 3: We multiply all the choice probabilities of the groups to find the choice probability of x, i.e. the choice probability of x is defined by: $\Pr ob(x) = \prod_{g_i \subseteq \| x \|} \Pr ob(g_i)$.

The response time is the time needed to finish a certain process. A process consists of many steps including the start step. The place corresponding to the start step is called Start Place. If we designate a different place as Start Place then response time is also changed. Therefore, we can conclude that we have to designate the Start Place properly considering the semantics of the FTCP. We define a FTCP which is consistent and has a Start Place a well-formed FTCP.

Given a well-formed FTCP, the algorithm shown in Table 2 finds the response time of the computer system modeled as the FTCP.

Table 2. Algorithm for response time analysis of FTCP

Input:
N: A well-formed FTCP,
n: the number of tokens at Start Place.
Step 1: We find all minimal support T-invariants of N.
Step 2: We perform the following steps to form maximal minimal T-invSPs
2.1 Construct a sub-FTCP for each minimal support T-invariant
2.2. While there exists a sub-FTCP$_i$ whose covering T-invariant is not a T-invSP
2.2.1. Merge sub-FTCP$_i$ to its adjacent sub-FTCP.
2.3. For each sub-FTCP, we obtain its covering T-invariant (* All of them are maximal minimal T-invSP *)
Step 3: For each T-invariant x_i obtained at Step 2.3, we obtain its choice probability $prob_i$.
Step 4: For each T-invariant x_i obtained at Step 2.3, we obtain $prob_i*n*x_i$.
Step 5: We obtain a positive T-invariant x by summing up all $prob_i*n*x_i$.
Step 6: Apply equation (1) $\tau_{\min} = \max_k \{ y_k^T (A^-)^T D x / y_k^T M_0 \}$.

[Theorem] Sum of all the choice probabilities of all the T-invariants obtained at Step 2 is 1.

<Proof> We prove this theorem with mathematical induction.

Base Step: Since the given FTCP is well-formed, it has at least one T-invariant. If it has exactly one T-invariant, then for all p_i P, $|p_i\bullet|=1$. $|p_i\bullet|=1$ implies that for all t_i T, $Prob(t_i)=1$. Therefore, the choice probability of x is 1*1*...*1 = 1.

Induction Hypothesis: If there are k T-invSPs, then the sum of choice probabilities of them is 1.

Induction Step: We have to prove that if one more T-invSP is added to existing k T-invSPs then the sum of choice probabilities is still 1. Let the k existing T-invariants be $x_1, x_2, ..., x_k$ and $Prob(x_1)+Prob(x_2)+...+Prob(x_k) = 1$. According to Theorem 1, addition of $k+1^{th}$ T-invSP requires one addition of choice relation. Let $choice_i(t_i, t_j)$ be

the additional choice relation. Then, either t_i or t_j must be a member of the support of one of the existing k T-invariants and the other one must be a member of the support of k+1th T-invSP. Let t_i $\|x_h\|$ and t_j $\|x_{k+1}\|$. Adding $choice_i(t_i, t_j)$ changes the $prob(t_i)$. Therefore, we name the x_h with new choice probability of t_i $newx_h$.

Let $\{t_{i1}, t_{i2}, ..., t_i, t_{i+1}, ..., t_{iq}\} = \|newx_h\|$, $\{t_{j1}, t_{j2}, ... t_j, t_{j+1},..., t_{jr}\} = \|x_{k+1}\|$, and the transitions are listed in the order of distance from Start Place. Then, the following must be true. Otherwise, the number of T-invSP is increased by more than one.

$t_{i1}= t_{j1}$, $t_{i2}= t_{j2}$, ..., $t_{i-1}= t_{j-1}$ and $prob(t_{j+1})=1$, ..., $prob(t_{iq})=1$, $prob(t_{j+1})=1$,..., $prob(t_{jr})=1$.

Therefore, $Prob(newx_h) = Prob(x_h)*prob(t_i)$, and $Prob(x_{k+1}) = Prob(x_h)*prob(t_j)$. Now, $Prob(newx_h) + Prob(x_{k+1}) = Prob(x_h)*prob(t_i) + Prob(x_h)*prob(t_j) = Prob(x_h)*\{prob(t_i) + prob(t_j)\}$. But, $prob(t_i) + prob(t_j) = 1$. Therfore, $Prob(x_h)*\{prob(t_i) + prob(t_j)\} = Prob(x_h)*1 = Prob(x_h)$.

4 Application of FTCP on Web Service System

An FTCP model of a web service system is shown in Fig. 3.

At Step 1 of the algorithm shown in Table 2, we find the following minimal support T-invariants:

$x_1 = (1\ 1\ 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0)^T$,
$x_2 = (1\ 1\ 0\ 1\ 0\ 1\ 0\ 1\ 1\ 1\ 2)^T$,
$x_3 = (1\ 1\ 0\ 1\ 1\ 0\ 1\ 1\ 1\ 1\ 2)^T$.

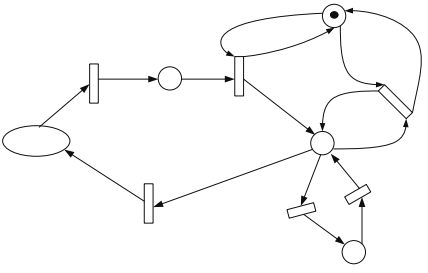


Fig. 3. An FTCP model for Web service system

At Step 2, we find that the minimal support T-invariants are also maximal minimal T-invSPs.

At Step 3, we find the probability of choice for $x_1 = 1*1*0.8 = 0.8$, for $x_2 = 1*1*0.2*0.7*1... = 0.14$, and for $x_3 = 0.2*0.3 = 0.06$.

As the result of Step 4 and 5, considering the probabilities of choice and the number of users, n, we obtain the following (2) positive T-invariant. A sequence of firing t_2 n times, t_1 n times, t_3 0.8n times, and so on is possible and returns back to the initial marking.

$$x = (n\ n\ 0.8n\ 0.2n\ 0.06n\ 0.14n\ 0.06n\ 0.2n\ 0.2n\ 0.2n\ 0.4n)^T \quad (2)$$

Step 6 is the existing minimum cycle time method, and we can obtain the response time as follows:

$$\text{Max } \{nd_1/n,\ 0.2n(d_2+d_5)/r,\ 0.06nd_3/m,\ 0.2nd_4/k,\ (nd_1+0.2nd_2+0.06nd_3+0.2nd_4+0.2nd_5)/n\} \quad (3)$$

For a numerical example, consider the following possibility distribution functions and numbers of resources:

$d_1 = (59{,}900,\ 60{,}000,\ 60{,}000,\ 60{,}100)$, $d_2 = (90,\ 100,\ 100,\ 110)$, $d_3 = (490,\ 500,\ 500,\ 530)$, $d_4 = (990,\ 1{,}000,\ 1{,}000,\ 1200)$, $d_5 = (99,\ 100,\ 100,\ 120)$, $n = 1{,}000$, $k = 1$, $m=1$, $r=1$.

The response time is:

Max{(59,900, 60,000, 60,000, 60,100), (37,800, 40,000, 40,000, 46,000), (29,400, 30,000, 30,000, 31,800), (198,000, 200,000, 200,000, 240,000), (60,165.2, 60,270, 60,270, 60,417.8)} = (198,000, 200,000, 200,000, 240,000).

From the response time, we can conclude that the following statements hold:

1) The system mostly responses within 220,000 time units.
2) The system definitely responses within 240,000 time units.
3) The system unlikely responses within 200,000 time units.
4) The system never responses within 198,000 time units.

The result implies that the system's response time is dominated by the content provider. That is, a user of the system will waste time for waiting response from the LBS system. In order to solve this problem, the designer has to do one of the following: 1) reduce subscribers, 2) speed up the content provider 3) increase the number of processors of the content provider.

As another numerical example, let us change the number of subscribers to 50 and the other factors leave the same as the above example. Then the minimum cycle time will be determined by d1, the time interval needed by a user to choose a menu. In this case, the system's performance is better than necessary and we can do one of the followings: 1) reduce implementation cost by using less expensive resources, or 2) serve more subscribers.

5 Conclusions

In the real world, the duration time from the start of an event to the end of the event is not a crisp number but a fuzzy number. The events are usually in conflict in that the occurrence of event A prevents event B from occurrence. However, the existing minimum cycle time method which deals with timed nets considers neither temporal fuzziness nor choice probabilities of events. On the other hand, Fuzzy-timing Petri net can handle temporal uncertainty but the analysis of the net is time consuming.

We proposed Fuzzy-Timing Petri net with Choice Probability (FTCP) in this paper. FTCP is a new version of timed net in which a transition can be associated with both possibility distribution function and choice probability. Therefore, FTCP has both Fuzzy-Timing Petri net's advantage and Timed net's advantage, namely the response time analysis method of FTCP is as efficient as the minimum cycle time method and the result of response time analysis is not a crisp number but a possibility distribution function.

We have built a FTCP model of a Web Service System and analyzed it with our response time analysis algorithm. The analysis result showed that FTCP indeed possesses both the advantages of Fuzzy-Timing Petri net and of Timed Net.

References

[1] S. Fernandes, W. Silva, M. Silva, N. Rosa, P. Maciel, and D. Sadok, "On the generalised stochastic Petri net modeling of message-oriented middleware systems," 2004 IEEE International Conference on Performance, Computing, and Communications, Cidade Univ., Recife, Brazil, 2004, pp. 783–788

[2] H. Gao, J. Zeng, G. Sun, Y. Wu, "New matrix method for analyzing steel making process," 2002 International Conference on Machine Learning and Cybernetics, Vol. 2, Nov. 4-5, 2002, pp. 1070-1074.

[3] T. Murata, "Temporal Uncertainty and Fuzzy-Timing High-Level Petri Nets," Application and Theory of Petri Nets 1996, pp.11-28.

[4] Y. Zhou and T. Murata, "Modeling and Analysis of Distributed Multimedia Synchronization by Extended Fuzzy-Timing Petri Nets," Transactions of the SDPS (Society for Design and Process Science), Vol. 5, No. 4, Dec. 2001, pp. 23-37.

[5] Y. Zhou, T. Murata, and T. DeFanti: Modeling and Performance Analysis Using Extended Fuzzy-Timing Petri Nets for Networked Virtual Environments. IEEE Transactions on Systems, Man, and Cybernetics – Part B: Cybernetics, Vol. 30, No. 5, October 2000, pp.737-756.

[6] Hervé P. Hillion, "Timed Petri Nets and Application to Multi-Stage Production Systems.", Lecture Notes in Computer Science, Vol. 424; Advances in Petri Nets 1989, Berlin, Germany, Springer-Verlag, 1990, pp. 281-305.

[7] A. Dasdan, R.K. Gupta, "Faster maximum and minimum mean cycle algorithms for system performance analysis," IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems, Vol. 17, No. 10, 1998, pp. 889-899.

[8] M. Nakamura, M. Silva, "Cycle time computation in deterministically timed weighted marked graphs," Proceedings of the 7th IEEE International Conference on Emerging Technologies and Factory Automation, Vol. 2, Oct. 18-21, 1999, pp. 1037-1046.

[9] J. Rodriguez-Beltran, A. Ramfrez-Trevino, "Minimum initial marking in timed marked graphs," 2000 IEEE International Conference on Systems, Man, and Cybernetics, Vol. 4, Oct. 8-11, 2000, pp. 3004-3008.

[10] T. Murata, "Petri nets: Properties, analysis and applications," Proceedings of the IEEE, Vol. 77. no. 4, April 1989, pp. 541-580

[11] J. L. Peterson, Petri Net Theory and the Modeling of Systems, Prentice-Hall, N.J., 1981, ISBN: 0-13-661983-5.

[12] W. Reisig, Petri Nets, An Introduction, EATCS, Monographs on Theoretical Computer Science, W.Brauer, G. Rozenberg, A. Salomaa (Eds.), Springer Verlag, Berlin, 1985.

[13] Jörg Desel, Wolfgang Reisig, Grzegorz Rozenberg (Eds.), Lectures on Concurrency and Petri Nets, Advances in Petri Nets, Lecture Notes in Computer Science, vol. 3098, Springer-Verlag, 2004, ISBN: 3-540-22261-8.

A Telematics Service System Based on the Linux Cluster

Junghoon Lee[1], Gyung-Leen Park[1], Hanil Kim[2], Young-Kyu Yang[3], Pankoo Kim[4], and Sang-Wook Kim[5,⋆]

[1] Dept. of Computer Science and Statistics, Cheju National University,
[2] Dept. of Computer Education, Cheju National University,
[3] Graduate School of Software, Kyungwon University,
[4] Dept. of Computer Engineering, Chosun University,
[5] College of Information and Communications, Hanyang University
{jhlee,glpark,hikim}@cheju.ac.kr, ykyang@kyungwon.ac.kr,
pkkim@chosun.ac.kr, wook@hanyang.ac.kr

Abstract. This paper designs and implements a taxi telematics service system, aiming at providing an efficient framework by means of a Linux cluster to host emerging telematics services that need intensive computing. Combined with global positioning system and radio communication technology, the taxi telematics service system traces the position of taxis, finds a time saving route between start and destination points, dispatches the nearest taxi to the service call point based on the latest traffic information, and finally decides an efficient route for multiple destinations. The performance measurement result demonstrates that the implemented system can process up to 200 map matches for every minute, keeping average response time for other requests below 1.5 seconds.

1 Introduction

[1] Telematics is the blending of computers and wireless telecommunication technologies, with the goal of efficiently conveying information over vast networks to improve a host of business functions or government-related public services[1]. This term later evolved to refer to automation in automobiles, namely, vehicle telematics. Due to the mobility of vehicles, the telematics service is necessarily combined with satellite navigation and geographic information. Additionally, GPS (Global Positioning System) and radio communication technologies have become indispensable in providing various telematics services such as vehicle location tracking, automatic collision notification, and location-driven driver information[2].

One of essential applications of telematics networks is vehicle tracking, which is capable of monitoring the location, movement, status, and behavior of a vehicle. The vehicles may be taxis, rent-a-cars, delivery trucks, and any other

⋆ Corresponding author.

[1] This research was supported by the MIC, Rep. of Korea, under the ITRC support program supervised by the IITA (IITA-2006-C1090-0603-0040).

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 660–667, 2007.
© Springer-Verlag Berlin Heidelberg 2007

automobiles. Particularly, for the call taxi company which runs a lot of taxis, the knowledge on the real-time position of each taxi enables to make a decision on the closest taxi to a service call, the best route, and so on. In addition, with the periodic report on the current speed from each taxi, information on the up-to-date traffic conditions, accidents, and jams can be dynamically monitored and estimated[3]. The collected data can be also exploited to some challenging traffic information technologies such as vehicle relationship management, traffic forecast, and mobility analysis[4].

With such requirements, this paper designs and implements the taxi telematics service system capable of tracing the positions of taxis, locating the nearest taxi to a service call, finding an economic path based on the latest traffic information, and discovering an efficient route for multiple destinations on top of a *Linux cluster*. The Linux cluster is known to be the most cost-effective and scalable way to provide large amounts of hardware and software resources. This parallel processing framework seems prospective since the existing and future telematics service will not only handle large amount of data but also need a significant computing power. The functions mentioned above can be implemented using various fundamental techniques including road network management, map matching heuristics[5], diverse versions of path finding algorithms[6], stable database management, and so on.

Using the taxi telematics system, company managers are able to supervise and guide every driver to travel in the optimal route, which is fuel-efficient. The customers need not wait for a longer period since the system aims to dispatch the nearest available taxi. They can also reach the destination by means of the optimal time saving route. The work of the driver is relieved since all guidance is provided by the automated server. Finally, a newly developed service can be easily integrated into the system.

This paper is organized as follows: After issuing the problem in Section 1, Section 2 provides some background for the target taxi telematics system. Then, Section 3 describes the system architecture and implementation details. After discussing the performance measurement results in Section 4, Section 5 concludes this paper with a brief description on future work.

2 Background

The taxi telematics system has been developed as a milestone project for the *Jeju Telematics City* enterprise in Republic of Korea. As one of the most famous tourist attractions in East Asia, Jeju Island is a popular vacation spot for Koreans and many international visitors. It has a well maintained road network which essentially follows the entire coast (200 km) and crisscrosses between the island's major points. In terms of a road network, there are about 18,000 intersections and 27,000 road segments. This means that the road graph can be built with 18,000 nodes and 27,000 links along with some additional data structures such as POI (Point of Interest), allowing the implementor to store the whole graph in the high-speed main memory, not using low-speed file systems or databases

in disk. Hence, almost every functions can be carried out only within the main memory. With the enterprise mentioned above, most of rent-a-cars have already installed in-vehicle telematics for the purpose of tour guide, navigation, safety service, entertainment, and so on. Moreover, some taxi companies began to employ a call taxi control system which tracks the location of taxis, decides the nearest taxi to reduce the waiting time for a customer, maintains the diverse service records for each taxi[7].

The in-vehicle telematics in taxis contains a GPS receiver as well as air interface, which is CDMA (Code Division Multiple Access) protocol in Korea. Each taxi reports its location and speed every minute with a reasonable communication cost negotiated with a telecommunication company. A remote server is responsible for receiving, managing, and exploiting this information. However, the current system just emphasizes the function of vehicle tracking and taxi selection according to the *Euclidean distance* to the call point. They do not even take advantage of the freshness of traffic data in path finding or taxi selection.

In the mean while, all-day hire of a taxi is one of the most popular tour patterns in Jeju, especially for the first time visitors. The customer can suggest multiple destinations, restaurant preference, and so on. Then, the driver makes a tour plan with this requirement and his own experiences. The tour plan essentially starts from and ends at the customer's hotel as long as the customer does not change his hotel, so it can be considered to be a TSP (Traveling Salesman Problem)[8]. TSP is one of the widely studied NP-hard combinational optimization problems and can be described as follows: To begin with, let $G = (V, E)$ be a graph where V is a set of nodes and E is a set of links, and $C = (c_{ij})$ be the distance or cost matrix associated with E. TSP is to find the cheapest way of visiting all of the nodes and return to the starting point.

Among diverse heuristics to solve TSP, the most famous one is Lin-Kernighan's [9,10]. It is known for its success in efficiently finding near-optimal results, while the core of this scheme consists of link exchange in a tour. A research web site of *Concorde* provides diverse TSP solutions including Lin-Kernighan's in a source level, enabling users to download and integrate it to their own system[9]. Though this software can calculate the route plan efficiently, it needs cost matrix each element of which has the respective cost from node i to j. Each element can be calculated by one-to-many version of the Dijkstra algorithm or A* heuristics[6]. Because the calculating step for each element is independent, each calculation can be performed in parallel using a Linux cluster.

3 System Architecture and Implementation

Our telematics service system consists of the telematics device on each taxi, call control server, telematics server, and Linux cluster as shown in Fig. 1. With this architecture, this service system provides such functions as the dynamic link cost update, taxi assignment, path finding between the two points, and path planning for multiple destinations.

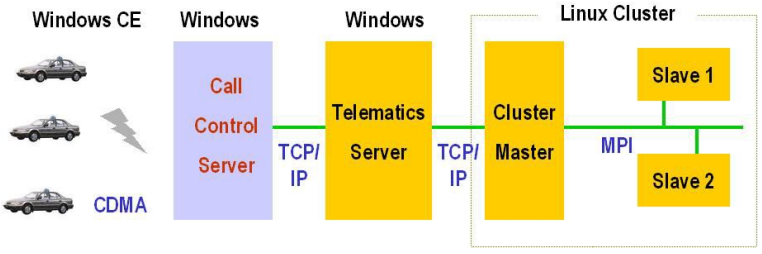


Fig. 1. The architecture of the taxi telematics service system

3.1 Telematics Device and Call Control Server

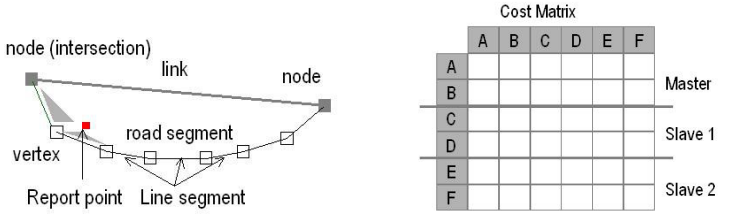
Basically, the telematics device provides taxi drivers with a user interface for the remaining telematics service system. So, the driver can invoke the function he wants via this device. Each telematics device deploys Windows CE 5.0 operating system, and receives information on its location every 1 second from the GPS receiver observing the NMEA specification. The information includes standard time, latitude, longitude, current speed, and moving direction. The telematics device reports these data to the *call control server* residing within the Internet domain. The servers on the wired network can be accessed from each taxi through CDMA air interface, as Windows CE supports TCP/IP communication protocol over the RAS (Remote Access Service) utility.

The call control server tracks and stores the current location of each taxi in the (latitude, longitude) coordinate system as well as to report such data to the *telematics server* for the advanced processing. Currently, up to 200 taxis are tracked simultaneously within our system, and the number of taxis will grow soon. It regenerates a fine-grained request to the telematics server upon receiving a specific function from the human operators or taxi drivers. For example, when a service call arrives from a customer, this server first extracts the list of taxis within 10 km radius from the call point and then sends the list along with the call point itself to the telematics server. After all, this server performs a preliminary functions and plays a role of a gateway to the telematics system.

3.2 Telematics Server

The telematics server provides enhanced functions to the telematics. There are two major data maintained in this server: (1) a digital map as a form of ESRI shape file and (2) a road network as a form of a main memory data structure in a multiple adjacent list graph, respectively. The digital map has full information on each road segment represented as a sequence of lines and vertices. On the other hand, the road network has only nodes and links, which are intersection and two end points of the road segment, respectively. In Fig. 2(a) which illustrates a road segment, a road network in main memory stores just nodes and link, while the digital map stores all nodes, vertices, and line segments. While the digital map is used just for the map matching purpose, the road network is exploited

for sophisticated functions such as taxi dispatching and path finding. Since the road network is the fundamental data to the system, the efficient management of the road network is very crucial to the overall performance of the system. For the town of small or medium size like Jeju, the data is not too large, so it can be stored, processed, searched efficiently only within main memory, avoiding the time-consuming file system calls in a disk.



(a) Digital map and road network. (b) Cost matrix.

Fig. 2. Basic assumptions

The telematics server provides the following functions. First, for the speed report message from the telematics through the call control server, it finds the link that corresponds to the location specified in the message as the (latitude, longitude) pair. As the report can be generated at any place of each road segment along the actual road, the whole digital map file should be searched. For a road segment, the area of each triangle which consists of two end points of each line segment and a report point is calculated, as shown in Fig. 2(a). Then, we can decide the distance from the line segment and the report point and if the distance is less than a given limit, the ID of that link is returned. After finding the corresponding link, the server updates the cost of that link with the reported speed and then stores the record in the system database table in the master node of the Linux cluster. This map matching is endowed with the highest priority as it is periodic and time-sensitive.

Second, for the taxi dispatch request, the server also receives the location of service call and the list of candidate taxis within a radius of 10 km. It decides the nearest taxi to the location of the service call in terms of the network distance by performing the classical Dijkstra's algorithm for multiple destinations. To this end, the telematics server maps all relevant locations represented in the (latitude, longitude) coordinate system to the nearest nodes. This version begins from the location of a service call and proceeds to the current location of candidate taxis by the breadth first search until it reaches any one of them.

Third, for the path finding request with the specification of a source and a destination, it employs a well-known A* algorithm in which the Euclidean distance is exploited for the remaining estimation and network distance for the accumulated cost[6]. In addition, the path finding scheme provides another option which takes into account the current moving direction of a vehicle. After matching the angle between the road segment and the taxi's direction, we can decide the node the taxi will arrive. In the angle matching procedure, the digital map file is used. We also calculate the distance and estimate the time duration from start point to end point.

Finally, receiving a multi-destination planning request, it transfers the request to the Linux cluster after converting the (latitude, longitude) coordinates to the node ID's that are shared with the Linux cluster.

3.3 Linux Cluster

To run a Lin-Kernighan algorithm for the given set of nodes, we need the cost matrix that contains the cost of a path for every node pair in the given set as shown in Fig. 2(b). It takes a significant time to compute the matrix but Linux cluster provides a cost-effective way to enhance the performance of such computation. When the master receives the list of nodes from the telematics server, it partitions the work and distributes the subtasks to the slaves via the MPI (Message Passing Interface) communication primitive[11]. Each computer runs the A* or multi-destination Dijkstra algorithm which calculates the cost from one given node to multiple destinations. This version begins from the start point and proceeds until it reaches all of the destinations. After building the cost matrix, the master runs the Lin-Kernighan algorithm. As it takes less than a second for less than 20 nodes, we don't have to consider the parallel version of a traditional TSP solver.

The cluster master has installed the MySQL DBMS to store the speed report records[12]. As a relational DBMS, MySQL enables to access databases anywhere on the internet. Hence, after installing a MySQL client version on the telematics server, it can not only store the record into the database table, but also make it possible for other analysis server to access the data at any time.

4 Performance Measurement Result

The telematics server runs on the IBM ThinkPad X31 personal computer which has a 1.4 GHz Pentium processor with 512 MB main memory and 100 Mbps Ethernet interface. In the Linux cluster, each of 3 nodes is equipped with a 700 MHz Pentium 3 CPU and 384 MB memory with 100 MBps Ethernet interface. In addition, the NETGEAR 8-port Fast Ethernet switch connects all of cluster nodes to build a private network. Finally, all nodes installed Redhat Linux version 9.0 and LAM-MPI version 7.1.2.

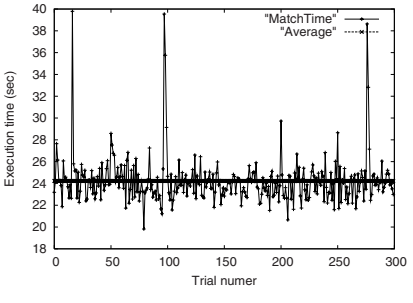


Fig. 3. Performance of map matching

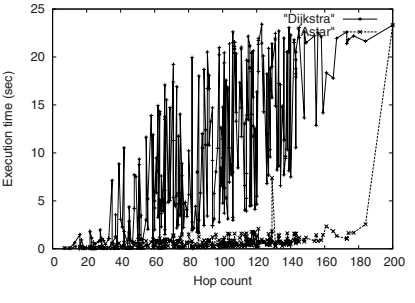


Fig. 4. Performance of path finding

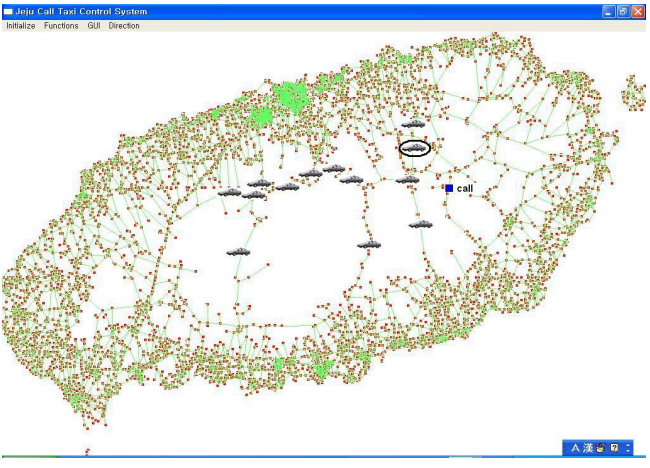


Fig. 5. Finding the nearest taxi to the service call

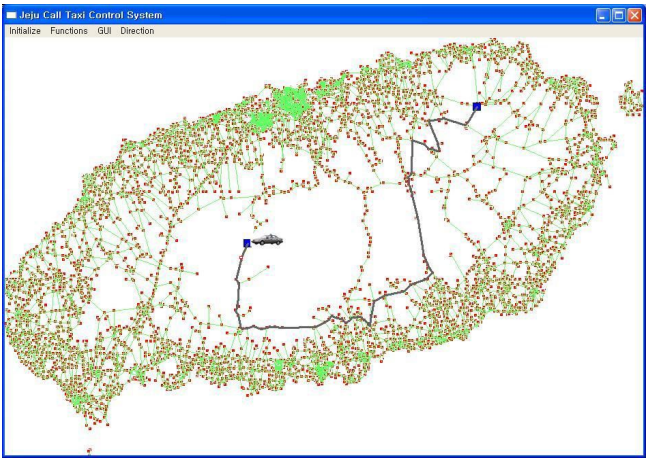


Fig. 6. Multi-destination planning

Fig. 3 plots the execution time of 200 map matches for every 1 minute. The execution time does not exceed 40 seconds (in about 24 seconds on average), giving a sufficient margin to other functions. Fig. 4 plots the execution time of path finding compared with Dijkstra's in the map of the Jeju area. In case there is no feasible route between two points, both schemes show the same execution time, but A* can generally performs 10 $\sim$ 100 times faster than Dijkstra's, providing a reasonable route.

Fig. 5 shows the example of a taxi dispatch result. The closest taxi marked with a circle is found for the candidate to serve the call, as it is the closest in the network distance. The execution time depends on how close the call point is to

such a taxi. Fig. 6 shows the example of a TSP execution result. The execution time also depends on the network distance between the destinations. These GUI displays are implemented in the call control server.

5 Concluding Remarks

In this paper, we designed and implemented a taxi telematics service system for a mid-size city, aiming at providing an efficient framework with a Linux cluster to host a computing-centric telematics service. Taking advantage of the moderate number of nodes and links, we implemented the taxi telematics service system capable of updating the current cost of a link, tracing the position of taxis, finding an efficient path, dispatching the nearest taxi to the service call based on the latest traffic information, and deciding an economic route for multiple destinations. Even though the system is made up with relatively low-performance components, this system can not only meet the system requirements and functional specifications but also be upgraded to achieve a better performance.

As a future work, we are planning to develop a more sophisticated telematics service such as path recommendation, combination of call taxi and tourist information ontology, and a multimedia delivery over the telematics network for advertisement and entertainment.

References

1. http://en.wikipedia.org/wiki/Telematics
2. Kiruthivasan, S., Madan Deepakumar, C., Althaf, S.: Decision Support System For Call Taxi Navigation Using GIS-GPS Integration, MAP India (2006)
3. Lee, S., Lee, B., Yang, Y.: Estimation of Link Speed Using Pattern Classification of GPS Probe Car Data. Proc. International Conference on Computational Science and its Applications (2006) 495-504
4. Jeong, S., Kim, S., Kim, K., Choi, B.: An Effective Method for Approximating the Euclidean Distance in High-Dimensional Space. Proc. International Conference on Database and Expert Systems Applications (2006) 863-872
5. Marchal, F., Hackney, J., Axhausen, K.: Efficient map matching of large global positioning system data sets: Tests on speed-monitoring experiment in Zürich. Journal of the Transportation Research Board (2005) 93-100
6. Goldberg, A., Kaplan, H., Werneck, R.: Reach for A*: Efficient point-to-point shortest path algorithms. MSR-TR-2005-132. Microsoft (2005)
7. Liao, Z.: Real-time taxi dispatching using global positioning systems. Communication of the ACM, **46** (2003) 81-83.
8. Winter, S.: Modeling Costs of Turns in Route Planning. GeoInformatica **6** (2002) 363-380
9. http://www.tsp.gatech.edu/concorde.html
10. Haghani, A., Jung, S.: A dynamic vehicle routing problem with time-dependent travel times. Computer & Operation Research **32** (2005) 2959-2986
11. Pacheco, P.: Parallel Programming with MPI. Morgan Kaufmann Publishers, Inc (1996)
12. Zawodny, J., Balling, D.: High Performance MySQL. O'Reilly Media (2004)

Unequal Error Recovery Scheme for Multimedia Streaming in Application-Level Multicast

Joonhyoung Lee, Youngha Jung, and Yoonsik Choe

Department of Electrical and Electronic Engineering, Yonsei University,
134 Shinchon-dong Seodaemun-gu, Seoul, 120-749, South Korea
{jhlee019,crosscom,yschoe}@yonsei.ac.kr

Abstract. As a way of implementing multicast delivery, application-level multicast (ALM) is known for its various advantages including high flexibility and easy deployment. However, one of its key limitations in achieving reliability and quality-of-service (QoS) for streaming services is the hierarchical error propagation caused by its tree structure. As a solution to this error propagation problem, we propose an unequal error recovery (UER) scheme where we apply different number of retransmissions to the different levels of the ALM tree. The error recovery scheme is implemented using reliable-UDP (RUDP). For each node of the ALM tree, the optimal number of the maximum retransmissions that minimizes the residual loss rate is obtained using Lagrange theory for a given delay constraint. Compared to the equal error recovery scheme where the equal number of maximum retransmissions is applied throughout the ALM tree, the proposed UER scheme shows up to 10% improvement in residual loss rate while maintaining acceptable delay overhead for real-time streaming applications.

Keywords: ALM, multicast, multimedia streaming, error recovery.

1 Introduction

Multicast delivery has been the preferred choice for the middle ground application between point-to-point unicast and large-audience broadcast [1]. In implementing the multicast functionality, two approaches have been widely used: *IP multicast* and *application-level multicast* (ALM). While IP multicast can provide the best performance in terms of the bandwidth on physical links, it has several drawbacks hindering its wider deployment. These include slow pace of deployment and the difficulties in reliability, congestion control and flow control. On the contrary, ALM can provide higher flexibility in streaming applications by implementing the multicast functionalities at the end hosts connected to the network. That is why it started receiving much attraction recently [1].

In general, ALM protocol is comprised of a logical tree rooted at source, where all the nodes are hosts and each link in the tree represents a network path. In this hierarchical structure, packet loss occurred in any links can be propagated to all the descendant links. The packet losses accumulated at the lower part of the tree can then cause severe performance degrading to its leaf nodes. Also,

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 668–675, 2007.
© Springer-Verlag Berlin Heidelberg 2007

leaf nodes can suffer from a large amount of packet loss even though the link connected to the leaf node itself has no packet loss at all. Therefore, packet loss occurred in each link in the ALM tree has different impacts depending on its level. For this reason, we can construct more efficient and reliable ALM tree by making the higher levels of the tree more robust than its lower levels.

In order to recover packet losses in the ALM tree, many approaches have been proposed including forward error correction (FEC) based on redundant parity packets [3,4] and the retransmission scheme based on feedback mechanism [5]. While all these approaches show good error recovery performances in each link of the ALM tree, they still show the inherent limitation caused by the hierarchical error propagation. Recently, a lateral error recovery (LER) scheme [2] was proposed to improve the error recovery performance of the ALM tree. In LER, a node recovers its losses by retransmission not from its ancestor node but from some nearby recovery neighbors of other planes, which contain independent ALM trees. However, the performance of LER scheme can still be improved by taking the hierarchical error propagation into account. While the retransmission scheme can improve the performance, it has the side effects of increasing end-to-end delays. For this reason, simply applying the same error recovery scheme to all the levels in the ALM tree can result in the large delay overhead especially at the leaf nodes.

In this paper, we propose a unequal error recovery (UER) scheme based on reliable-UDP (RUDP) [7]. UER has been designed and developed with the trade-off between performance and delay overhead in mind. In other words, high performance scheme is applied to the higher levels of the ALM tree and low performance scheme is applied to the lower levels. To do this, the effect of the configurable parameters of RUDP on the performance is first analyzed and one of the dominant parameters - maximum number of retransmission - is selected to model the performance of the error recovery scheme. Then, optimal solution for the number of retransmission that minimizes overall packet loss rate is obtained. The optimization is performed using Lagrange theory with delay penalty as a constraint. Compared to the uniform error recovery scheme (where same type of recovery scheme is applied to all the levels in the ALM tree), UER shows up to 10% improvement in residual loss rate while maintaining a similar delay penalty.

Although we use RUDP in our UER approach, the retransmission scheme can be replaced by any kind of error recovery scheme as long as it can provide a means to apply the unequal performances depending on the level of the ALM tree. Also, different kinds of error recovery schemes can be applied to the different levels of the ALM tree in order to further improve the performance.

The remainder of this paper is organized as follows. In Sect. 2, basic error characteristics of ALM are presented. In Sect. 3, the performance of retransmission scheme with RUDP is modeled using its configurable parameters and an optimal solution for the parameter is obtained for UER using Lagrange theory. Experimental results are shown in Sect. 4 followed by the conclusion in Sect. 5.

2 Error Characteristics in ALM Tree

Among various ALM tree construction algorithms, the binary tree is selected to construct an ALM tree in order to analyze the error characteristics. In Fig. 1, the packet 3, which is lost between source S and node A becomes lost packets in all descendants. While the loss of packet 8 from node B to node C affects node C only, the loss of packet 3 between source S and node A consequently means packet loss to the other six nodes. That is, the number of nodes affected by packet loss strongly depends on the level of the hierarchical ALM tree.

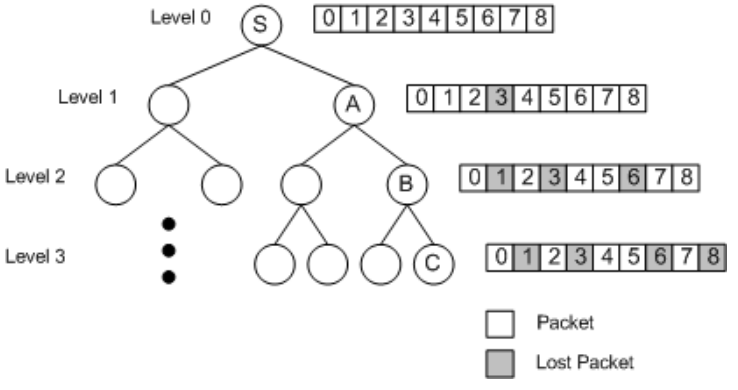
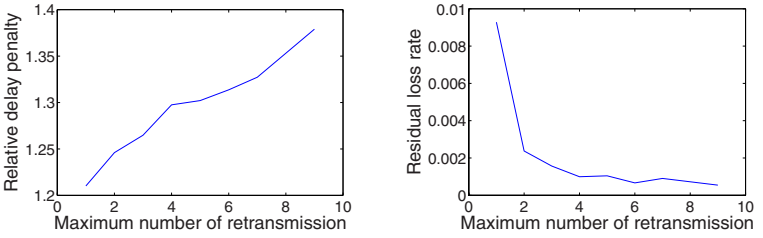


Fig. 1. Error propagation property through the link of the ALM tree

By considering the error propagation described above, decreasing packet losses in the higher level of the ALM tree can improve the performance of the ALM. While minimizing the packet loss rate in all levels of an ALM tree is the best solution, improvement of error recovery at all levels might cause the following side effects: 1) For FEC, increasing redundant packet reduces packet loss rate, while bandwidth efficiency is reduced. 2) In case of retransmission scheme, incrementing the number of retransmission accompanies additional delay in network path although it improves the error recovery performance. There is a strong possibility that these accumulated side effects in the ALM tree deteriorate the service quality of streaming service severely. For this reason, these side effects should be considered all together to optimize the performance of error recovery scheme.

3 Unequal Error Recovery (UER) Scheme

In this section we analyze the retransmission scheme and model its delay and packet loss rate in terms of the RUDP parameters. Based on the analysis, we find the optimal values of the parameter, the maximum number of retransmission, which minimizes the packet loss rate subject to the delay constraint.



(a) RDP (b) RLR

Fig. 2. The performance of error recovery scheme in terms of the maximum number of retransmission

3.1 Modeling the Performance of Retransmission Scheme

Although TCP provides 100% reliability by means of retransmission, it is not suitable for real-time streaming application [2]. UDP is widely used for real-time streaming application. However it does not provide any error recovery scheme at all. We use RUDP [7] as an error recovery scheme because it provides retransmission mechanism and high flexibility with various configurable parameters. It can also be easily and effectively applied to any real-time streaming applications and show acceptable performance.

Error recovery performance and delay of error recovery scheme are easily controlled by several parameters in RUDP. Because these parameters have trade-off between error recovery performance and delay, we need to analyze the effects of these parameters on the performance of error recovery scheme. To analyze the performance, we use the following two performance measures.

- *Relative Delay Penalty* (RDP); it is defined as the ratio of the delay in the overlay to the delay in error-free delivery.
- *Residual Loss Rate* (RLR); it is defined as the overall packet loss rate for all packets

In order to measure the performance of RUDP in terms of each parameter, large amount of experiments are done with various combination of the parameters for various packet loss rates. These experiments show that the maximum number of retransmission, N, is the dominant parameter in terms of its impacts on the error recovery performance and delay characteristics. This is also backed up by the fact that setting all the other RUDP parameters (except N) as recommended by [7] does not show performance variations compared to setting them to the other values (this results are omitted due to the limit of this paper)

As an example, Fig. 2 shows the performance of error recovery scheme in terms of the maximum number of retransmission, N. We set the packet loss rate of links to 10% in this experiment. As it can be inferred from Fig. 2, RDP

has a linear relationship to N and RLR follows the inverse square curve of N. Accordingly, RDP and RLR can be modeled as follows.

$$D_i = \alpha N_i + \beta \ . \tag{1}$$

$$R_i = \frac{\gamma}{N_i^2} \ . \tag{2}$$

where D_i and R_i are RDP and RLR measured at link i, respectively, N_i is the maximum number of retransmission, and α, β, γ are modeling coefficients.

3.2 Optimal Unequal Error Recovery Scheme

Using the basic models in (1) and (2), the performance of error recovery schemes along a single network path from a source to a leaf node can be expressed as follows.

$$D = \sum_{i=1}^{T} D_i = \sum_{i=1}^{T} (\alpha N_i + \beta) \ . \tag{3}$$

$$R = \sum_{i=1}^{T} w_i R_i = \sum_{i=1}^{T} \frac{w_i \gamma}{N_i^2} \ . \tag{4}$$

where D and R are RDP and RLR measured along a single network path from the source to the leaf node, respectively, T is the number of links and w_i is the weighting factor that represents the number of descendants that are affected by the packet lost in the ith link. Now we want to find the expression for the maximum number of retransmission for each level, $N_1^*, \ldots, N_T^*$, that minimize RLR in (4) while satisfying the condition such that the delay from the source to the leaf node must be equal to or lower than the delay constraint, D_{total}:

$$N_1^*, \ldots, N_T^* = \underset{\substack{N_1,\ldots,N_T \\ \sum_{i=1}^{T} D_i \le D_{\text{total}}}}{\operatorname{argmin}} \sum_{i=1}^{T} \frac{w_i \gamma}{N_i^2} \ . \tag{5}$$

Observe that, since we are minimizing a convex, differentiable function on a convex set, there is a unique solution that can be obtained using Lagrange theory [6]. To do this, we define λ and λ^* as the Lagrange multiplier and its optimal value, respectively. Then the optimization problem in (5) can be re-written in its equivalent (unconstrained) form:

$$N_1^*, \ldots, N_T^*, \lambda^* = \underset{N_1,\ldots,N_T,\lambda}{\operatorname{argmin}} \sum_{i=1}^{T} \frac{w_i \gamma}{N_i^2} + \lambda \left[\sum_{i=1}^{T} (\alpha N_i + \beta) - D_{\text{total}} \right] \ . \tag{6}$$

By setting partial derivatives to zero in (6), we can obtain the following expression for the optimized value for the maximum number of retransmission at each level.

$$N_i^* = \sqrt[3]{\frac{2 w_i \gamma}{\alpha}} \cdot \frac{D_{\text{total}} - T\beta}{\sum_{i=1}^{T} \sqrt[3]{2\alpha^2 w_i \beta}} \ . \tag{7}$$

The experiments using this optimal value in the ALM tree is explained in the next section.

4 Experimental Results

For the performance experiments, we generated a number of ALM trees using the NS-2 simulator [8]. Each node in the tree is connected to a router with 1 ms delay, and each router is connected to the other routers with 30 ms delay. For an ALM tree, the basic form of binary tree was chosen due to its simplicity for analysis. We used MPEG-4 compressed bitstream for streaming application and RTP as a streaming protocol.

Packets transmitted by multicast were randomly dropped in physical links. Packet loss rates were selected in accordance with LER [2]: with the probability of 0.95 the packet loss rate of a link is uniformly distributed between 0 and 1%, and with the probability of 0.05 the packet loss rate is uniformly distributed between 5 and 10%. Packets were divided into two types, such as data packets for video bitstream and control packets for control message. We assumed that control packets cannot be lost by applying an error model to downlink only.

As a measure of performance, RDP and RLR explained in Sect. 3.1 were used and we set the coefficients for (7) such that $\alpha = 0.0029$, $\beta = 1.0304$, $\gamma = 0.000583$. These numbers were obtained through least-square-error model fitting of the various experiments. All the other parameters of RUDP, except the maximum number of retransmission, were set as recommended in [7].

The performance of the proposed UER scheme was compared with the equal error recovery (EER) scheme where the maximum number of retransmission of 3 ($N = 3$) was applied to all levels of the ALM tree. Delay constraint, D_{total}, was set as the RDP value from the EER with $N = 3$, meaning that we want to improve the error recovery performance of the UER under the same delay penalty condition as in the EER scheme. The optimal values of the maximum number of retransmission at each level were obtained from (7) and represented in Table 1.

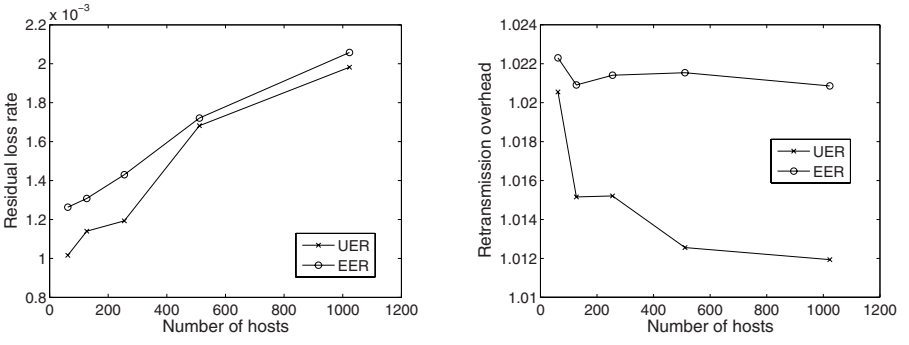
Relative delay penalties of the proposed UER and the EER scheme are compared in Table 2. It can be seen from the table that the RDP values of the UER are almost same as the ones of the EER. This is simply because the UER is designed assuming the delay constraint of the EER. We can also see from the

Table 1. Optimal values of the maximum number of retransmission at each level

The number of links	Level								
from source to leaf	0	1	2	3	4	5	6	7	8
5	5	4	3	2	2	-	-	-	-
6	5	4	3	2	2	1	-	-	-
7	6	4	3	3	2	2	1	-	-
8	6	5	4	3	2	2	1	1	-
9	6	5	4	3	3	2	2	1	1

Table 2. Comparison of relative delay penalty

The number of hosts (levels)	UER	EER
63 (6)	5.39259	5.393329
127 (7)	6.424505	6.431873
255 (8)	7.488069	7.47746
511 (9)	8.592417	8.583579
1023 (10)	9.694796	9.686914



(a) Residual loss rate (b) Retransmission overhead

Fig. 3. UER performance versus number of hosts

table that the increment of the delay using RUDP is acceptable for streaming application. The increment of delay penalty is less than 10% compared to the error-free environments.

The residual loss rate versus the number of hosts was also compared between the UER and the EER and it is shown in Fig. 3(a). As the plot shows, the performance of the UER provides better residual loss rate compared to that of the EER. However, the performance gain of the UER is not perfectly linearly dependent on the number of hosts (there is a dip near 255 hosts). This is because the residual loss rate of UER strongly relies on the distribution of high packet loss rate link. The performance gain of the UER will be higher if the network paths near the source node have high packet loss rates.

In addition to the two performance measures (RDP, RLR) that we discussed in Sect. 3.1, we also analyzed the retransmission overhead of each error recovery scheme as shown in Fig. 3(b). Retransmission overhead represents the ratio of the number of all packets traversing a physical link to the number of source packets in the ALM tree. As seen from the Fig. 3(b), the retransmission overhead of the UER is much smaller than that of the EER. This is because the UER uses smaller value for the number of maximum retransmission in the lower part of the ALM tree and the lower part has much more number of links connected than its higher part. This also explains why the retransmission overhead becomes smaller as the number of host increases.

It can also be inferred from Fig. 3(a) and Table 2 that error recovery performance and delay characteristic of RUDP is good enough to be applicable to any real-time streaming application based on the UDP. Optimization for dedicated streaming application is also very easy due to its simplicity and flexibility.

5 Conclusion

In this paper, we proposed and investigated an unequal error recovery (UER) scheme to recover packet loss in application-level multicast (ALM). The UER was designed to make use of the hierarchical structure of the ALM tree and showed good performance results. As an error recovery methodology, the retransmission-based RUDP was used on top of the UDP. The performance of RUDP is heavily dependent on the maximum number of retransmission so that we modeled the relative delay penalty and the residual loss rate using this parameter and optimal value was obtained using Lagrange theory. The simulation results demonstrated the benefit of the proposed UER approach. Under acceptable delay penalty, residual loss rate can be improved up to 10%, compared to the baseline error recovery scheme which has equal performance in all levels of the ALM tree.

References

1. Ganjam, A., Zhang, H.: Internet Multicast Video Delivery. Proc. IEEE **93** (2005) 159–170
2. Yiu, W.-P.K., Wong, K.F.S., Chan, S.-H.G., Wong, W.C., Zhang, Q., Zhu W.W., Zhang, Y.Q.: Lateral Error Recovery for Media Streaming in Application-Level Multicast. IEEE Trans. Multimedia **8** (2006) 219–232
3. Tan, W.-T., Zakhor, A.: Video Multicast Using Layered FEC and Scalable Compression. IEEE Trans. Circuits Syst. Video Technol. **11** (2001) 373–386
4. Noguchi, T., Yamamoto, M., Ikeda, H.: Reliable Multicast Protocol Applied Local FEC. Proc. IEEE ICC **8** (2001) 2348–2353
5. Towsley, D., Kurose, J., Pingali, S.: A Comparison of Sender-Initiated and Receiver-Initiated Reliable Multicast Protocols. IEEE J. Sel. Areas Commun. **15** (1997) 398–406
6. Pierre, D.A.: Optimization Theory with Applications. New York:Dover (1986)
7. Bova, T., Krivoruchka, T.: Reliable UDP Protocol. Internet Draft, Network Working Group [OnLine]. Available: draft-ietf-sigtran-reliable-udp-00.txt
8. NS-2: network simulator [OnLine]. `http://www.isi.edu/nsnam/ns`

A Fast Handoff Scheme Between PDSNs in 3G Network

Jae-hong Ryu[1] and Dong-Won Kim[2,*]

[1] Electronics and Telecommunications Research Institute (ETRI)
161 Gajeong-dong, Yuseong-gu, Daejeon, 305-350, Korea
[2] Department of Information and Communication, Chungbuk
Provincial University, 40 Gumgu-ri, Okchon-gun, Chungbuk, 373-807, Korea
won@ctech.ac.kr

Abstract. This paper proposes a fast handoff scheme between PDSNs (Packet Data Serving Nodes) which provide packet services to a mobile node. The proposed handoff scheme does not require reestablishment of a PPP connection that may occur in the process of performing handoff between PDSNs. The handoff method between PDSNs requires that the PDSNs should receive subscribers' information about mobile nodes from their neighbor PDSNs forming a communication network. When the PDSN recognizes the mobile node moving into its coverage area, it can quickly establish a communication channel with the mobile node based on the subscriber information received already. As a result, the handoff is performed without reestablishing PPP. Therefore, the handoff between PDSNs can be performed faster removing time needed for establishing a PPP session with a terminal and for terminating a previously set up PPP session.

Keywords: Handoff, Packet Data Serving Node, PPP connection.

1 Introduction

The number of users accessing the Internet through a mobile phone or a Personal Data Assistant (PDA) is increasing rapidly. And CDMA2000 is one of the leading technologies for the third-generation (3G) wireless communications and is being standardized by the 3G Partnership Project 2 (3GPP2)[1]. The CDMA 2000 network can provide both circuit-switched and packet-switched data services [2]. While the circuit-oriented data service is suitable for applications that need bulk data transfer, most data applications will be more efficiently provided by the packet-switched data service.

Since mobile subscribers of a wireless network can roam and move from one cell to another, there is a need for handoff support. Depending on channel usage, there are three types of handoff in the CDMA 2000 system, namely, hard handoff, soft handoff, and softer handoff. A hard handoff requires the MS (Mobile Station) to reestablish synchronization when it enters a new cell. It requires a complex signaling operation between an MS and a network node in order to reestablish a radio channel in the new

* Corresponding author.

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 676–684, 2007.
© Springer-Verlag Berlin Heidelberg 2007

cell. Moreover, after the cellular handoff stage is completed, the MS needs to perform Mobile IP signaling if the care-of-address is changed due to the cellular handoff.

The CDMA2000 works with Mobile IP protocol [3] to provide mobility in the IP layer. For packet data service in the CDMA network, the Mobile IP supports the handoff of the mobile station moving from one cell to another. The Mobile IP is standardized by the IETF, allows a mobile station to transparently access the Internet without changing its IP address when the mobile station moves to another IP domain [3]. The data packets that are destined to the MS cannot reach the MS if the MS has changed its attachment from one PDSN in the CDMA 2000 system to another. A fast moving MS will require a timely completion of the handoff operation so that little data is lost during handoff. Therefore, handoff performance in terms of completion time and data loss prevention has been the key to improving the quality of service [11]. While the CDMA 2000 handoff scheme [2, 4] may be suitable for a packet-switched data service for moderate speed mobile users, there is considerable room for improvement such that the system can provide better performance to fast moving users that can benefit from low handoff latency.

This paper proposes a fast handoff scheme between PDSNs (Packet Data Serving Nodes) for a mobile node that can reduce signaling latency and user data loss during handoff. This improvement is achieved by exchanging subscribers' information about mobile nodes among their neighbor PDSNs forming a communication network.

A number of suggestions have been made so far to improve handoff performance. [5] has suggested the use of a previous foreign agent notification extension allowing data packets in flight to the MS's previous foreign agent to be forwarded to its new location. This reduces the packet loss during the Mobile IP handoff stage. In [6], the authors suggest hierarchical domains of Internet routers in which the Mobile IP registration to the home network can be avoided when the MS hands off regionally within the domain thereby reducing the frequency of access to the home network. This approach reduces the time delay of Mobile IP handoff made within the same domain, but it is not effective when the handoff takes place between different domains. Similar concepts of hierarchy of Internet routers and regional registration have been used in [7, 8, 9].

Our scheme, on the other hand, does not rely on the domain concept and yet it yields less delay regardless of which router the MS hands off to. In addition to the above efforts to improve the performance of Mobile IP handoff, [10] suggests a cellular handoff scheme for CDMA 2000 that replaces a part of the signaling steps carried out by Mobile IP packets with link level signaling messages, thereby reducing message processing delay in the Mobile IP layer.

[11] proposes an improved hard handoff scheme by making a part of the signaling procedure carried out concurrently with other parts in an effort to reduce the total time that it takes to complete the signaling. Our scheme gives better performance than other schemes as shown in Section 4.

The remainder of the paper is organized as follows. A brief introduction to the current CDMA 2000 handoff scheme is given in Section 2. The proposed fast handoff scheme is described in Section 3. Section 4 presents performance evaluation of the proposed signaling scheme with the standard one, and some concluding remarks are given in Section 5.

2 Handoff Scheme of CDMA2000 System

The CDMA2000 consists of radio networks and IP core networks. Figure 1 shows the architecture of the CDMA2000 network. In the radio network, a Base Station Controller (BSC) communicates with a packet data serving node (PDSN) through the packet control function (PCF), which is responsible for handling a logical link between the BSC and the PDSN. Mobile Switching Center (MSC) controls user traffic channels in the radio network.

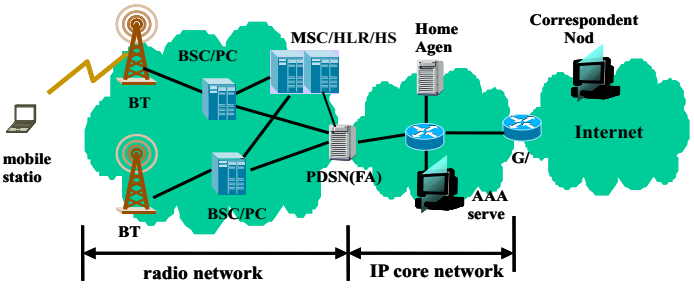


Fig. 1. Architecture of the CDMA2000 Network

The functional elements for the packet data service in the IP core network consist of PDSN, home agent (HA) for Mobile IP service, HLR/HSS and Authentication, Authorization and Accounting (AAA) server. The PDSN provides the interface between the radio network and the IP core network, and performs foreign agent (FA) functions of the Mobile IP protocol. It is responsible for point-to-point protocol (PPP) connection to the mobile station and for logical link (R-P session) to the PCF. It is also responsible for the mapping between the PPP connection and the R-P session. For Mobile IP service, the PDSN is a default router for the mobile station. The mobile station registers with its home agent via the PDSN. After successful registering, the PDSN de-encapsulates data packets which are sent by a correspondent node to the mobile station and are tunneled from the HA. And the PDSN forwards them to the mobile station.

The AAA server supports both the Challenge Handshake Authentication Protocol (CHAP) and the Password Authentication Protocol (PAP) based authentication while a mobile station establishes a PPP connection to the PDSN and requests the Mobile IP registering to the PDSN. The HA maintains the current location of a registered mobile station in its binding table. It is also responsible for the delivery of data to a registered mobile station based on its binding table. The binding table includes home address assigned to a mobile station and COA, which is the address of PDSN that a mobile station has connected on the PPP link.

Handoff methods in CDMA 2000 based wireless data communications can be broken down into handoff between PCFs (Packet Control Functions) and handoff between PDSNs. The handoff between PCFs occurs in a case where the movement of a mobile node between PCFs happens, and PCFs gets changed before and after the moving but are in an identical PDSN area. In this case, care-of-address used by the mobile node does not change so that the mobility is guaranteed.

The handoff between PDSNs occurs in a case where a mobile node moves to another PDSN area. In this case, since care-of-address changes, the IP protocol of the mobile node should be reset. For this, by using a FA (Foreign Agent) function of the PDSN, an IP should be reassigned to the mobile node, and the mobile node should register in a new FA. Only after this, the moving of the mobile node is completed. In other words, when a mobile node moves to another PCF in a PDSN area in a state where the mobile node is registered for Mobile IP in the PDSN. The PCF traces the mobility. A packet transmitted from an IP network to the mobile node, is transmitted to the PDSN through a HA (Home Agent) according to the IP address of the mobile node. Then, the PDSN transfers the newly registered packet to the PCF through a radio and packet (R-P) interface and the PCF transfers it to the mobile node.

However, when a mobile node moves into another PDSN area, the MIP is registered in the new PDSN. This registration informs the new PDSN of the mobile node's IP address moving. In this case, since a new FA registers in an HA, an IP packet to be transferred to the mobile node is transferred to the new PDSN from the HA and then transferred to the mobile node. Accordingly, loss of the user packets occurs during the delay time because this PPP re-registration cannot be prevented.

That is, generally a contact point corresponding to the PPP of an MS is a PDSN and since the MS has moved into a new PDSN, a process which reestablishes a PPP session should be performed in order to obtain a new PPP contact point.

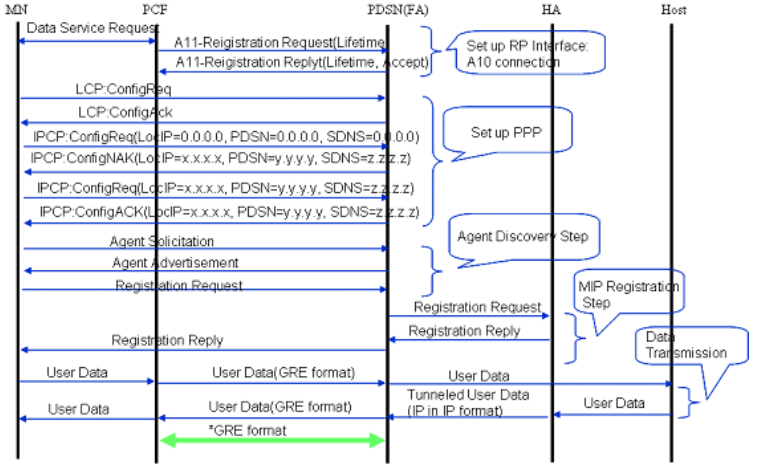


Fig. 2. Standard CDMA2000 call process for a mobile node to initially connect a call

Fig. 2 is a schematic diagram of a signal flow showing a standard CDMA2000 call process for a mobile node to initially connect a call. Referring to Fig. 2, when a mobile node first connects a call, the mobile node transfers a calling message containing a data service request to a PCF. When a call is established, a PCF plays a role to exchange signals and traffic information between an AP (Access Point) and a PDSN. The PCF receiving the message transmits this information to a PDSN to establish RP connection and performs a PPP connection setup procedure. Here, RP

connection means an interface connection between a PCF and a PDSN for signaling (A11) and user traffic (A10). The PDSN performs a FA function in an MIP and an NAS (Network Access Server) function for setting up a PPP with a terminal. At this time, after the PDSN allocates an address to the mobile node, the mobile node completes PPP setup. The PDSN transmits an advertisement message to the mobile node periodically. Through this message, the mobile node can confirm its current Internet contact point.

Meanwhile, when the mobile node receives this advertisement message and transmits a mobile IP registration request (MIP RRQ) to the PDSN, the PDSN and the HA determine whether or not the subscriber is one of those subscribers who are qualified for the MIP support, and then perform authentication. If the subscriber cannot be authenticated, the PDSN includes an error in the mobile IP registration reply (MIP RPL) code, transmits the code to the mobile node and terminates the call. However, if the subscriber is authenticated to perform normal MIP, the PDSN maintains visitor information and by informing it to the mobile node, the registration procedure is finished. Thus, if a PPP is set up and a call is effectively established, actual data communications between the mobile node and a host are performed.

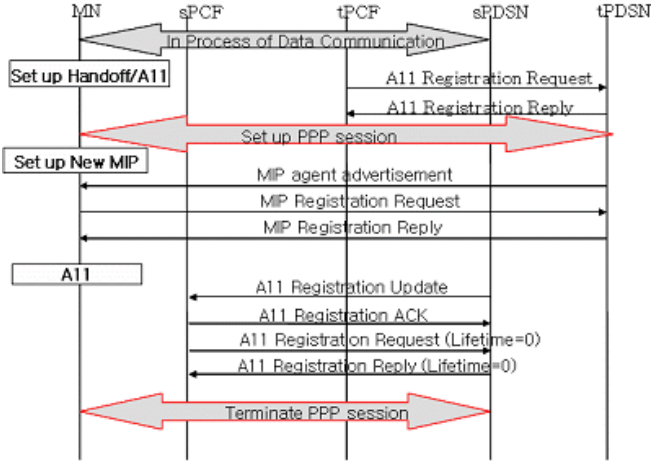


Fig. 3. Handoff procedure of standard CDMA2000 network

A handoff example will now be explained referring to Fig. 3. Fig. 3 is a schematic diagram showing a process to support mobility of a mobile node in which the process of Fig. 2 is performed twice. For the following explanation, a couple of terms is defined. A tPDSN (target PDSN) is a PDSN providing packet services to a newly connected mobile node which is moved into its concerned area. An sPDSN (source PDSN) is a PDSN to which, before moving, the mobile node was connected. It is assumed that at present a mobile node receives data services through sPCF (source PCF) and sPDSN after connected to the Internet. Since the subscriber moves and handoff between sPCF and tPCF (target PCF), and tPDSN occurs, a new RP session and PPP session should be set up. Here, a target PCF and a target PDSN mean a PCF

and a PDSN of a network to which the subscriber should be connected because of the moving. This process will now be explained in detail. First, tPCF transmits A11 RRQ message to tPDSN, and according to this, tPDSN responds to PCF with an RLP message. If this process is successful, a PPP session is reestablished between the mobile node and tPDSN. Then, tPDSN transmits an MIP advertisement message to the mobile node, and according to this, the mobile node transmits an MIP registration request to tPDSN. In response to this, tPDSN again transmits an MIP registration reply to the mobile node, and by doing so, a new MIP is set up. Then, A11 interface for signaling between sPCF and sPDSN is performed and the existing PPP session is terminated. Thus, when handoff between PDSNs is performed, processes for setting up a new PPP session and terminating an existing PPP session are needed such that unnecessary time and resources may be expended.

In this situation, the proposed method can perform handoff in a short time without performing unnecessary PPP reestablishment that may occur in handoff between PDSNs.

3 Proposed Fast Handoff Scheme

A preferred embodiment of the proposed scheme to remove this problem will now be explained referring to Fig. 4. Fig. 4 is a schematic diagram showing a detailed process such that the additional PPP connection process in the handoff method of Fig. 3 can be omitted when a mobile node moved.

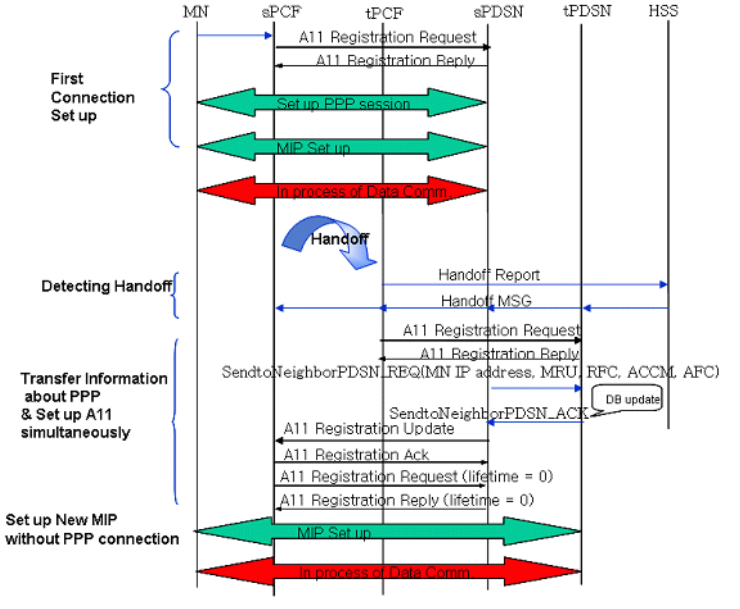


Fig. 4. Signal flow of proposed handoff scheme

Let's consider a situation where a mobile node performs data communications with sPDSN before moving and then moves to another network, so the handoff is needed. When the mobile node is first trying to connect a packet call, the sPDSN makes subscriber information which the subscriber sets initially, shared by all PDSN neighboring the sPDSN. Here, subscriber information indicates an IP address of the subscriber, options defined when ICP and IPCP are tried, and so on. In this state, if subscriber movement occurs, the tPDSN senses that the subscriber moved into its concerned area and transfers the subscriber number and IP address of the mobile node to the sPDSN to which the subscriber was connected before the moving. Upon receiving this, the sPDSN transfers all information on the subscriber to the tPDSN. Upon receiving the information, the tPDSN stores the subscriber information as database items, and performs a procedure related to MIP with the mobile node.

Referring to Fig. 4, the embodiment will now be explained in detail as a signal flow between the mobile node and apparatuses in the wireless data communications network. PCF transmits A11 registration request message to tPDSN, and according to this, tPDSN transmits A11 registration response message to tPCF. Thus, the process for RP setup is the same as in Fig. 3. However, instead of the PPP session setup procedure as shown in Fig. 3, the PPP session related data – e.g., MN IP address, MRU, protocol control field compression (PFC), async. control character map (ACCM), address control field compression (AFC) – are received from sPDSN, to which the mobile node was connected before the moving, and utilized as PPP setup resources of the mobile node. The tPDSN transmits an MIP advertisement message to the mobile node which has moved into its concerned area, and according to this, the mobile node transmits an MIP registration request to tPDSN. In response to this, tPDSN again transmits an MIP registration reply to the mobile node, and by doing so, a new MIP is set up. Then, A11 interface for signaling between sPCF, to which the mobile node belonged before the moving, and sPDSN is updated.

Meanwhile, when the MIP setup is normally allocated, the PPP resources set to sPDSN may request termination or reestablishment by itself without negotiation with the mobile node, and accordingly, to the extent that the information is managed, the information is retained.

4 Performance Evaluation

According to the method for handoff between PDSNs of the proposed scheme as described above, handoff is performed without reestablishing PPP and accordingly handoff between PDSNs can be performed faster by a time (T_s) needed for establishing a PPP session with a terminal and a time (T_t) for terminating a previously set up PPP session. Here, since at least 10 messages are transmitted and received during LCP and IPCP negotiation for establishing and terminating PPP session according to the standard, T_s and T_t are given, respectively, by

$$T_s = 2T_{Lcp_ConfigReq} + 2T_{Lcp_ConfigAck} + 2T_{Ipcp_ConfigReq} + 2T_{Ipcp_ConfigAck} \quad (1)$$

$$T_t = T_{Lcp_TermReq} + T_{Lcp_TermAck} \quad (2)$$

The time reduction (T_r) effect can be achieved as follow.

$$T_r = T_s + T_t + T_{nl} \tag{3}$$

Where, T_{nl} means the network latency time. It depends on processing capacity of network nodes and network traffic loads.

We measured the delay performance of the proposed scheme on our testbed [12]. Fig.5. shows the comparison results of connection time. When the normalized traffic load is 0.5, we can save over 1.1 seconds. Generally, two cases don't show a big difference under the lower traffic load. But under the higher traffic load, our proposed scheme has the better performance by skipping PPP reestablishment and also reducing packet loss during handoff.

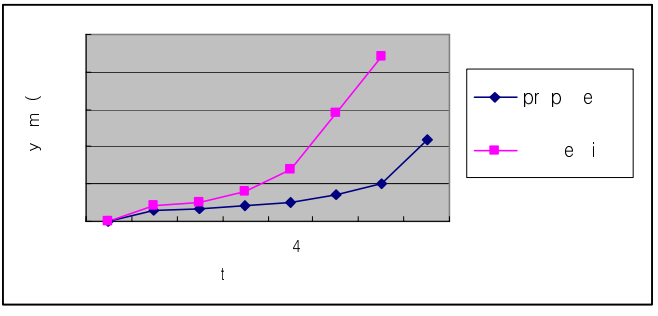


Fig. 5. Comparison of connection delay between the conventional scheme(*diamond*) and the proposed scheme(*square*)

5 Conclusions

In this paper, we have proposed a method for handoff between PDSNs which can reduce signaling latency and user data loss during handoff. This improvement is achieved by exchanging subscribers' information about mobile nodes among their neighbor PDSNs forming a communication network.

When a PDSN recognizes a mobile node moving into its coverage area, it can quickly establish a communication channel with the mobile node based on the received subscriber information. As a result, handoff is performed without reestablishing PPP. Therefore, handoff between PDSNs can be performed faster, removing time needed for establishing a PPP session with a terminal and for terminating a previously set up PPP session.

References

1. 3GPP2/TSG-P, P.S0001-A-1.DOC Version 1.0 Version Date: December 15, 2000
2. Inter-Operability Specification (IOS) for CDMA 2000 Access Network Interfaces (PN-4545), Ballot version, Jun. (2000)
3. C.E. Perkins, (ed.): Ipv4 Mobility Support, RFC 2002, Oct. (1996)

4. 3rd Generation Partnership Project 2.: Wireless IP Architecture Based on IETF Protocols, 3GPP2 P.R0001, Version 1.0.0, Jul. (2000)
5. C. Perkins and D. Johnson: Route Optimization in Mobile IP, IETF draft-ietf-mobileip-optim-09.txt, Feb. (2000)
6. R. Ramjee et al.: IP micro-mobility support using HAWAII, IETF draft-ietf-mobileip-hawaii-01.txt, Jul. (2000)
7. C. Perkins: Mobile-IP Local Registration with Hierarchical Foreign Agents, IETF draft-perkinsmobileip-hierfa-00.txt, Feb. (1996)
8. E. Gustafsson et al.: Mobile IP Regional Registration, IETF draft-ietf-mobileip-reg-tunnel-02.txt, Mar. (2000)
9. A. Campbell et al.: Cellular IP, IETF draft-ietf-mobileipcellularip-00.txt, Jan. (2000)
10. S. Thalanany and A. Singh: Quick handoff scheme in a 3G Wireless Network, IETF draft-thalanany-mobileipqh- 00.txt, Jul. (2000)
11. Hoon Choi, Nader Moayeri: A Fast Handoff Scheme for Packet Data Service in the CDMA 2000 System, GLOBECOM (2001) 1747-1753
12. Eunjun Rhee, Jaehong Ryu, Won Ryu: Implementation of Packet Data Serving Node in the CDMA2000 Network, APIS (2002)

Privacy Protection for a Secure u-City Life

Changjin Lee, Bong Gyou Lee, and Youngil Kong

Graduate School of Information, Yonsei University
134 Shinchondong, Seoul 120-749, Korea
{cjlee,bglee,okay777}@yonsei.ac.kr

Abstract. Recently, projects to construct an innovative brand-new city, termed a u-City (ubiquitous City), are being carried out in many countries. The u-City is a future-oriented city which combines ubiquitous information services with city. It has emerged as an alternative for improving the quality of life for human beings and reaching the balanced development among cities. Although a number of privacy problems are anticipated, there are very few researches on the protection of privacy in the u-City. In this study, we describe the privacy guidelines for modern cities as well as the privacy issues in the u-City, and provide suggestions for privacy protection to establish a safe u-City.

Keywords: Ubiquitous Computing, u-City, Personal Information, Privacy Protection.

1 Introduction

Ubiquitous computing technologies have been applied to many fields including public service, education, health, and transportation. These technologies are being combined with city, and then make it possible to build up a new city, u-City. The u-City is a sophisticated and intelligent city where information can be easily exchanged among people, objects and environments inside the city based on the ubiquitous network. The u-City is aimed to improve the quality of people's life in the city by building up a human-centered city, and to accomplish balanced development among cities by specializing their own strength [1].

However, in the u-City, private information becomes readily available and as such, privacy issues may become a serious problem. Since most projects to construct the u-City are in the very early stage, there is few research on privacy issues in the u-City. In this research, therefore, we look into various issues in privacy violation which should be considered in the process of u-City planning, and then make suggestions for a safer u-City life.

We briefly explain the features of u-City in chapter 2. The privacy issues in the u-City are described in chapter 3. In chapter 4, we provide some suggestions for privacy protection in the u-City. Conclusion and future direction of a secure u-City are presented in chapter 5.

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 685–692, 2007.
© Springer-Verlag Berlin Heidelberg 2007

2 What Is the u-City?

The u-City is different from the Digital Cities which are currently being constructed in many countries over the world. The Digital City collects and organizes the digital information of the corresponding physical city, and provides a public information space for people living in and visiting the city to interact with each other [2].

However, the u-City is a complex of u-Home, u-Work, u-Transportation, u-Environment, u-Health, u-Education, and u-Government based on ubiquitous computing technologies such as USN (Ubiquitous Sensor Network), RFID (Radio Frequency Identification), and FTTH (Fiber To The Home) as shown in Figure 1 [3].

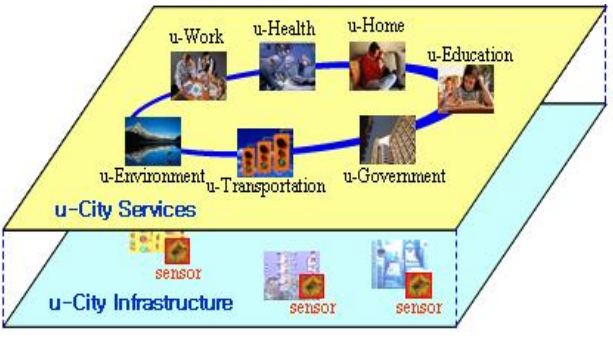


Fig. 1. The Concept of u-City

The differences between a modern city, Digital City and u-City are summarized in Table 1.

Table 1. Modern City vs. Digital City vs. u-City

Modern city	Digital city	u-City
Distance centric	Information centric	Human centric
Spatial, temporal limitation	Spatial, temporal limitation relieved	No spatial, temporal limitation
Real space	Division of real and virtual space	Convergence of real and virtual space
Environment and energy problem	Environment and energy problem	Environment friendly city
Limitation on accommodating new function	Limitation on accommodating new function	Efficient city management
Producer centric	Producer centric	Consumer centric
Unbalanced development among region	Unbalanced development among region	Balanced development among region

3 Privacy Issues in the u-City

3.1 Guidelines for Privacy Protection by International Organizations

Traditionally, 'privacy' is defined in various ways. It does not have specific or definite legal connotation [4] and as such the concept of 'privacy' can be too broad. Thus, this paper will focus on the information privacy related to protection of personal information. 'Privacy' means the right or obligation of an individual or entities to collect, use and disclose personal information [5], and personal information refers to any type of information that identify or can identify an individual or an entity [6].

So far, international efforts have focused on the protection of privacy with regards to collecting, processing and distribution of personal information. OECD was first to provide guidelines on the protection of privacy (Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data) in 1980 as the development of electronic data processing technique made mass data transfer possible [7]. Since then, development of new guidelines on privacy has followed in the UN, EU and APEC with a view to providing their members with references on privacy protection.

The OECD Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data is composed of 8 principles ; 1) Collection Limitation Principle, 2) Data Quality Principle, 3) Purpose Specification Principle, 4) Use Limitation Principle, 5) Security Safeguard Principle, 6) Openness Principle, 7) Individual Participation Principle, 8) Accountability Principle. The Guidelines provide recommendations to the member countries although no binding force exists.

The OECD Guidelines had a direct effect on the privacy guidelines or rules developed later such as the UN Guidelines for the Regulation of Computerized Personal Data Files (1990) [8], EU Directive on the Protection of Individuals with regard to the Processing of Personal Data and on the Free Movement of Such Data (1995) [9] and APEC Information Privacy Principles (2004) [10]. Thus, these guidelines accommodate most of the 8 principles suggested in the OECD Guidelines and build upon the OECD Guidelines by adding new text relevant to each situation.

The OECD Guidelines function as the basic principle for privacy protection and can be seen in the Privacy Laws of many countries such as England, Sweden, Canada, Hong Kong, Australia, and New Zealand. UN Guidelines recommend that member countries reference the guidelines when making privacy laws or procedures of their countries. The EU Directive provides principles and standards for processing personal data and the Directive has been adopted by the domestic laws of its member countries.

Also, in October 2004, APEC adopted guidelines on privacy in October 2004. Based on the OECD Guidelines, the APEC "Information Privacy Principles" added 3 new principles; Preventing Harm , Notice, Choice.

3.2 Privacy Issues in the u-City

The u-City is a new space where physical space and electronic space are integrated into one, quite different from the concept of digital city, in which a modern city or a physical space is simply being reflected or applied to an electronic space. In other words, in the u-City, computers, which are interconnected by ubiquitous networks, are embedded in human beings, objects, and environments and offer further convenience

for residents. However, the side-effect is that private information becomes readily available and as such, privacy issues may become a serious problem. Privacy issues that may appear in the u-City will be reviewed based on the OCED Principles for the Protection of Privacy.

First, in the u-City, there is a high possibility for the Collection Limitation Principle to be violated. One of the key technologies in the u-City is sensing and tag technologies such as RFID, which enable large amounts of private information to be collected in real-time regardless of one's own will. For example, by analyzing RFID information of a purchased item by an individual, one can speculate about the individual's patterns of consumption, social status, current physical location, or health-related information regardless of the individual's will. Thus, personal data can be collected through illegal means without the consent of the subject. Furthermore, under the environment of the u-City, it may sometimes be impossible to notify in advance and ask for consent from the subject for collecting data. As a result, the u-City is highly vulnerable to infringement of the Collection Limitation Principle.

Second, in the u-City, there is also a high possibility of violating the Data Quality Principle. As previously explained in the Collection Limitation Principle, in the u-City, immense amounts of personal data—both necessary and unnecessary—are collected in real-time through computers that are embedded in every object. These data may not all be relevant to the purpose of use, causing inconsistencies in data. This problem can raise problems with the quality of personal data.

Third, the u-City can raise problems with the Purpose Specification Principle. For example, in addition to purpose-specified personal data, other data such as individual location can be collected by identifying the course of movement of an individual through using a telematic service. Also, using RFID chips, sensitive information such as medical conditions or patterns of consumption can also be collected.

Fourth, in terms of the Use Limitation Principle, the u-City is not much different from a modern city since both are exposed to the same danger of using personal data for purposes other than those specified in accordance with the Purpose Specification Principle. The only difference between the two cities is the amount of personal data collected.

Fifth, compared with modern cities, the u-City has a higher possibility of violation of the Security Safeguard Principle. In the u-City, computers are embedded in every device and as each device is interconnected, more personal data is exposed to larger number of people than in modern city. Thus, the danger of personal data being used by malicious users such as hackers increases in the u-City, causing such risks as unauthorized access, destruction, use, modification or disclosure of data.

Sixth, the Openness Principle is also easily violated in the u-City compared to the modern city. In modern cities, which are internet-based, personal data is collected openly mainly through websites or via off-line. Most websites or data controllers notify their subscribers about the policies of protection of private information. However, in the u-City, personal data can often be collected without previous consent or recognition of the data subject and the means to notify the subscribers of policies on private information protection are limited. Thus, in the u-City, the Openness Principles are more difficult to implement than in modern cities.

Table 2. Possibility for Privacy Violation(Modern City vs. u-City)

OECD Principle	Modern city	u-City	Example of privacy invasion in u-City
Collection limitation principle	Low	Very high	Impossible to notify a subject about collection of her or his information
Data quality Principle	Medium	High	Discordance among collected private information
Purpose specification principle	Low	High	Possibility of data analogy without purpose of collection
Use limitation Principle	High	High	Abuse of private information in monitoring and pursuing a subject
Security safeguard Principle	High	Very high	Possibility of security violation by hacking mobile networks
Openness principle	Medium	High	Limited means to notify privacy protection policy to a subject
Individual participation principle	Medium	High	Impossible to identify data controller
Accountability principle	Low	High	Vagueness of data controller and the location of responsibilities

Seventh, in the u-City, a large spectrum of personal data—not only one's identification information but also one's current location, current status and so on—are collected and processed in real-time basis regardless of one's own will. Therefore, it is impossible in reality for the subject to identify data controllers, who have his or her personal data. Thus, in the u-City, the Individual Participation Principle is difficult to adhere to.

Table 3. Privacy Issues in the u-City

Category	Privacy issues
Information collection	Information collection without consent by an information subject
	A breach of notification during the collection of private information
	Collection of private information by illegal means such as hacking
	Collection of private information by illegal monitoring or pressure
Information processing	Outflow of private information by internal data controller
	Outflow of private information by an unauthorized person
	Outflow of private information by careless manipulation
	Outflow of private information by lack of technical measures
	Rejection of an information subject' requests to change and delete their private information
	Retention of private information after duration of use is expired
Information dissemination	Abuse of private information without notice to an information subject
	Information sharing with third party without an information subject's agreement
	Violation of security agreement of private information
	Outflow of sensitive information that can put subjects in danger or hurt the fame
	Illegal appropriation of other's ID
	Intentional outflow of other people's incorrect information

Last, in the u-City, the data controller for specific data is often unclear and difficult to be identified since there are so many channels available to gain access to personal data. Thus, personal data may become poorly managed within the u-City, which in turn, may give rise to infringement of the protection of private information.

Based on the 8 Principles for the Protection of Privacy, Table 2 below summarizes the possibilities of privacy violations that can occur under modern city versus u-City.

Solove (2005) presented the taxonomy of harmful activities on privacy, consisting of information collection, information processing and information dissemination. This taxonomy provides a framework to classify risks with regards to privacy. Thus, privacy issues in u-City can be arranged using the taxonomy as Table 3.

4 Suggestions for Privacy Protection in the u-City

In order to achieve privacy-protected and safe u-City life, measures to protect privacy should be considered in various aspects including regulatory and technological framework based on the privacy issues mentioned on the previous chapter.

4.1 Regulatory and Institutional Suggestions in Building a Safe u-City

Firstly, in order to strengthen privacy protection in u-City, relevant regulation that reflects the specific characteristics of u-City should be introduced to provide with clear procedures for obtaining consent, providing notification, etc. In u-City, personal activities are exposed to pervasive sensors. As it is sometimes difficult to obtain consent or make notification to the data subject in the u-City environment, the problem of personal information control arises.

However, obtaining consent of the data subject in collecting personal data, specifying the purpose for which personal data are collected, and providing notice regarding the collection and use of personal data is the essential principles that should be observed in order to protect privacy. These basic principles for privacy protection should be ensured in the u-City environment as well. In this context, efforts should be paid to seek appropriate measures to protect privacy in the u-City by providing clear process for obtaining consent and making notification that reflects the circumstances of u-City.

Secondly, regulatory framework should be prepared to protect anonymity of delicate personal data that can be collected in the u-City environment. In the u-City environment, delicate information such as a person's locational and circumstantial information can be induced by combining information collected through sensors, RFID Tags. For example, by analyzing the RFID tag of a medical product which is purchased by an AIDS patient, information that the person is HIV–positive can be induced. When such information is disclosed, the data subject can face the danger of being excluded from society as a result of privacy infringement

Thirdly, regulatory tool to protect personal data from development of new technologies should be devised. In the u-City environment, the continuous evolution of ubiquitous computing infra technology can result in infringement of privacy as new devices and media are being developed. However, with the current legal system, it is inevitable to face certain dead zones due to the fast development of technology that

outruns the development of new regulation. Thus, institution of neutral regulation that can protect privacy regardless of emergence of new technology is required.

Lastly, clear legal standard should be provided on the limitation of personal data collection. Privacy infringement can also occur in the u-City environment by service providers' profiling of wide range of personal data through various channels. Therefore, legal framework to regulate profiling personal data should be introduced.

4.2 Technical Suggestions in Building a Safe u-City

Another challenge that should be tackled in building safe u-city besides shaping up new legal and institutional framework is developing new technologies to protect privacy. Alongside the already existing PETs (Privacy Enhancing Technologies) such as P3P, Privacy Policy Statements Generators, cookie control, SW encryption, and anonymity technology, advanced technologies for the u-city environment need to be developed. Continuous R&D efforts for the development of advanced technologies (such as technologies on encryption to prevent information leakage, safe information processing in the wire and wireless network, PETs on data transfer and containment, authentication, and personal identification management) are called for.

In addition, the protection of personal locational and circumstantial information should be ensured in the u-City environment by realizing the integrated PETs that can protect anonymity in addition to the current ID management technology. Anonymity, tracking and traffic analysis are the fields which have not deserved a due consideration yet, but in the u-City environment, these areas are the essential requirement that need to be ensured to protect personal data.

Also, as the existing infra technologies keep developing and new technologies keep appearing, seamless R&D efforts to deal with the novel technological challenge is required to build a safe u-city.

5 Conclusion

Different from traditional cities, the u-City is the forthcoming intelligent city, engrafting highly sophisticated ubiquitous IT technologies such as sensing & RFID technologies, context-aware technology, mobile network technology, and so on. However, due to these advancements, the u-City raises many new problems relating to private information. Because the concept of the u-City is quite new, not only is there a lack of related research in general, but studies on the implication of the u-City on privacy protection are also scarce.

Thus this paper has carefully observed several types of privacy issues that may arise under the environment of the u-City in comparison with traditional cities. It also has suggested legal, institutional and technical complementary measures to solve privacy problems in the u-City. In terms of legal and institutional aspects, measures should be taken to provide previous notification and ask for the consent of the data subject to enhance privacy protection in the u-City. Also, the necessities of taking measures relating to guarantee of anonymity for sensitive information, establishment of value-neutral laws to prepare for new technologies and laws on profiling of

personal data are explained in the paper. For technical aspects, the paper has proposed a future direction of how the PETs should develop under the environment of the u-City.

In addition, for establishing the safe u-City, raising individual awareness for protection of private information, expanding self-regulatory movements, and training specialists for personal data protection should be pursued along with legal, institutional and technical measures. Furthermore, overemphasis on privacy protection may have a danger of undermining the overall efforts to achieve the u-City. Therefore, achieving a stable equilibrium of private information protection and a wise use of personal data for informatization should always be at our focal interest.

References

1. Kim, M. S., New City in the Era of Ubiquitous: u-City, Korean National Assembly Library Press, Vol. 325 (2006) 40-45
2. Ishida, T., Isbister, K. (ed.): Digital Cities: Experiences, Technologies and Future Perspectives, Lecture Notes in Computer Science, Vol. 1765. Springer-Verlag (2000) 350-363
3. Park, J. S., Lim, H. B., The Concept of u-City and Business Strategy, Communication Market, Korea Telecom Management Institute, Vol. 59 (2005) 3-4
4. Solove, D. J., A Taxonomy of Privacy, University of Pennsylvania Law Review, Vol. 154, No. 3 (2006) 477-560, Available at SSRN: http://ssrn.com/abstract=667622.
5. Kim, T. J., Lee, S. W., Lee, E. Y., Privacy Engineering in ubiComp, ICCSA 2005, Lecture Notes in Computer Science, Vol. 3482 (2005) 1279-1288
6. US: US Safe Harbor Privacy Principles (2000)
7. OECD, Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data (1980)
8. UN, Guidelines for the Regulation of Computerized Personal Data Files (1990)
9. EU, Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the Protection of Individuals with regard to the Processing of Personal Data and on the Free Movement of Such Data (1995)
10. APEC, APEC Privacy Framework (2004)

Hybrid Tag Anti-collision Algorithms in RFID Systems

Jae-Dong Shin[1], Sang-Soo Yeo[2], Tai-Hoon Kim[3], and Sung Kwon Kim[1]

[1] School of Computer Science & Engineering, Chung-Ang University, Seoul, Korea
mulli2@alg.cse.cau.ac.kr, skkim@cau.ac.kr
[2] Department of Computer Science & Communication Engineering, Kyushu University, Fukuoka, Japan
ssyeo@itslab.csce.kyushu-u.ac.jp
[3] Division of Computer Information Communication & Engineering, Ewha Womans University, Seoul, Korea
taihoonn@empal.com

Abstract. RFID, Radio Frequency Identification, technology is a contactless automatic identification technology about which a lot of researches and developments are recently progressing. For this RFID technology to be widely spread, the problem of multiple tag identification, which a reader identifies a multiple number of tags in a very short time, has to be solved. So far, several anti-collision algorithms are developed. And those can be largely divided into ALOHA based algorithm and tree based algorithm. In this paper, two new anti-collision algorithms combining the characteristics of these two categories are presented. And the performances of the two algorithms are simulated.

1 Introduction

RFID technology is an automatic identification technology of contactless method that identifies electronic tags attached to goods [1]. For this RFID technology to be widely used, multiple tag identification problem must be solved in the first place. This problem is defined as a one-to-many communication problem between a reader and tags. That has to identify by receiving the information transmitted from tags without collision in case a multiple number of tags exist within the identification area of a reader.

Tag anti-collision algorithms can be categorized into ALOHA based algorithms and tree based algorithms. ALOHA based algorithms usually refer to the slotted ALOHA algorithm, an algorithm in which makes only one tag respond in a slot, in the response of tags, by dividing a time into slot units. On the other hand, tree based algorithms make trees while performing the tag identification procedure using a unique ID of each tag.

In this paper, two new algorithms combining the framed slotted ALOHA algorithm, a typical ALOHA based algorithm, and the query tree algorithm, a typical tree based algorithm, are presented. And the performances of the two algorithms are compared through simulations with existing other anti-collision ones used in RFID systems.

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 693–700, 2007.
© Springer-Verlag Berlin Heidelberg 2007

Forward link	Request	Slot1	Ack 1011	Slot2	Slot3	Slot4	Ack 0111	Request
Return link		1011		Collision		0111		
Tag1				0010				
Tag2						0111		
Tag3		1011						
Tag4				1110				

Frame size = 4

Fig. 1. An example of tag identification process in FS-ALOHA algorithm

2 Related Work

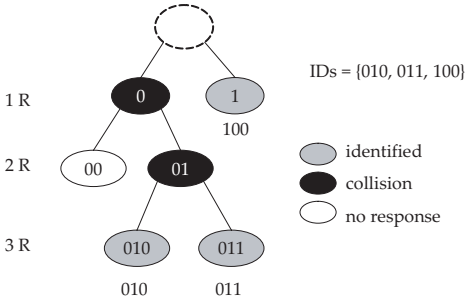
2.1 Framed Slotted ALOHA Algorithm

FS-ALOHA algorithm [2] is the most well-known of anti-collision algorithms used for solving the collision of tags in RFID systems. In FS-ALOHA algorithm, when a reader requests tags to transmit their ID, it also transmits a frame size(FS). On receiving ID transmission request from a reader, a tag randomly decides its own transmission slot within the frame size, and then transmits its ID after waiting until its turn. On the reader side, three kinds of cases can occur. To begin with, there is a case receiving no response to the slot. This is referred to as "no response", and the number of no responses in a frame is expressed as C_0. The second one is a case only one tag has responded. This is referred to as "identification", and the number of identifications in a frame is expressed as C_1. Lastly, there is a case two tags or more attempt to transmit in the same slot. So, a collision takes place and the data transmitted by tags is lost. This is referred to as "collision", and the number of collisions in a frame is expressed as C_k.

Fig.1 illustrates the operation of FS-ALOHA algorithm using four tags. The reader requests tags to transmit their ID along with sending 4 as a frame size, and then each tag selects its own slot and attempts to transmit its ID. In Slot1 and Slot4, only one tag attempted to transmit and thus the reader successfully identifies Tag2 and Tag3, then send the tags an Ack command informing it has identified, in order to keep tags from responding in the next frame. In Slot3, however, as there is no tag, it becomes no response. While in Slot2, an collision occurs since Tag1 and Tag4 sent their ID at the same time. Finishing a frame, the reader requests the retransmission of ID to the remaining tags. At this point, the number of remaining tags is estimated using the C_0, C_1, and C_k, and then the next frame begins by changing to a frame size suitable to the number of remaining tags [3].

2.2 Query Tree Algorithm

QT algorithm [5] is a typical one of tree based algorithms. When requesting tags to transmit their ID, a reader sends a prefix P_k of k bits together.



Query	0	1	00	01	010	011
Response	collision	identified	no response	collision	identified	identified

Fig. 2. An example of identification process in QT [4]

Then each tag confirms whether it is the same as the beginning part of its own ID, and responds its own ID to the reader if it is the same. Likewise as in FS-ALOHA algorithm, at this point, the three cases, "no response", "identification", and "collision", can occur. Here, when a collision takes place, the reader knows there are many tags with the same prefix. In that case, two new prefix P_{k+1} of $k+1$ bits, which "0" and "1" are added at the end of the prefix that has just been transmitted, are made and placed in the queue. The prefix placed in the queue is queried again later. The initial value of the queue is "0" and "1".

Fig.2 is an example of having executed QT algorithm on the assumption of three tags whose identifying IDs are "010", "011", and "100", respectively. Prefix "0" and "1" are set up in the initial queue, and the reader queries tags by taking out prefixes in the queue. For a start, when "0" is queried, Tags "010" and "011" respond at the same time because the prefix is the same as their ID. Then the reader judges there are two tags or more starting with "0" and enters "00" and "01" in the queue. Thereafter, prefix "1" is taken out from the queue and queried. Since there is only one tag, "100", it is identified normally to the reader. The first round ends by the method like this. And another round begins for "00" and "01" which has been previously placed in the queue. This algorithm ends when the queue is empty.

3 Hybrid Anti-collision Algorithms

In FS-ALOHA and QT algorithm explained above, when there are many tags a reader wants to identify, there exists many tags that will respond at the same time and thus many collisions take place. This ultimately makes the time of identifying tags longer. Framed query tree(FQT) algorithm and query tree ALOHA(QT-ALOHA) algorithm presented in this paper are hybrid forms of FS-ALOHA algorithm and QT algorithm.

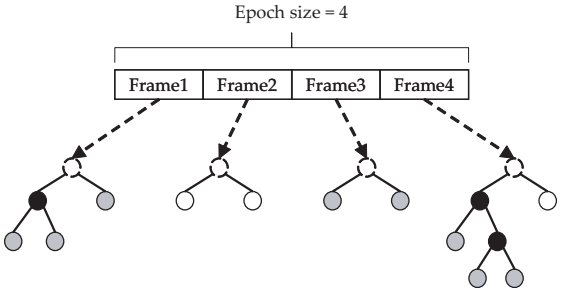


Fig. 3. An example of identification process in FQT algorithm

3.1 Framed Query Tree Algorithm

FQT algorithm divides tags randomly into frame units. And within this unit, tags are identified using QT algorithm for them. The actual operation is as follows:

When requesting tags to transmit their ID, a reader also send an epoch size, the number of the total frames. Then each tag decides its own participating frame randomly and responds only when the reader queries its own frame. The reader executes its identification process within each frame using QT algorithm. However, the reader transmit to tags the number of frame as well as the prefix of ID. Each tag confirms whether the number of frame is the same as its own selected number of frame, and if it is the same, transmits its own ID when looking the prefix and it is consistent, as in the existing QT algorithm. After identifying all tags within a frame through QT algorithm identification process in the frame, the reader proceeds to the next frame. This process is carried out repetitively for every frame.

The example of Fig.3 shows the identification process of FQT algorithm when the number of tags to be identified is 8 and the epoch size is 4. To begin with, a reader transmits the epoch size to tags and then each tag selects its own frame randomly. For three tags which selected Frame1, QT algorithm is used for identifying tags. Thereafter, tags are identified while proceeding to the next frames by the same method.

To determine the most appropriate epoch size is very important in improving performance. By intuition, an epoch size which a tree depth will not exceed 2 when executing QT algorithm exhibits the best performance. The reason is that when executing "0", "1" in the initial queue of QT algorithm, the case that two tags in total, one tag beginning with "0" and another tag beginning with "1", are identified is the most ideal. Frame3 is exactly the best case. Assuming the number of tags to be identified is N and the epoch size is ES, the most ideal ES can be expressed as follow:

$$N = 2 * ES$$

This can be actually verified through the simulation in section 4.1. But there is a big problem with this case. For it is difficult to determine an suitable epoch size from the beginning since the tag identification procedure is initiated under the condition not knowing N, in other words, how many tags are to be identified. From that reason, the final FQT algorithm uses FFT(First Frame Test). The FFT begins from a small epoch size and stops its identification process when the first frame has collisions exceeding a collision threshold, and then resumes its identification of tags by increasing the epoch size. As it is assumed that all tags are randomly divided in frames, if many collisions occur in the first frame, remaining frames are more likely to have such a trend.

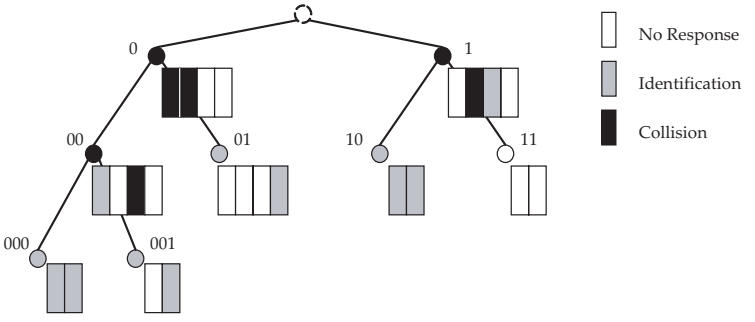
As shown in the above best case, Frame3 in Fig.3 , a tree can have the best performance when its tags are two with the depth being 1. Thus a threshold is needed in order to prevent the tree from becoming deeper than this. The collision threshold is a constant based on the concept that the more collisions happen, the deeper the tree becomes. As a result of simulations conducted many times, we have verified that an appropriate epoch size is approached faster than any other cases when this collision threshold is set at 3. In case it is smaller than 3, even when the epoch size is adequate, the epoch size becomes big and thus may be passed over. On the contrary, when it is bigger than 3, the epoch size gets too big and thus the speed to increase to an suitable epoch size becomes too slow. So the collision threshold is assumed to be 3 in this paper.

3.2 Query Tree ALOHA Algorithm

QT-ALOHA algorithm is another hybrid form of FS-ALOHA and QT algorithm. FQT algorithm basically implements FS-ALOHA algorithm and uses QT algorithm as actual tag identification process. On the other hand, in QT-ALOHA algorithm, QT algorithm is a big picture, while actual tag identification process progresses with FS-ALOHA algorithm. This operation is as follows:

On requesting tags to transmit their ID, a reader sends a prefix and a frame size together. Then only tags which are consistent with their own prefix proceed to FS-ALOHA algorithm with the transmitted frame size. And in the progress of FS-ALOHA algorithm, if a collision takes place even in a single slot, it is interpreted as a collision of QT algorithm, and then a new prefix is made and entered in the queue. At this point, a difference from QT algorithm is to calculate a frame size [3] to be transmitted next and also place this in the queue.

Fig.4 is an example of QT-ALOHA algorithm. It is assumed that the number of tags to be identified is 8 and the initial frame size begins from 4. In the first round, the reader transmits to tags a prefix "0" and a frame size 4. And collisions have occurred in the frame. Then, "00" and "01" are entered in the queue and the frame size is determined as 4 through the calculation process of frame size. And then "1" is taken out from the queue. When the queue becomes empty, the algorithm ends.



Query	0	1	00	01	10	11	000	001
Frame Size	4	4	4	4	2	2	2	2
Response	collision	collision	collision	identified	identified	no res.	identified	identified

Fig. 4. An example of identification process in QT-ALOHA algorithm

4 Simulations

The ID of tags used in simulations was set at a size of 64 bits according to the international standard and was created using random number generator. And for raising the reliability of tag simulation results, 100-time simulations were conducted for every same environment and average it.

4.1 Epoch Size of FQT Algorithm

The first simulation was conducted with the epoch size being changing under the condition of 100 tags. Table.1 shows the results of this simulation. Reviewing this results, they can be divided into three cases according to the epoch size. The first case is at 32 and 64, close to about 50, an ideal epoch size. This case demonstrates the best performance as we mentioned in section 3.1. The second is a case smaller than the ideal epoch size. But in this case, an optimum epoch size is found by the first frame test. And the last is a case bigger than the ideal epoch size. As there is no operation to reduce the epoch size in FQT algorithm and thus the initial epoch have to be entirely executed.

Table 1. Performance comparison according to Epoch Size(ES)

Frame Size	Query	C_0	C_1	C_k
16	246.6	55.1	100.0	81.2
32	238.2	56.9	100.0	78.2
64	241.3	85.9	100.0	54.1
128	327.2	193.6	100.0	32.4
256	550.2	430.6	100.0	18.6

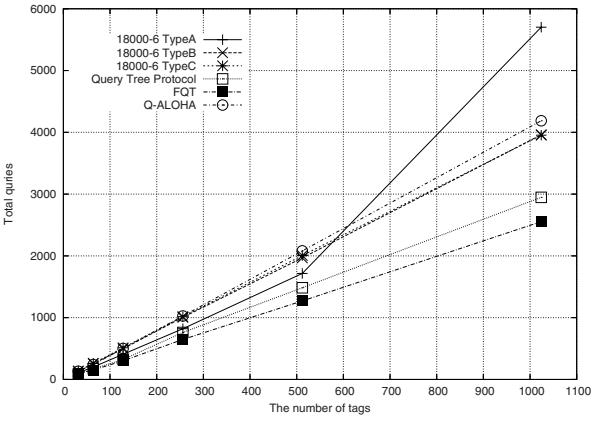


Fig. 5. Comparison of query-response number

4.2 Performance Comparison Between Presented Algorithms and Other Ones

The second simulation has compared the presented FQT algorithm and QT-ALOHA algorithm with other anti-collision algorithms, 18000-6 [6] Type A, Type B, Type C, and QT algorithm. This comparison sets the query-response number of times between a reader and tags as a comparison value while changing the number of tags from 32 to 1,024. On the other hand, in the algorithms based on FS-ALOHA algorithm, such as Type A, Type C and QT-ALOHA algorithm, the initial frame size is started arbitrarily at 32 regardless of the number of tags. The epoch size in FQT algorithm also begins at 32.

Fig.5 and 6 are the results of simulations. 18000-6 Type A and Type C use the same operating method, FS-ALOHA algorithm. But as the maximum frame size of Type A is 256, if executing the algorithm at the number of tags higher than it, a significant degradation in performance can be seen. Type C, however, if there occurs many collisions or no-responses even in the middle of a frame, stops the on-going frame and proceeds to the next frame, thereby improving the performance. Type B uses binary tree algorithm, one of the tree based algorithms.

The results show that the tree based algorithm typically is lower in the query-response number of times than the ALOHA based algorithms. For tags identified in a frame of the ALOHA based algorithm send tags a command they have been identified so they cannot be included in the identification process of the next frame.

Fig.6 illustrates how many queries-responses are needed in order to identify one tag. Looking into this, it can be seen that FQT algorithm queries less for identifying one tag than any other algorithms. Its performance improvement is 10 to 50 percent of times than many other existing anti-collision algorithms.

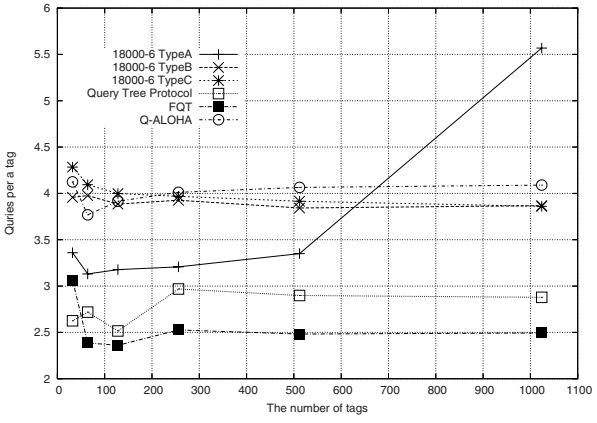


Fig. 6. Comparison of query-response number needed for identification per tag

5 Conclusions

Two new algorithms combining the characteristics of the two categories, ALOHA based and tree based algorithm, have been presented in this paper. FQT algorithm of the two, in particular, has shown a big performance improvement. In the near future, methods to find a faster optimum epoch size in FQT algorithm are to be studied.

Acknowledgement

This work was supported by grant No. R01-2005-000-10568-0 from the Basic Research Program of the Korea Science & Engineering Foundation.

References

1. K. Finkenzeller, "RFID handbook", *John Wiley & Sons*, 1999.
2. F. C. Schoute, "Control of ALOHA Signalling in a Mobile Radio Trunking System", *International Conference on Radio Spectrum Conservation Techniques*, IEEE, pp.38-42, 1980.
3. H. Vogt, "Multiple object identification with passive RFID tags", *In IEEE International Conference on Systems, Man and Cybernetics (SMC'02)*,October 2002.
4. J. Myung and W. Lee, "An Adaptive Memoryless Tag Anti-Collision Protocol for RFID Networks", *IEEE 24th Conference on Computer Communications (INFOCOM'05)*, March 2005.
5. C. Law, K. Lee, and K. Siu, "Efficient Memoryless Protocol for Tag Identification", *4th International Workshop on Discrete Algorithms and Methods for Mobile Computing and Communications*, pp.75-84, ACM, August 2000.
6. *ISO/IEC 18000-6:2004/Amd 1:2006*, International Organization for Standardization, June 2006.

Design and Implement Controllable Multicast Based Audio/Video Collaboration

Xuan Zhang[1,2], Dongtao Liu[1], and Xing Li[1,2]

[1] Network Research Center
[2] Department of Electronic Engineering
Tsinghua University ,Beijing China, 100084
zhangx@cernet.edu.cn , ldt@mails.tsinghua.edu.cn,
xing@cernet.edu.cn

Abstract. Multicast based audio/video collaboration system is one of representative applications in next generation internet. Adopting multicast technique could save bandwidth for multipoint-to-multipoint audio/video communication. Lack of ubiquitous native multicast limits the application, and current multicast based A/V collaboration systems lack effective control mechanism. This paper introduces one controllable audio/video collaboration system based on multicast. The control and management on audio/video collaboration are presented. The system has been implemented and applied on CERNET.

Keywords: Controllable collaboration, audio/video, multicast.

1 Introduction

The multipoint audio/video collaboration systems have played important roles in Next Generation Internet [1][2]. Among collaboration techniques IP multicast has advantage on saving bandwidth for group communication. This makes it advanced comparing to centralized system on large scale multi-party A/V collaboration.

Some IP multicast based A/V collaboration applications Access-grid[3] INDIVIA[4], have been employed in internet. But problems remain to be solved to these current systems. The control and manage on current multicast-based A/V collaboration systems are not effective and not easy to be employed when the scale is large, for example, the voice echo and noise problem during collaboration.

This paper introduces the IP multicast based A/V collaboration system. The control and management on multicast based audio/video tools are presented, performance monitor and congestion control are described. We have implemented and applied the controllable audio/video collaboration system on CERNET successfully.

2 Control and Management on Audio/Video Collaboration

In multicast based audio/video collaboration system, each user could receive all video and audio streaming from other users in the group. Users talk to each other equally, there are no centralized control or manage. This equality in many-to-many advances

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 701–704, 2007.
© Springer-Verlag Berlin Heidelberg 2007

interactive quality for collaboration. On the contrary, lack of control and management cause problems during collaboration such as noise or echo. How to focus current speaker's video among the dazzling video windows of the group is another issue.

To manage and control the participants' audio/video in collaboration, we propose one mechanism based on chairman control panel to control and manage the A/V during collaboration. The control panel is managed by chairman. During session course the control panel could monitor the audio/video states and control video/audio by sending control message. The messages are sent via unicast or multicast way according to request.

The multicast based audio/video tools we adopted are originally from rat/vic [5].We modified the tools to meet the goal of management and control.

2.1 Control and Manage on Audio

In audio collaboration, users in group could speak equally and freely. Avoiding the unwanted voice such as noise or echo is hoped. Chairman or administrator should be able to control it when abnormal voices happened. Here we adopt the scheme combining voice volume monitoring and remote control as figure 1.

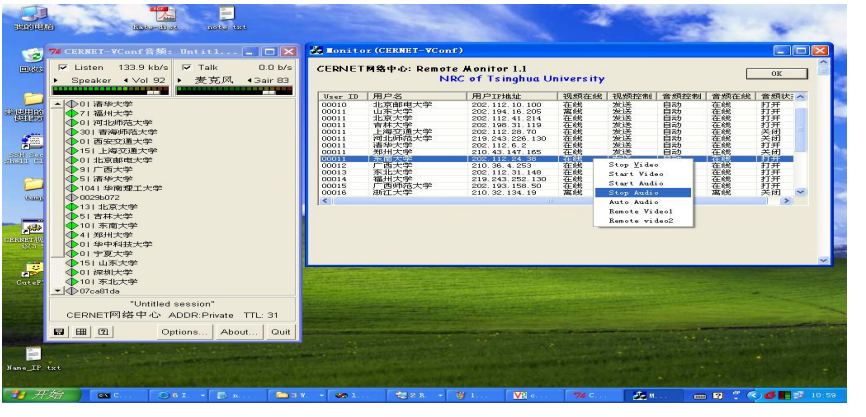


Fig. 1. The chairman control panel with audio monitor interface

The left of figure 1 presents the audio monitor interface, the users' voice volume values ranging from 0 to 16383 are displayed in user-list. We call the voice volume as *powermeter*. According to *powermeter* chairman could judge which user is speaking or making noise. When any speaker makes noise or echo, chairman could mute the speaker remotely by sending *Audio/mute* message to the speaker from control panel. Figure3 shows how chairman mutes user (202.112.24.38) remotely via control panel.

2.2 Control and Management on Video

During multicast based video collaboration, users could receive all users' video streaming and see all the others' video windows simultaneously. When the number of video windows is large, how to focus the current speaker video window is necessary.

Commonly, the current speakers, chairman and local video should be focused and their video windows should be enlarged.

Figure 2 shows one layout for video windows, the main speakers, local video and slider window are enlarged, the other participants' video remain as stamp windows.

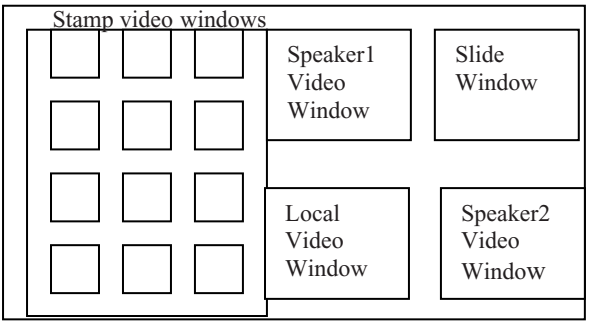


Fig. 2. Video window layout during collaboration

How to build and maintain the focusing-based video layout automatically for all users is important for group collaboration. Typically, when the current speakers change, the enlarged video windows for speakers should be changed accordingly to all users synchronously.

We implement the video layout by defining four types of Enlarged Video Windows (EVW) class: *EVW/local, EVW/slide, EVW/speaker1* and *EVW/speaker2*. The EVW classes are identified by video source identification SSRC. Among them *EVW/local* and *EVW/slide* are usually constant and can be designated at beginning.

But *EVW/speaker1* and *EVW/speaker2* would change during collaboration course. We d\efine messages *Video/speaker1* and *Video/speaker2* (as table 1) as EVW control message for *EVW/speaker1* and *EVW/speaker2*. When one new speaker begins to talk, the chairman could select the new speaker from the user list in control panel (as figure 1), and send the current speaker's SSRC to all users by sending *Video/speaker1* and *Video/speaker2* message. When end user receives the newly *Video/speaker* messages, the end system would change their current speaker EVW automatically according to received control message. The *Video/speaker1* and *Video/speaker2* messages are multicast to all users in the group, so all the users could switch their speakers' EVW synchronously. Figure 3 shows one instance of video layout with EVW and stamp video, during one forum on IPv6.

Table 1. Control messages for video

Message type	Description
Video/bit-rate-send	bit rate of video sending for flow control
Video/speaker1	EVW control message for *EVW/speaker1* with SSRC
Video/speaker2	EVW control message for *EVW/speaker2* with SSRC

3 Application Cases

The multicast based audio/video collaboration systems have been employed to more than 38 cities covering all provincial capitals of China. In 2006, some checking and report meetings for research project were held via A/V collaboration system. Figure 3 shows one scene on IPv6 forum via the collaboration system on CERNET.



Fig. 3. Instance of video layout with EVW, one application case on IPv6 forum

4 Conclusion

The controlling and managing are important for many-to-many A/V collaboration. In this paper, we introduce one controllable multicast based audio/video collaboration systems. The audio/video collaboration controlling and managing mechanisms are discussed. We have implement and applied the collaboration systems on CERNET. The application would extend to CERNET2 (the Chinese next generation internet).

References

1. Internet2 consortium, http://internet2.edu/ .
2. Geoffrey Fox, Wenjun Wu, Ahmet Uyar, Hasan Bulut,Shrideep Pallickara, “Global Multimedia Collaboration System”, 1st International Workshop on Middleware forGrid Computing, Rio de Janeiro, Brazil, (June 2003)
3. Access Grid Project, http://www.accessgrid.org
4. W.T. Ooi, P. Pletcher, and L.A. Rowe,INDIVA: Middleware for Managing a Distributed Media Environment, SPIE Multimedia Computing and Networking, (January 2004).
5. http://www-mice.cs.ucl.ac.uk/multimedia/software/

Solving a Problem in Grid Applications: Using Aspect Oriented Programming

Hyuck Han, Shingyu Kim, Hyungsoo Jung, and Heon Y. Yeom

School of Computer Science and Engineering,
Seoul National University,
Seoul 151-742, Korea
{hhyuck,sgkim,jhs,yeom}@dcslab.snu.ac.kr

Abstract. Aspect Oriented Programming (AOP) was introduced 10 years ago and many research projects have focused on broadening AOP and its target areas. However, few applications in the Grid computing world adopt AOP in comparison with very vigorous research of AOP. Therefore, we present a case study that covers a general networking problem in the Grid computing. AOP provides a novel solution of the problem without modifying existing source code. Aspects that we define are simple, intuitive and reusable. We believe that our implementation is very useful in developing other Grid computing software platforms, and AOP can be a powerful method in modularizing source codes and solving problems of software architectures.

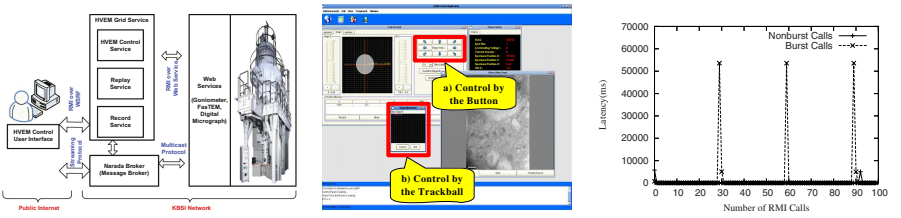
1 Introduction

Many scientists in e-Science [1] currently utilize computing resources as part of their research and will utilize more powerful computing resources across the Grid [2] infrastructure. They also have access to very large data sets and are able to perform real-time experiments through Grid applications. Distributed global collaborations over the Internet such real-time experiments require high bandwidth and low latency. However, firewalls in front of networks of research institutes often show unexpected long latency because many SOAP (Messaging Protocol of Grid computing) messages in a short period can be regarded as a DoS attack. Therefore, we devised a solution to overcome the problem. Then, we added a new module without any modification of legacy software by utilizing the Aspect Oriented Programming (AOP) [3] technique.

AOP is a new technology for separating crosscutting concerns that are usually hard to do in object-oriented programming (OOP). AOP complements OOP by allowing the developer to dynamically modify the static OO model to create a system that can grow to meet new requirements. Just as objects in the real world can change their states during their lifecycles, an application can adopt new characteristics as it develops.

In this paper we describe our AOP-based solution to overcome the firewall problem. Aspects that we define are simple, intuitive and reusable enough to be applied to other Grid applications.

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 705–708, 2007.
© Springer-Verlag Berlin Heidelberg 2007



(a) System Architecture (b) User Interface (c) Latencies of RMI

Fig. 1. HVEM Grid System and its Problem

2 Overall System

2.1 Motivation

High-Voltage Electron Microscope (HVEM) allows scientists to see objects at a magnification greater than the actual sample, and we use HVEM that the Korea Basic Science Institute (KBSI)[4] operates. Figure 1(a) shows the architecture of the HVEM Grid System. Our system consists of three tiers - HVEM Control User Interface (HVEM Control UI), HVEM Grid Service and Web services which encapsulate Vendor-provided applications. XML messaging over HTTP is used to communicate between HVEM Control UI and HVEM Grid Service. KBSI operates firewalls between the public internet (HVEM Control UIs) and the KBSI network (HVEM Grid Service and legacy systems). Figure 1(b) shows HVEM Control User Interface. Part (a) calls HVEM Grid Service for each mouse click, and part (b) calls it continuously while trackball is moving (numerous and continuous calls in a short period). In Figure 1(c), lines noted as Nonburst Calls and Burst Calls correspond to latencies of RMI by part (a) and part (b) respectively. In case of Burst Call, there are some peaks exceeding 50 seconds at 28th, 57th and 90th call. This is due to the firewall which is located in front of HVEM Grid Service.

When the transport layer of SOAP is HTTP, the flow of each RMI call follows.

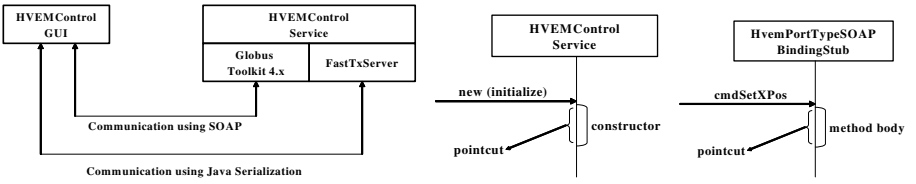
- Connect to host which runs Globus Toolkit
- Transfer XML-based SOAP messages to the corresponding connection
- Globus Toolkit invokes the requested method of the requested service.
- Globus Toolkit sends results of the method call to the connection
- Disconnect the connection

The burst of RMI calls by part (b) in Figure 1(b) suggests that the number of connection requests to the server host increases. The peaks in Figure 1(c) come from the feature that the firewall regards the burst as a DoS attack.

2.2 Improvements

Our primary goal of this study is to guarantee short and coherent latencies, but it is also important to minimize modifications of HVEM Grid Service and

HVEM Control User Interface. To achieve these goals, it is the best approach to add a new transport layer in the same level in which skeleton and stub are located in HVEM Grid Service and in HVEM Control User Interface.



(a) Improved System Architecture (b) Server-side pointcut (c) Client-side pointcut

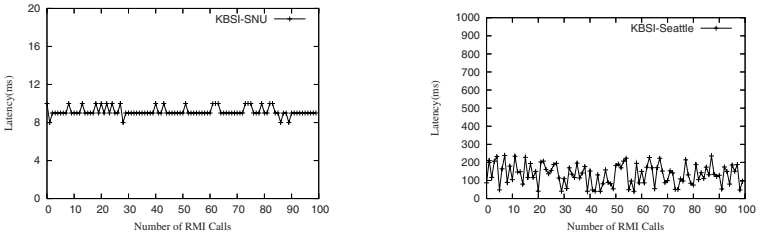
Fig. 2. Overall System

Figure 2(a) illustrates the improved system. In this system, communications between HVEM Control User Interface and HVEM Grid Service can use two ways. One is to use SOAP and HTTP and the other is to use Java Serialization. The former is for part (a) and the latter for part (b) in Figure 1(b). The server which utilizes Java Serialization is called FastTxServer.Unlike Globus Toolkit, the FastTxServer does not disconnect client after it responds to a request. This feature makes the firewall not to consider continuous calls as a DoS attack. To meet the goal of minimizing modifications, we capture execution points and add new functions in these points using AOP. In other words, we capture executions of HVEM Grid Service and HVEM Control User Interface, and replace them with desired methods. Figure 2(b) and 2(c) describe pointcuts in the program flow. In server-side, the constructor of HVEM Grid Service which captured as a pointcut is passed to the FastTxServer as argument. Then, the FastTxServer can execute a method through Java Reflection when a method call request is arrived at the FastTxServer. In client-side, entire methods, which generate the firewall problem, are captured as pointcuts. Then, invocations based on SOAP over HTTP are replaced with calls to the FastTxServer.

3 Evaluation

We evaluated our improved system using 3 PCs in Seoul National University (SNU), KBSI and Seattle. The machines were Pentium 4 2.80GHz in SNU and Seattle, and the machine was Pentium 4 3.0GHz in KBSI, running Linux 2.4.18. HVEM Grid Service server was in KBSI machine, and HVEM Control UIs were in SNU and Seattle. In case of Seattle, clients and the server are connected over the GLORIAD network[5].

Figure 3 shows latencies of RMI with optimization in stub and skeleton levels. Latencies were measured by calling RMI 100 times sequentially. The average latency was 10ms in SNU, and 120ms in Seattle. These short latencies are due



(a) Client in SNU (b) Client in Seattle

Fig. 3. Latencies of RMI with Improvement in Stub and Skeleton level

to the lack of XML translation in Java Serialization. In addition, firewalls of KBSI do not recognize any calls as DoS attacks. By this improvement, overseas users can use our HVEM Grid Service with a reasonable speed.

4 Conclusion

AOP is regarded as a powerful method for modularizing software and solving problems of software due to its attractive features. This article shows a case study that covers a firewall problem in Grid computing. We devised a AOP-based solution to overcome the problem without modification of legacy software or existing services and our results show that our improved system guarantees fast and coherent latencies.

Acknowledgment

The ICT at Seoul National University provides research facilities for this study.

References

1. Oxford e-Science Centre: (e-Science Definitions) `http://e-science.ox.ac.uk/public/general/definitions.xml`
2. Foster, I., Kesselman, C., Tuecke, S.: The anatomy of the grid: Enabling scalable virtual organizations. International Journal of SuperComputer Applications **15**(3) (2001)
3. Palo Alto Research Center: (The AspectJ(TM) Programming Guide) `http://eclipse.org/aspectj/`.
4. Korea Basic Science Institute: (KBSI microscopes & facilitates) `http://hvem.kbsi.re.kr/eng/index.htm`.
5. GLORIAD Korea: (GLORIAD-KR) `http://www.gloriad-kr.org/eng/index_eng.htm`.

Energy-Aware QoS Adjustment of Multimedia Tasks with Uncertain Execution Time*

Wan Yeon Lee[1], Heejo Lee[2], and Hyogon Kim[2]

[1] Hallym University, Chunchon 200-702, South Korea
[2] Korea University, Seoul 136-701, South Korea
wanlee@hallym.ac.kr, {heejo,hyogon}@korea.ac.kr

Abstract. In order to make the best use of available energy budget, we propose a QoS adjustment method which maximizes the total QoS-provisioning of multimedia tasks with uncertain execution time. This method utilizes the probability distribution of task's execution time to determine an instant QoS level. Our experiments show that the proposed method gives 52% more QoS-provisioning than the conventional method using a constant QoS level derived from the worst-case time.

1 Introduction

A necessary feature for the mass computation of multimedia applications on wireless electronic devices is acceptable battery lifetime. If critical computation suddenly stops before its completion due to the shortage of available energy, it may result in great loss. Generally, we can reduce the energy consumption rate of battery and thus extend the battery's lifetime by decreasing the computation amount in a running application. In most cases, however, more computation is necessary to provide better quality of service(QoS). As a result, there is a demand to control both the computation amount of an application and its battery lifetime. Another issue to be considered is the uncertain execution time of multimedia applications. Their execution time heavily depends on their input data, but the information of input data is not available before starting execution.

In this paper, we propose an energy-aware QoS adjustment of multimedia tasks with uncertain execution time, which guarantees their worst-case execution time and maximizes the total QoS-provisioning gained from performing their computation in limited-energy environments. The proposed method assigns the highest QoS level to the earliest processing part of a task and decreases the QoS level gradually as the task progresses its execution. Even though the later running parts provide lower QoS level when being executed, they are rarely performed due to their lower probability. Statistically, this can support more QoS-provisioning than assigning a constant QoS level derived from the assumption that the task is always executed for the worst-case execution time.

* "This work was supported by Hallym University Research Fund, 2007(HRF-2007-039), the ITRC program of the Korea Ministry of Information & Communications, and the Basic Research Program of the Korea Science & Engineering Foundation."

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 709–712, 2007.
© Springer-Verlag Berlin Heidelberg 2007

Graceful degradation [1] or stopping the unimportant task [2] can allow mission-critical tasks to run for a longer period of time when available energy is low. These methods considered the energy management of tasks with fixed execution time but not that with uncertain execution time. Lee *et al.* [3] addressed a similar method to maximize QoS-provisioning while using the limited amount of energy. They proposed a general approach, however, there is no relationship between QoS and energy consumption.

2 Preliminaries

The notion of QoS in multimedia tasks includes various characteristics such as resolution, delay, jitter, loss rate, etc. Among several properties of QoS, the amount of data computed by a task is referred to as *fidelity*. For example, the display size of a video player or the amount of lossy compression applied to a video stream can be the fidelity of the video streaming task. Experimental results [1,4] showed that the energy consumption of a multimedia task is proportional to its fidelity such as the image size being processed, the transmission rate on wireless networks, and its running time.

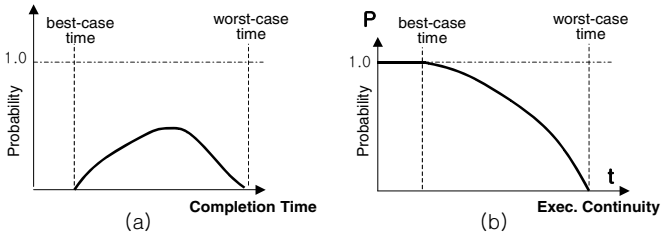


Fig. 1. Probability distribution of task's execution times

The systems can be designed for scheduling multimedia tasks with their worst-case execution times, resulting in significant waste of energy. Approaches based on the average case are likely to suffer from lacking resource, particularly in the worst case. More accurate models are based on the probability distribution of execution times. Figure 1(a) shows an example of probability distribution of task's execution times and Figure 1(b) shows the tail distribution of its cumulative probability distribution, denoted as P_t at a time t. P_t is the probability that the task continues its execution for at least time t.

3 Proposed Method

In this paper, we study how to maximize the fidelity benefit of a single multimedia task with uncertain execution time on a limited-energy system. It is assumed

that the other tasks running on the system have nearly fixed execution times and their operations are stable in most cases. Then, we can formulate the total energy consumption amount of all tasks as follows:

$$\int_0^{T_w} m \cdot \mathcal{F}(t) \cdot dt + b \cdot T_w$$

where T_w is the worst-case execution time, m and b are an application-specific coefficient associated with the multimedia task and a constant coefficient associated with background tasks respectively, and $\mathcal{F}(t)$ is the instantaneous fidelity of the multimedia task at a time t. When E_{max} denotes the amount of available energy budget, $\int_0^{T_w} m \cdot \mathcal{F}(t) \cdot dt + b \cdot T_w \leq E_{max}$. For simplicity, we define B_{max} as $\int_0^{T_w} \mathcal{F}(t) \cdot dt \leq \frac{E_{max} - T_w \cdot b}{m} = B_{max}$.

As a benefit measure of fidelity, we define *Perceptional Resolution* of a task as the resolution of 2-dimensional images computed by the task. User perception against 2-dimensional images is proportional to the square root of their sizes. The problem dealt with in this paper is to maximize the total Perceptional Resolution of a task during its execution, subject to the constraint that the amount of total energy consumption during its execution is no larger than the given energy budget. The problem can be formulated as follows:

$$Maximize \int_0^{T_w} \sqrt{\mathcal{F}(t)} \cdot P_t \cdot dt = \int_0^{T_w} (\mathcal{F}(t)/P_t^2)^{1/2} \cdot P_t^2 \cdot dt \tag{1}$$

subject to

$$\int_0^{T_w} \mathcal{F}(t) \cdot dt = \int_0^{T_w} (\mathcal{F}(t)/P_t^2) \cdot P_t^2 \cdot dt \leq B_{max}. \tag{2}$$

By Jensen's inequality [5], this maximization occurs when all values of $\mathcal{F}(t)/P_t^2$ are the same. Then Equation (1) has an upper bound as follows:

$$\int_0^{T_w} \sqrt{\mathcal{F}(t)} \cdot P_t \cdot dt \leq \sqrt{C} \cdot \int_0^{T_w} P_t^2 \cdot dt = \sqrt{B_{max}} \cdot \sqrt{\int_0^{T_w} P_t^2 \cdot dt}$$
$$\text{if and only if } \mathcal{F}(t) = C \cdot P_t^2 = \frac{B_{max}}{\int_0^{T_w} P_t^2 \cdot dt} \cdot P_t^2.$$

Since P_t always decreases as t increases, a task decelerates its fidelity as its execution goes on in the optimal schedule. The previous work [1,6] showed that the overhead to dynamically decelerate fidelity is negligible. This approach can be applicable to similar problems which try to maximize or minimize another metric of QoS, instead of Perceptional Resolution.

4 Evaluation

The proposed method determines an instant QoS level based on the distribution of task's execution times while the conventional method determines a constant QoS level based on the worst-case execution time. For evaluation metric, we define *QoS Increment* as $\frac{PS_p - PS_w}{PS_w} \times 100$, where PS_p and PS_w are the total Perceptional Resolutions in the proposed method and in the conventional method,

respectively. We implement the proposed method and the conventional method on a practical multimedia application and compare their performance. In these experiments, we consider the case that a user enjoys watching a live broadcasting of the 2006 Major League Baseball(MLB) on a mobile device with limited energy budget. Figure 2(a) shows the playing time distribution of the former 83 games and the latter 83 games in the 2006 season of the New York Yankees team [7]. Figure 2(b) shows the average performance of the latter 83 games when they utilizes the proposed method based on the distribution information of the former 83 games. The longest playing time of MLB (i.e., 486 minutes) is used for the worst-case playing time. These experiments show that the proposed method provides 52% QoS Increment when T_w is 486 minutes. As the value of T_w is decreased, its performance goes down but the risk to irresistibly stop the broadcasting in the middle of games is also increased.

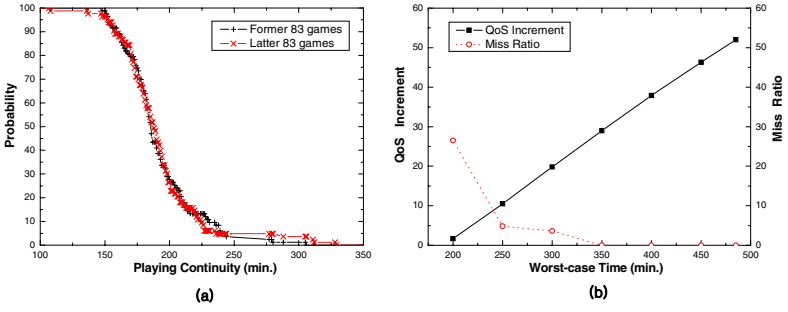


Fig. 2. Experiment results of a multimedia application

References

1. Flinn, J., Satyanarayanan, M.: Managing battery lifetime with energy-aware adaptation. ACM Trans. Comp. Syst. **22**(2) (May 2004) 137–179
2. Tamai, M., Sun, T., Yasumoto, K., Shibata, N., Ito, M.: Energy-aware video streaming with QoS control for portable computing devices. In: ACM NOSSDAV. (2004) 68–73
3. Lee, C., Lehoczky, J., Rajkumar, R., Siewiorek, D.: On quality of service optimization with discrete QoS options. In: IEEE RTAS. (June 1999) 276–286
4. Feeney, L.M., Nilsson, M.: Investigating the energy consumption of a wireless network interface in an ad hoc networking environment. In: IEEE INFOCOM. (April 2001) 1548–1557
5. Krantz, S., Kress, S., Kress, R.: Jenen's Inequality. Birkhauser (1999)
6. Yuan, W., Nahrstedt, K.: Energy-efficient soft real-time CPU scheduling for mobile multmedia systems. In: ACM SOSP. (August 2003) 149–163
7. ESPN: MLB scoreboard. http://sports-ak.espn.go.com/mlb/scoreboard

SCA-Based Reconfigurable Access Terminal

Junsik Kim[1], Sangchul Oh[1], Eunseon Cho[1], Namhoon Park[1], and Nam Kim[2]

[1] Mobile Telecommunication Research Division
Electronics and Telecommunications Research Institute, Daejeon, Korea
{junsik,scoh,escho,nhpark}@etri.re.kr
[2] Dept. of Comput. & Commun. Eng., Chung-Buk Nat. Univ., Cheongju, Korea
namkim@chungbuk.ac.kr

Abstract. In this paper, we propose a Reconfigurable Access Terminal (RAT) which is composed of SDR(Software Defined Radio) hardware platform test bed and SCA(Software Communication Architecture) - based software platform. Specifically, we propose a design of the SDR Access Terminal middleware, wireless protocol software component and the procedure of reconfiguration which changes one mode into the other mode. Our RAT system shows that it is capable of changing mode between WiMAX(World Interoperability for Microwave Access) and HSDPA(High Speed Downlink Packet Access). The radio access protocol and application components of RAT are designed through a SCA adapter.

Keywords: SCA; SDR, Reconfiguration, Software Component, Middleware.

1 Introduction

The SDR, which is a communications device whose functionality is defined in software, has made itself become a key enabling technology in order to realize such a flexible and reconfigurable radio system. Several research works have involved in the development of SDR systems and their efforts focus on hardware design and software framework. However, the current SDR system is insufficient to provide their full potential due to the performance problems. The SCA specification by JTRS(Joint Tactical Radio System) establishes a hardware-independent development framework with baseline requirements for the software definable radios [1]. The SCA has been published to provide a common open architecture that can be used to build a family of radios across multiple domains. The SCA also supports software reusability. Meanwhile, the SCA aims to define a middleware that allows baseband, modulation, protocols modules working together[2][3].

The remainder of this paper is as follows. We provide an overview of the SCA which is a software platform for the RAT and provides a flexible and reconfigurable environment in Section 2. In Section 3, we present the RAT system architecture and functionalities which includes hardware platform and software platform architecture. Then, we describe the procedure of reconfiguration function. Finally, conclusions are drawn.

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 713–716, 2007.
© Springer-Verlag Berlin Heidelberg 2007

2 SCA Overview

The SCA defines an Operating Environment (OE) and specifies the services and interfaces that the applications use from the environment. The interfaces are defined by using CORBA IDL and graphical representations are made by using Unified Modeling Language (UML). The OE consists of a Core Framework (CF) that is the essential set for the open interface, a CORBA middleware and a POSIX-based OS. The CF describes the interfaces, their purposes and operations[1]. It provides an abstraction of the underlying software and hardware layers for software application developers. The SCA compatible systems must implement these interfaces. The application components of the SCA divide into two parts, CORBA and Non-CORBA components. The communication between CORBA components and Non-CORBA components is possible through a SCA adapter[4].

3 Reconfigurable Access Terminal

We describe our sample RAT platform architecture based on SCA. Fig. 1 describes the reference implementation system architecture for demonstrating SCA-based RAT software download function between WiMAX and HSDPA. On the hardware platform, it is organized by the air interface waveform software components from the software download center using over the air download scheme[2].

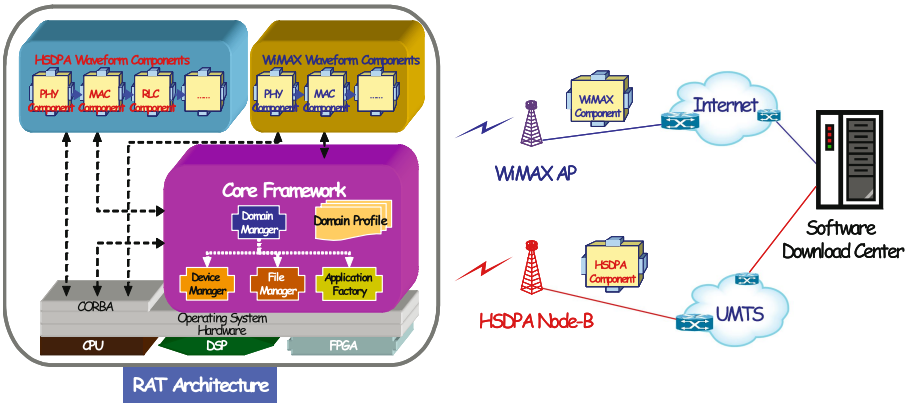


Fig. 1. Reconfigurable Access Terminal

3.1 RAT Hardware Platform Architecture

The functional modules on hardware platform are composed of mother board, processor board with GPP(General Purpose Processor), FPGA for base band modem, IF board, RF module and reserved DSP board. As user interface ports, Ether-net port, USE2.0 port, UART port and JTAG for FPGA downloading/debugging are provided. As main control processor, PowerPC embedded in XC2VP30 operates at 150 MHz (Upto 300MHz is available in Xilinx data sheet). 128 Mbyte FLASH is used for

program ROM, 128 Mbyte SDRAM is used for program RAM. Two 1Mbit DPRAM (CY7C028) are provided for reserved data memory for MAC hardware. The both of RF and HSDPA modem is under construction, therefore, at the HSDPA mode, the lower transport layer of MAC connected to LAN through a modem simulator and tested.

3.2. RAT Software Platform Architecture

The RTOS, CORBA, and SCA CF reside on the main board memory and application components are downloaded from the download center or other auxiliary memory. The radio protocol and application components of the RAT system divide into two parts, CORBA and Non-CORBA components. The communication between CORBA components and Non-CORBA components is possible through the SCA adapters.

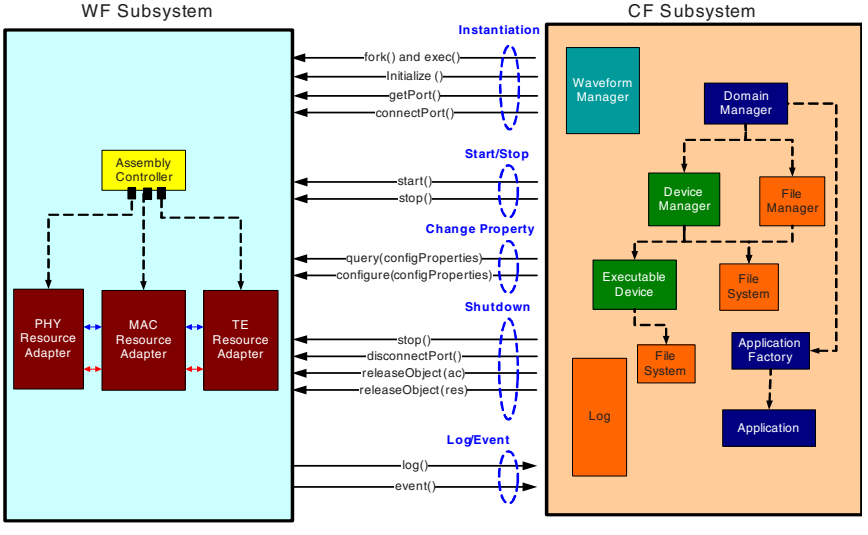


Fig. 2. RAT middleware operation on the Software platform

Fig. 2 shows the middleware operation on the software platform and interfaces between the subsystems of our RAT system. The software platform is specified by two main subsystems which are CF (Core Framework) subsystem and WF (Waveform) subsystem. The CF subsystem, provides the SCA software environment that explained before, makes the efficient operation each manager possible as the middleware which comprises the waveform which is the application managing the wireless access and which it runs. The WF subsystem is comprised of the wireless access protocol and application components, and it is applied in the position of a middleware with the object of a drive. The CF subsystem installs and comprises the wireless access software components on the WF subsystem through a command including the start / stop, the change property for the parameter delivery, a shutdown, the history log / event for message passing, and etc. and runs.

4 Reconfigure Procedure

In previous section, we presented the hardware and software platform of the RAT system. Each platform has the proper structure to reconfigure the system. In this section, we describe the mode change function as the core function of the SDR. The mode change function is initiated from the RATM which is software block, and responsible for the management of the RAT. As mentioned above, the RAT can provide the service of two modes i.e. WiMAX and HSDPA independently. When the RAT is currently on service with the WiMAX mode, the user (or operator) via the RATM wants to change the HSDPA mode. When the RATM decides to change the mode, it sends the change mode request with a HSDPA software package to the MC which is software block and responsible for the mode control of the RAT, and it unpacks and installs the HSDPA software package into the specified device modules. The MC sends the change mode request to the device modules. Once receiving the change mode request, the device modules let the WiMAX software modules stop and shutdown. If the change mode procedure of the WiMAX software modules is complete, the MC sends the reconfigure request to the component update unit. After the update completion, the MC sends the execution request to the device module. Once receiving the execution request, the device modules let the software modules of the HSDPA run. If the entire change mode procedure is complete, the MC responses the RATM.

5 Conclusion

This paper represents a Reconfigurable Access Terminal developed on the basis of the SCA standard architecture and it is capable of providing the WiMAX and the HSDPA service. The RAT currently adopted an interim architecture using a SCA adaptor. However, if there are more advances in the SCA technology as well as the SDR hardware, the architecture of the RAT will be upgraded for the system performance. Specifically, we designed SCA-based RAT software platform and illustrated the procedure of reconfiguration which changes one mode into the other mode. The research of reducing the reconfiguration time, FPGA componentization, and software modem improvement should be studied later.

References

1. JTRS website, http://www.jtrs.army.mil
2. SDR Forum website, http://www.sdrforum.org
3. Saehwa Kim, Jamison Masse, Seongsoo Hong, and Naehyuck Chang : SCA-based Component Framework for Software Defined Radio, Proceedings of the IEEE Workshop on Software Technologies for Future Embedded Systems 2003, May 15-16, (2003) 3-6
4. EunSeon Cho, ChangKi Kim, YeonSeung Shin, and JinUp Kim : SCA-based multi-LAN application development, Proceedings of VTC2004-Fall, 26-29 Sept. (2004) 1978 - 1982

Investigating Media Streaming in Multipath Multihop Wireless Network

Binod Vaidya[1], SangDuck Lee[2], Eung-Kon Kim[3], JongAn Park[2], and SeungJo Han[2,⋆]

[1] Dept. of Electronics & Computer Eng., Tribhuvan Univ., Nepal
vaidyab@aol.com
[2] Dept. of Information & Communication Eng., Chosun Univ., Korea
dandylsd@hanmail.net, japark@chosun.ac.kr, sjbhan@chosun.ac.kr
[3] Dept. of Computer Science, Sunchon National Univ., Korea
kek@sunchon.ac.kr

Abstract. Mobile Ad hoc Networks (MANETs) are very attractive for many applications. However, media streaming over MANET is quite challenging task. In this paper, we depict a framework for audio streaming over multihop wireless network. And we propose multipath routing for MANET and investigated media streaming using different scalable speech coding techniques. With the simulation results, performance of such a framework is evaluated.[1]

1 Introduction

As mobile ad hoc networks (MANETs) are self-organizing, rapidly deployable, and have no centralized control and administration, they are very attractive for many applications, such as battlefield communication, personal area networking and search-and-rescue. However, media streaming over MANET is quite challenging because of dynamic network topology, limited wireless bandwidth, and high bit error rate of wireless links.

When real-time multimedia is streamed over MANET, packet loss rate can be very high under adverse conditions, thus in turn, communication may be lost. In this paper, we propose a multipath routing protocol for multihop wireless network that can be used efficiently for media streaming. Till now, many multipath routing protocols have been proposed for wireless ad hoc networks. Some of multipath protocols based on AODV [1] are AOMDV [2], NDMR [3], and AODV-BR [4]. While streaming multimedia through multiple paths, the content can be divided into multiple minor flows and stream through available paths as in MDSR [5]. A scalable speech coding technique [6] is considered as it is designed for adaptable real-time traffics over lossy networks. Main intention of our investigation is to show the effect of different scalable speech coding techniques on MANET using proposed multipath routing protocol.

⋆ Corresponding author.

[1] This study was supported (in part) by research funds from Chosun University 2006.

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 717–720, 2007.
© Springer-Verlag Berlin Heidelberg 2007

2 Multipath Routing for Media Streaming over MANET

Proposed multipath routing protocol for MANET is basically intended for highly dynamic ad hoc networks in which communication faults frequently occur. It is basically modification of a single-path on-demand AODV routing protocol. As in AODV, three control messages are route request (RREQ), route reply (RREP) and route error (RERR). The proposed protocol has two basic phases, namely route discovery and route maintenance.

First, to find routes for a destination node D, a source node S broadcasts RREQ packet. RREQ packet structure of the proposed scheme is same as that of AODV, except presence of path accumulation list of the route path. The source node appends own address to route path in RREQ. Further, when RREQ is forward by intermediate nodes, each node appends its address to it. As RREQ ID and source address form a unique identifier for RREQ, a node checks a received RREQ if it is from same source and with same RREQ ID. If a node receives first RREQ packet, it records a reverse route in its routing table. An intermediate node receiving RREQ, replies by sending a RREP if it has a route to the destination. In our proposed scheme, intermediate nodes forward duplicate RREQs that came from at most two different neighbors. This is essential to discover a number of alternate route paths.

In this scheme, the destination is responsible for selecting multiple alternate route paths. When receiving first RREQ, the destination records route paths of RREQ. Then after copying route paths of RREQ to a RREP packet, the destination node sends it to source node via its route paths. When the destination receives a duplicate RREQ, it will compare route paths of RREQ to that of the routing table. If only source node and destination node are same between them, a path is node-disjoint with primary path. If at least one of intermediate nodes in route paths in the routing table is different from all of nodes in route paths of RREQ, a route is partially disjoint path, which is defined as fail-safe [7]. Similarly, the destination sends RREP to the source along route path of RREP.

In case of route maintenance, the proposed scheme is capable of recovering broken routes immediately. When a node fails to deliver data packets to the next hop, it removes entries with broken link in its routing table and if it has another entry for the destination, data packets is delivered through the alternate route. If it has no another entry, it sends a RERR packet to the upstream node. When the source has no entry for the destination, it would initiate a new route discovery.

Scalable speech coding consists of a minimum rate bit stream that provides acceptable coded speech quality, along with one or more enhancement bit streams, which when combined with a lower rate coded bit stream, provide improved speech quality. The standards for scalable speech coding are G.727 [8], and MPEG-4 speech coding [9]. The G.727 speech coding is based upon adaptive differential pulse code modulation (ADPCM) with data rates of 16kbps to 40kbps [8]. The core bitrate is 16kbps, and up to three 8kbps enhancement layers can be included. MPEG-4 Natural Speech Coding Tool Set [9] provides a generic speech coding framework having bitrate from 2kbps to 23.4kbps. MPEG-4 speech

coding scheme uses CELP (Code excited linear predictive coding) for bitrates higher than 3.85kbps. The bitrate scalability of the core layer bitrate is possible up to three enhancement layers.

In this framework, we have evaluated G.727 scheme with 16kbps core layer and one 8kbps enhancement layer, for a total of 24kbps and MPEG-4 CELP scheme with a 6kbps core layer and one 2kbps enhancement layer, for a total of 8kbps. Since we use scalable speech coding for media streaming, the source node begins to send core bit stream on primary path and enhancement bit stream on node-disjoint path. Since primary path and node-disjoint path are not correlated, source node uses node-disjoint path to provide load balancing. When forwarding paths break, nodes receiving core bit stream or enhancement bit stream may use different paths in the routing table to forward packets. Generally, a multihop wireless path is up or down for random periods of time, leading to bursty packet losses. A core layer packet loss is likely to be experiencing packet loss burst. So the proposed multipath protocol finds an alternate fail-safe path for each node on primary path as it has higher packet delivery rate.

3 Performance Evaluations

In order to evaluate performance of the proposed framework for media streaming over multipath wireless ad hoc network, we have designed experimental model and simulated using OPNET Modeler [10]. In the simulation, MANET consists of sixteen mobile nodes which are randomly placed inside a 600m x 600m region initially. We consider that a mobile node moves around continuously with using random waypoint mobility model with pause time of 0s and a maximum speed of 5m/s. The channel has a bandwidth of 1Mbps. Transmission range is 250m. Among these nodes, one is randomly chosen as a streaming source and another as the destination. Five UDP traffic flows are introduced as background traffics. Each of these flows has traffic rate of four packets per second. Size of data payload is 512 bytes. Each of nodes has queue size of 10 packets.

For the experimental purpose, two scenarios have been considered, that is, a framework for audio streaming over MANET using G.727 scheme whereas another using MPEG-4 speech coding technique; in both cases, the proposed multipath routing scheme is used. Performance metrics computed during simulations are packet loss rate and end-to-end packet delay.

In order to analyze the simulation results for above framework, we compare performance of two different scalable speech coding techniques in terms of packet loss rate with respect to bit error rate (BER) and end-to-end delay (latency). Fig. 1 shows the packet loss rate for the both scenarios with respect to BER. It can be seen that the packet loss rate in G.727 coding scheme increases rapidly than in case of MPEG-4 CELP scheme. At BER of 10^{-3}, packet loss rate for G.727 coding scheme is about 10% whereas for MPEG-4 CELP scheme it is about 8.1%. Fig. 2 illustrates the end-to-end delay for scalable speech coding techniques. It can be seen that end-to-end delay for G.727 coding scheme has higher average end-to-end delay than that of MPEG-4 CELP scheme. It can

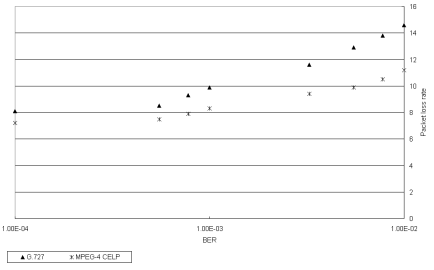


Fig. 1. Packet loss rate

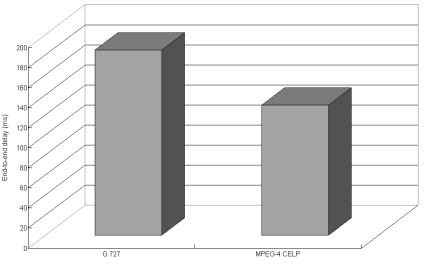


Fig. 2. End-to-end Delay

be derived that for the audio streaming over MANET, MPEG-4 CELP shows better performance than with G.727 coding scheme.

4 Conclusions and Future Work

In this paper we have depicted a framework for media streaming over multipath wireless ad hoc network. In order to investigate the performance of G.727 and MPEG-4 CELP speech coding schemes, we have simulated above framework using proposed multipath routing in MANET. It can be seen that the performance of MPEG-4 CELP scheme is much better that G.727 in terms of packet loss rate and end-to-end delay. In the future work, we will investigate above framework while applying selective encryption in scalable speech coding technique.

References

1. C. E Perkins, E. Royer: Ad-hoc on-demand distance vector routing, IEEE WMCSA 1999, Feb. 1999, pp 90-100
2. M.K Marina, S.R Das: Ad hoc on-demand multipath distance vector routing, Wiley Wireless Communications and Mobile Computing, Vol 6(7), 2006, pp. 969-988
3. X. Li, L. Cuthbert: Stable node-disjoint multipath routing with low overhead in mobile ad hoc networks, IEEE MASCOTS 2004, Oct 2004, pp. 184-191
4. S.J Lee, M. Gerla: AODV-BR: Backup routing in ad hoc networks, IEEE WCNC 2000, Vol 3, Sep 2000, pp. 1311-1316
5. S. Mao, *et al.*: Video transport over ad hoc networks: multistream coding with multipath transport, IEEE Journal on Selected Areas in Communications, Vol 21(10), Dec. 2003, pp. 1721-1737
6. H. Dong, *et al.*: SNR and bandwidth scalable speech coding, IEEE ISCAS 2002. Vol 2, pp. 859-862
7. L. R Reddy, S.V Raghavan: SMORT: Scalable multipath on-demand routing for mobile ad hoc networks, Elsevier Ad hoc Networks, Vol 5(2), 2007, pp. 162-188
8. ITU-T, 5-, 4-, 3-, and 2-bit/sample embedded adaptive differential pulse code modulation (ADPCM), Dec. 1990
9. ISO/IEC JTC1 SC29/WG11, ISO/IEC FDIS 14496-3, "Coding of Audio-Visual Objects Part 3: Audio", Oct. 1998
10. OPNET Modeler Simulation Software, http://www.opnet.com

A Low-Power 512-Bit EEPROM Design for UHF RFID Tag Chips

Jae-Hyung Lee[1], Gyu-Ho Lim[1], Ji-Hong Kim[1], Mu-Hun Park[1], Kyo-Hong Jin[1], Jeong-won Cha[1], Pan-Bong Ha[1], Yung-Jin Gang[2], and Young-Hee Kim[1]

[1] Changwon National University,
9 Sarim-dong, Changwon, Gyeongnam, 641-773, Korea
{tommo,ghlim}@changwon.ac.kr, lapael81@empal.com,
mhpark@sarim.changwon.ac.kr
{khjin,jcha,pha,youngkim}@changwon.ac.kr
[2] DavitDyne Co., Ltd.
B 901-3, Ssangyoung IT Twin Tower, Sangdaewon-dong,
Sungnam, Kyungki, 462-723, Korea
jim@davitdyne.com

Abstract. In this paper, a design for a low-power 512-bit synchronous EEPROM with flash cells for passive UHF RFID tag chip is presented. Applied are low-power schemes such as dual power supply voltage(VDD=1.5V and VDDP=2.5V), clocked inverter sensing, voltage-up converter, IO interface, and Dickson charge pump using schottky diode. An EEPROM is fabricated with the 0.25 EEPROM process. Simulation results show that power dissipations are 8.34 in the read cycle and 57.7 in the write cycle, respectively. The layout size is 449.3 × 480.67 .

Keywords: Low-Power, EEPROM, UHF RFID, Tag, Charge pump.

1 Introduction

RFID(Radio Frequency IDentification) is the technology to provide various services communicating between things by collecting, storing, and revising the information around these things through they installed RFID tags on item. RFID tags are classified according to communication method, battery existence, and read/write[1]. They are standardized by EPC global which sets EPC (Electronic Product Code). Currently, Generation 2 of Class 1 is one of the widely selected standards. It has advantages in terms of cost and area. This paper presented EEPROM design for a passive UHF RFID tag chip.

2 Circuit Design

Fig. 1 is a block diagram of a 512-bit synchronous EEPROM. The memory cell array is composed of flash cell[8]. The EEPROM has four operating modes : program, erase, read, and stand-by. It is synchronously operated by clock. Write mode means

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 721–724, 2007.
© Springer-Verlag Berlin Heidelberg 2007

program and erase modes. Dual power supply voltage, VDD(1.5V) and VDDP(2.5V), is used to reduce currents in read and write modes.

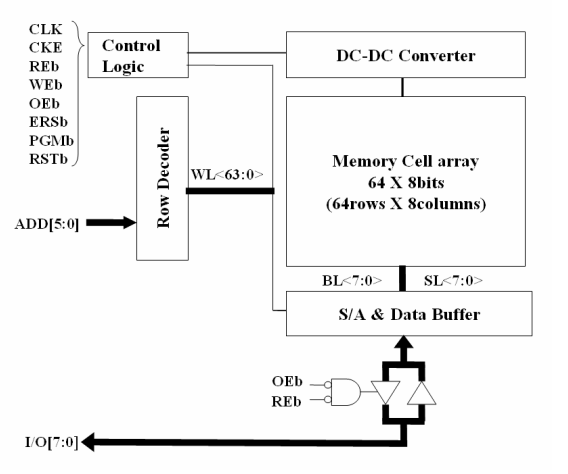


Fig. 1. Block diagram of 512-bit synchronous EEPROM

A clocked inverter sensing method[6] is applied to read data of EEPROM cell in the read mode. And a current sensing circuit is used for the non-volatile memory[4]. This is not proper in the EEPROM design for RFID tag chips because the current dissipation of the sensing circuit is large. Therefore, despite low speed, low-power RD(Read Data) sense amplifier is used without a reference current biasing circuit.

DC-DC converter uses Dickson charge pump[5] to generate high voltage in the write mode. VDDP power is used in the voltage-up converter, VPP control logic, and charge pump circuits[8]. VREF_VPP is reference voltage required for the DC-DC converter. VREF_VPP level is too high to generate it by using VDD, so a low-power voltage-up converter is added to the DC-DC converter. The voltage-up converter makes reference voltage double by using VDDP.

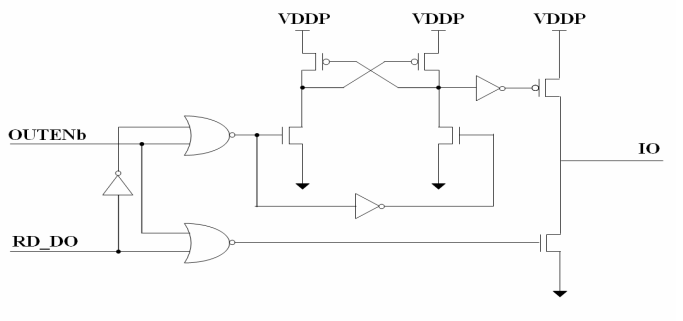


Fig. 2. IO Interface circuit

The voltage of RD_DO swings between VDD and VSS. The voltage of IO swings between VDDP and VSS. If VDD voltage transfers to the IO, they induce short circuit currents in other IO interface. So, level translator is newly applied to the IO interface.

Dickson charge pump generates high voltages, VPP and VPPL, in the write mode. The lower the forward bias voltage of the diode is used, the lower current flows in Dickson charge pump. For this reason, a schottky diode is used for the pump. Power dissipations of the charge pump are 67.7 with PN diode and 57.7 with schottky diode in the program mode. Approximately, the power dissipation with schottky diode drops by 12% compared with PN diode.

3 Simulation Results

Fig. 3 shows timing diagrams of CLK signal from analog circuit, command control signals, CKE, REb, OEb from logic circuit, and PRECHARGE, DLINE_LOADb and SAENb from control logic circuit as shown in Fig. 1. When read command enters at a rising edge, PCHARGE makes DLINE and BL precharge to VDD. WL is active after BL is precharged. When a data is transferred to the BL, a valid data comes out of I/O through the RD_DO within a half clock period with SAENb activated. Power dissipation simulation results are 57.7 in the program mode, 42.3 in the erase mode, and 8.34 in the read mode, respectively. Fig. 4 is a layout picture with 0.25 EEPROM process. Layout size is 449.3 × 480.67 .

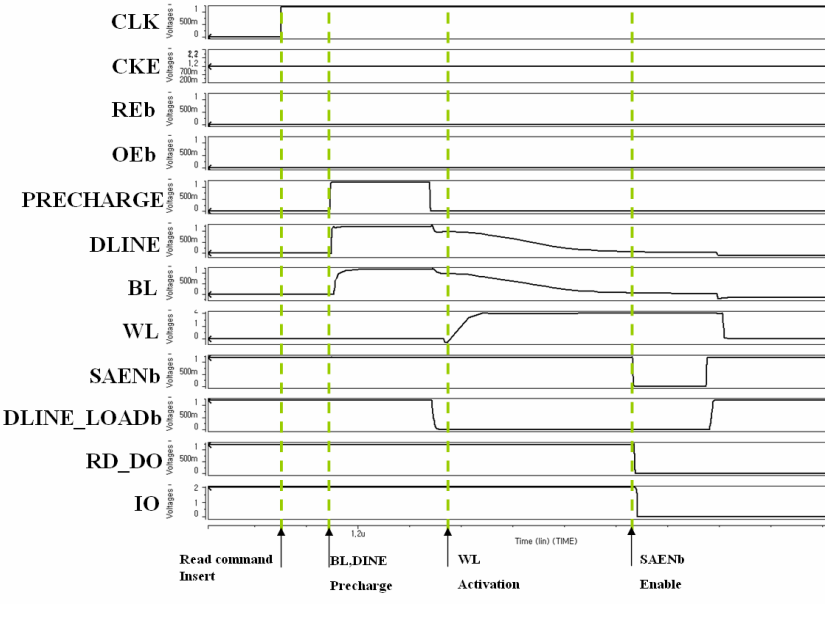


Fig. 3. Simulation result for the case of critical path in the read cycle

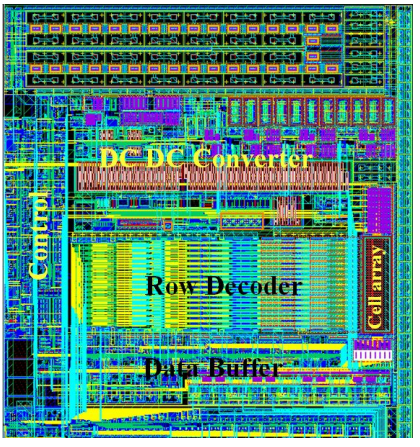


Fig. 4. EEPROM layout picture

4 Conclusions

The EEPROM is fabricated with the 0.25□ EEPROM process. In this paper, to reduce power dissipation in EEPROM, dual power supply voltage, VDD(1.5V) and VDDP(2.5V), is used to reduce the currents in the read and write modes. Also sensing method using clocked inverter in the read mode is applied. VREF_VPP is made by using the voltage-up converter in the write cycle. Level translator is newly applied to IO interface in order to reduce short circuit current. A schottky diode is used for lower power dissipation in Dickson charge pump. Simulation result shows that the designed EEPROM is suitable for UHF RFID Class 1 Generation 2 and the fabricated EEPROM will be verified by measurement in near future.

Acknowledgments. This work is supported by IT R&D Project funded by Korean Ministry of Information and Communications**.**

References

1. http://www.epcglobalinc.org
2. Weinstein, R.: RFID: a technical overview and its application to the enterprise. IT Professional, vol.7, Issue 3. (2005) 27-33
3. Junghwan Lee and Minkyung Ko: A novel EEPROM cell for smart card application. Microelectronic Engineering, vol.71, Issues 3-4. (2004) 283-287
4. Fei Xu, Xiangqing He, and Li Zhang: Key Design Techniques of a 40ns 16K Bits Embedded EEPROM Memory. ICCCAS 2004, vol. 2. (2004) 1516-1520
5. J.F.Dickson: On-Chip High-Voltage Generation in MNOS Integrated Circuits Using an Improved Voltage Multiplier Technique. IEEE JSSC, vol. 11. (1976) 374-378
6. YoungHee Kim et al.:A low-power EEPROM design for UHF RFID tag chip. Journal of KIMICS, vol.10, No.3 (2006) 486-495

VPDP: A Service Discovery Protocol for Ubiquitous Computing

Zhaomin Xu, Ming Cai, and Jinxiang Dong

Institute of Artificial Intelligence, Zhejiang University, 310027 Hangzhou, China
{xzm,cm,djx}@zju.edu.cn

Abstract. Many service discovery protocols have been proposed by far to support ubiquitous computing. But they don't always apply to ubiquitous computing because of their confines. In this paper, we propose a service discovery protocol named VPDP adopted in our DOM-based middleware architecture, which aims to support as many types of ubiquitous environments as possible. VPDP is important for our middleware architecture to accommodate heterogeneity and uncertainty of ubiquitous computing environments.

Keywords: Ubiquitous computing, Service discovery, Volunteer, Middleware.

1 Introduction

The essence of the vision about ubiquitous computing was the creation of environments saturated with computing and communication capability, yet gracefully integrated with human users [1]. The technological advances necessary to build a ubiquitous computing environment fall into four broad areas: devices, networking, middleware, and applications [2].

Service discovery is essential for ubiquitous computing environments to gracefully integrate networked computing devices. Service discovery protocols are designed to minimize administrative overhead and increase usability. They can also save ubiquitous system designers from trying to foresee and code all possible interactions and states among devices and programs at design time. [3]. In this paper, we propose a service discovery protocol called VPDP (Ubiquitous Discovery Protocol based on Volunteers) for ubiquitous computing, which aims to support as many types of ubiquitous environments as possible. This protocol can accommodate heterogeneity and uncertainty of ubiquitous computing environments. Volunteers in our middleware architecture are middleware nodes with less limited resources.

The rest of this paper is structured as follows: Section 2 discusses our service discovery protocol including volunteers election, service registration and service discovery. Section 3 shows the experimental result of service discovery using VPDP. Finally, we make a conclusion of this paper and describe our future work.

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 725–728, 2007.
© Springer-Verlag Berlin Heidelberg 2007

2 Ubiquitous Discovery Protocol Based on Volunteers

2.1 Volunteers Election

Volunteers in VSD [5] are elected within a one-hop network range using broadcasts. In our middleware architecture, volunteers are elected from middleware nodes within the range of a certain number of hops using multicasts. The number of hops (denoted as 'TTLs') is defined as a system parameter which can be changed by system administrators or by the middleware system automatically. The election method of volunteers in our middleware architecture is similar to VSD, but we have made a few modifications to its parameters. Fig. 1 shows the node state transition diagram of VPDP.

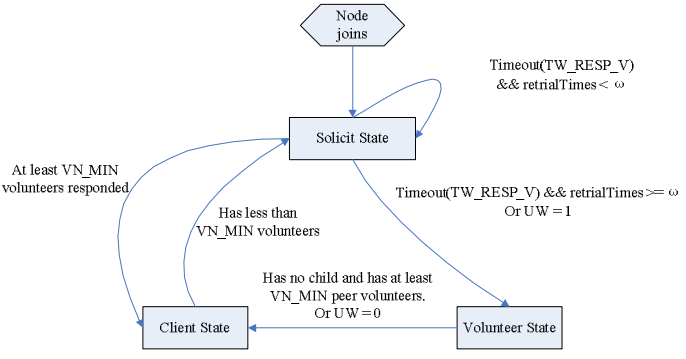


Fig.1. Node state transition diagram

Unlike VSD, in our middleware architecture, when a node repeats the solicit process after it can't register with k volunteers within a given time period, it will try to register with max {(k / 2), k_{min}} volunteers. Although at the start all nodes try to register with k volunteers, they may end up with different number of local volunteers that they have registered with. In other words, each client may have different number of local volunteers and each volunteer may maintain different number of the same directory entries for different clients.

In VSD, each node sets its own retrial times (denoted as 'ω') by considering its willingness, degree of mobility and amount of resource [5]. The lower value of ω the higher chance a node can take to be a volunteer. In our middleware architecture, each node turns on or restarts with the system parameter ω set to a default value, which will be changed by a evaluation function over time.

The evaluation function of ω depends on three parameters: the total running time in hours (TRTh) of the node since it turns on or restarts, changed times of the network address (CTNA) of the node, and the number of services (n_s) on the node. The evaluation function is invoked periodically or when some events happen (such as network address changed, new services mounted on the node). TRTh and CTNA indicate the degree of mobility of a node. The number of services indicates the

amount of resources on the node. The willingness of a node is hard to evaluate. We use the system parameter user willingness (UW, a value between 0 and 1, high value indicates high willingness) and n_s to evaluate a node's willingness. Currently, the evaluation function used in our middleware architecture is defined as below.

$$\omega = \frac{1}{UW \times (n_s + 1)} \times \frac{CTNA + 1}{TRTh / 7} \times \frac{6}{n_s + 1} \tag{1}$$

2.2 Service Registration and Discovery

When a SP registers its services, it sends service advertisements to its local volunteers one by one through unicasts. A SP will register its services with at least k_{min} volunteers, which form a logical overlay network. Volunteers extract service information from service advertisements received and store services' information in their service cache. Volunteers' announcements contain advertisements about their cached services, so that service advertisements can spread over the network.

A service requestor (SR) can send its service request messages if it has at least k_{min} local volunteers. There are two types of queries as stated in paper [4] in our middleware architecture: one query-one response (1/1) and one query-multiple responses (1/n). In 1/n queries, the service discovery process is almost the same as that described in paper [5].

In 1/1 queries, the SR selects a volunteer in its local volunteers list and sends a service request to it. If the SR gets a service response, it can then directly interact with the SP. Otherwise if it can't get any response within a certain amount of time, it will send the service request to the next volunteer in its local volunteers list. If all the volunteers have been tried and the SR still can't get any service response, the discovery process fails. On receiving a service request, a volunteer will lookup services in its service cache. The volunteer will send the service request to the first matched SP. If the SP accepts the request, it will send an acknowledgement to both the SP and the volunteer. The service discovery process ends successfully. Else if the SP doesn't respond within a certain amount of time, the volunteer will try to find another matched SP in its service directory. If the volunteer can't find any matched SP, it will forward the service request to its nearby volunteers. The discovery process will continue in this way until a matched SP is found or TTLr <= 0 (the time to live field of the request message) or all neighbor volunteers have been tried.

3 Experiment

We have developed a prototype of our DOM-based middleware architecture using Eclipse 3 and J2SDK 1.4.2. The experiment is taken in our campus network which is composed of many Local Area Networks. We have tested the mean response time of a service request within different number of nodes. Fig. 2 shows the results which indicate that VPDP is very efficient in service discovery.

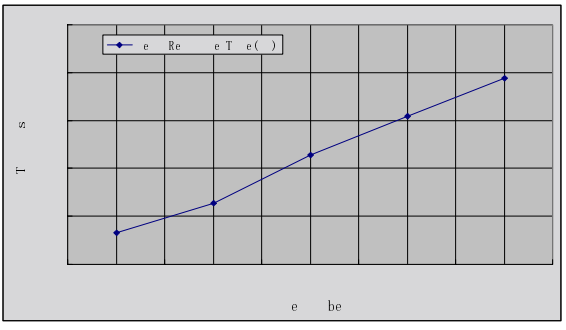


Fig. 2. Results of service discovery using VPDP

4 Conclusions and Future Work

Service discovery is essential for ubiquitous computing environments to gracefully integrate networked computing devices. We have discussed the VPDP protocol adopted in our middleware architecture, which aims to support as many types of ubiquitous environments as possible. Experimental results show that VPDP is efficient in service discovery, which is important for our middleware architecture to accommodate heterogeneity and uncertainty of ubiquitous computing environments. In the future, we will keep on improving VPDP to realize efficient service migration.

References

1. M., Satyanarayanan: Pervasive Computing: Vision and Challenges. IEEE Personal Communications, 8(4) (2001) 10-17
2. D., Saha, A., Mukherjee: Pervasive computing: a paradigm for the 21st century. Computer, 36(3) (2003) 25-31
3. Feng, Zhu, Matt, W., Mutka, Lionel, M., Ni: Service discovery in pervasive computing environments. IEEE Pervasive Computing, 4(4) (2005) 81-90
4. Celeste, Campo, Carlos, García-Rubio, Andrés, Marín López, Florina, Almenárez: PDP: A lightweight discovery protocol for local-scope interactions in wireless ad hoc networks. Computer Networks, 50(17) (2006) 3264-3283
5. M.J., Kim, M., Kumar, B.A., Shirazi: Service Discovery using Volunteer Nodes for Pervasive Environments. Proceedings of International Conference on Pervasive Services, (2005) 188-197

A Study on the Aspects of Successful Business Intelligence System Development

Il Seok Ko and Sarvar R. Abdullaev

Division of Computer and Multimedia, Dongguk University, 707 Seokjang-dong, Gyeongju-si, Gyeongsangbuk-do, 780-714 Korea
isko@dongguk.edu

Abstract. Business Intelligence (BI) Systems are today considered as a major strategic tool of many well-established companies like Lufthansa, TDC Telecom and AT&T. Of course, the knowledge derived from the experience of those companies could be helpful guide in building efficient BI Systems. Thus, this study highlights the main points arisen from the worldwide practice of building successful BI Systems. The managers usually expect from BI Systems the acceleration of their decision-making process while keeping the quality of each decision, the enhancement of their product development cycle, the maximization of the profit from existing product lines and the discovery of new opportunities, the establishment of better marketing with robust CRM. However, more than 50% of BI projects fail to meet these requirements. So this study investigates the critical mistakes commonly made by BI System developers and suggests profound solutions from the best BI practices.

Keywords: business intelligence, failure of BI, successful BI cases.

1 Introduction

What made people build BI Systems? IBM gives the clear definition of BI: "*Business intelligence means using your data assets to make better business decisions. It is about access, analysis, and uncovering new opportunities.*" By the end of 90s, big corporations found out that they have huge amount of unused historical data accumulated from their daily activities. So the people started to mine that useless data and consequently gained useful knowledge.

Generally speaking, we can split the BI evolution into 3 stages. Initially, there have been business information systems aimed at accomplishing limited range of operational business activities and, in the same way, storing the operational data. Almost all decisions had to be made leaning on operational data and sometimes that made a clash. At the second stage, the historical data has been separated from operational data into data warehouses which are specially tailored to store and provide quick access to such data. It also summarized the data in various cuboids and consequently eased the decision making process. Finally, today's BI systems have been discovered by involving data mining techniques and artificial intelligence in order to extract knowledge for decision making.

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 729–732, 2007.
© Springer-Verlag Berlin Heidelberg 2007

In this paper we will consider both successful cases of BI system deployment and also study commonly made critical mistakes while building BI system.

2 Why Do BI Projects Fail?

More than half of BI projects fail to meet the requirements stated at the beginning of the project. Commonly these are the requirements aimed at facilitating the process of decision making and identifying new business opportunities by the utilization of accumulated historical data. However these goals are not always met, and there are several reasons for that.

Mostly BI projects considered to be the inner project which should meet the company's internal requirements. However it should reflect the needs of customers and deal with external issues like market situation or customer's behavior. On the other hand, there is always gap between the actual implementers and the users of the system, and sometimes it becomes technology oriented project which aimed at outrunning in technological perspective. Nonetheless, it is a false ideology. BI should be done toward business perspective of the company. As the BI project is organization-wide project, it should involve members from each department especially from Marketing and Sales departments who are appeared to be the end-users of knowledge mined out from BI system.

BI projects are organization specific project, so they have to be unique implementation methodology where the project deliverables and goals should be stated clearly. BI development is an incremental process which never stops getting improved. Unlike OLTP systems, BI system never reaches the end, as it continues evolving and acquiring new functionalities. As BI used for achieving strategic goals, there is no definite set of deliverables which constitute the completeness of BI system and they always change. On the other hand, the project should be properly planned before kicking it off. It include stages like requirements gathering, identifying the condition of available data sources, cost estimation, risk assessment, identifying the success factors, preparation of project charter and issuance of high-level project plan.

Also, the standardization of each process and data makes the things much comprehensive and kills ambiguities. Therefore there should be a thorough standard which makes the teamwork coherent. Along with unique standard, there should be data quality control. As the quality of the knowledge directly connected to the quality of inputted data, the rules on cleansing the dirty data should be accepted. Additionally, there should be metadata repository which gives the context to stored data. We can divide metadata into two types: technical and business. Technical metadata defines the algorithms and the definition of database while business metadata gives unambiguous definition of business elements like sales, products and etc.

BI developers usually strive to implement everything, but in many cases it is hardly achieved and sometimes it causes many problems associated with the continuation of the project or maintaining the bundle of tools which are not rather used. There are three major applications should be included in any BI system:

1. ETL services
2. Data warehouse and integrated OLAP
3. Front-end application

It is worthwhile to say that if to take into account all issues that have been mentioned above, the BI project is likely to payoff and to reach its ultimate goals like increase in sales, efficient product development cycle, minimized expenses, better loyalty of customers and, of course, discovery of new business opportunities.

3 Successful BI Cases

In this section, we will discuss a couple of different BI cases implemented on various platforms.

3.1 Microsoft Case Study: TDC Telecom

TDC is the Denmark's telecommunication leader with annual revenue of $8 billion is spread over 12 countries. Recently it implemented SQL Server 2005 to integrate 6 terabytes of data from over 60 disparate sources into one data warehouse and to harvest much knowledge from that data. The company was in need of Business Intelligence system which would integrate the data from various sources and would possess following features:

- Integrate all data into a data warehouse through ETL to provide a single "version of the truth."
- Create multidimensional cubes to support data analysis.
- Reduce the cost of analyzing disparate data.

The BI solution provided by Microsoft Corp. with the use of technologies such as Microsoft SQL Server 2005 (64-bit) Enterprise Edition, Microsoft SQL Server Analysis, OLAP and Reporting Services on the Microsoft Windows Server 2003 (64-bit) Enterprise Edition had a new name called CUBUS. Firstly, CUBUS adopted single standard for interpreting the data and met the requirement of single "version of the truth" which was stated above. It integrated the available data into one source which, in turn, maintained the consistency of data. Also even the data is stored in one source the views for that data could differ upon the analyst's preference. The single standard also included the definition of each data construct stored in separate meta-data repository. That enabled business analysts to understand the context of each data and consequently make appropriate decisions. Similarly, with help of multidimensional cubes the terabytes of data has been efficiently stored and summarized which led to 80% reduction of processing time. Moreover the success of this BI project could be proven through high motivation of key business representatives in building such system.

3.2 Oracle Case Study: Etos

Etos is the major supermarkets chain scattered over Neitherland with its 450 outlets. Before it started to promote brands of other foreign stores, it had operated single database. But with the integration of different foreign stores, its overall IT infrastructure has become heterogeneous which consequently posed new challenges. Etos needed a centralized collection point from which to gather information on point-of-sale purchases, product range, pricing, and special offers.

The Oracle platform appeared to be the best choice as a ground for comprehensive business intelligence environment. Setting up a data warehouse with Oracle Warehouse Builder, Oracle Database, Oracle Portal, Oracle Reports and Oracle Discoverer, made it possible to access business data in any required combination. Now it has integrated data into one point and it could be viewed by means of graphs and tables. It makes possible to store historical data and facilitates the job of analysts while examining the precise trend.

The integration of information from procurement, logistics, and sales systems made it possible to monitor the actual picture of how every part of the retail business is performing. For example, BI discovered much knowledge about how particular products are selling well in a particular region and how their shelf-position affects sales.

4 Conclusion

To sum up, there are several challenging points in developing BI and they are often overlooked by BI developers which, in turn, could lead to failure of BI project. We can summarize these points one more:

- Not internal requirements, but market and customer requirements.
- Dedicated business representation from each department.
- Availability of skilled team members.
- Unique BI development methodology.
- Thorough project planning.
- Data standardization.
- Date quality control.
- Existence of metadata.
- Implementation of only required tools.

On the other hand, some BI projects survive and meet their final requirements. These BI projects usually accomplished by popular software vendors like Microsoft, Oracle, SAP and etc. But on the bottom line, the success of BI project depends on how BI team members cope with those challenging points mentioned above.

References

1. Shaku Atre, “The Top Ten Critical Challenges For Business Intelligence Success”, Computerworld Custom Publishing, (2003).
2. Larissa T. Moss, Shaku Atre, “Business Intelligence Roadmap – The Complete Project Lifecycle for Decision-Support Applications”, Addison Wesley, (2003)
3. Roman Bukary, “Top 10 BI Traps”, BI Strategic Directions, CIO, (2006)
4. Maria Sueili Almeida et al, “Getting Started with Data Warehouse and BI”, IBM, (1999)
5. Nils H. Rasmussen et al, “Financial Business Intelligence: Trends, Technology, Software Selection, and Implementation”, Wiley, (2002)
6. www.oracle.com, “Etos Adopts Business Intelligence Approach for Its 450 Outlets” (last accessed: 14.12.2006)
7. www.microsoft.com, “Denmark Telecom Leader Cuts with 6 Terabyte Data Warehouse on SQL Server 2005” (last accessed: 14.12.2006)

Robust Phase Tracking for High Capacity Wireless Multimedia Data Communication Networks

Taehyun Jeon

Seoul National University of Technology,
Dept. of Electrical Engineering,
172 Gongneung-2Dong, Nowon-Gu
139-743 Seoul, Korea
thjeon@snut.ac.kr

Abstract. This paper proposes an algorithm for tracking of the residual phase errors incurred by carrier and sampling frequency offset in the OFDM transmission systems which is suitable for high speed and high capacity wireless multimedia communication networks. The proposed scheme utilizes the degree of the channel fading in the frequency domain in the offset tracking procedure which improves the error estimation accuracy and the tracking performances. This scheme also contributes to the increase of the throughput and the network capacity by reduction of the packet error rate or equivalently probability of re-transmission events which are critical QoS requirements for multimedia data transmissions.

Keywords: ubiquitous network, QoS, carrier frequency offset, OFDM.

1 Introduction

The demands for access to the high quality multimedia data any time any where have been increased dramatically with the introduction of the concept of ubiquitous networks. The requirements for such a high capacity data transmission include enhanced data retrieval under various channel and interference environments as well as the data rate increase up to hundreds or even to gigabits per second ranges depending on the application areas. The orthogonal frequency division multiplexing (OFDM) system achieves high data rates by simultaneous transmission of multiple data symbols through subcarriers which are orthogonal to each others [1]. As in the general synchronous digital communication systems, the carrier frequency difference between transmitter and receiver sides in the OFDM plays an important role in the entire system performance. Another important impairment is the sampling frequency offset which is generated in the process of digital-to-analog (D/A) and analog-to-digital (A/D) conversion in both sides, respectively [2].

This paper proposes a reliable estimation and tracking method for phase errors based on the channel information in the OFDM systems. In this scheme the received signal processed by the frequency domain equalizer is used for the error estimation and the channel gain for each subcarrier is utilized to provide the improved reliability.

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 733–736, 2007.
© Springer-Verlag Berlin Heidelberg 2007

2 Phase Error Detectors

In this section the detection methods for phase errors are described which are caused by the local oscillators' carrier frequency difference and sampling clock frequency difference at transmitter and receiver sides. For better understanding of the proposed algorithm one of the widely used OFDM based local area wireless data transmission systems is assumed [3].

2.1 Phase Errors by Carrier Frequency Offset

The phase error caused by carrier frequency offset (CFO) can be approximated as the same for all subcarriers in the frequency domain within one OFDM symbols. Here the CFO phase error can be estimated as the average value over all corresponding subcarriers within one symbol period. In case only the pilot subcarriers are utilized for the error estimation the estimated phase error by CFO is represented as follows:

$$\hat{\theta}_i = \frac{1}{N_p} \sum_{j \in \{pilot_index\}} \phi_{i,j} \tag{1}$$

where $pilot_index$ represents a set of subcarrier indices where the pilots are located and N_p the total number of pilot subcarriers within one OFDM symbol. Also $\phi_{i,j}$ is the estimated phase error values contained in the received signal sample where i and j represent indices for OFDM symbol and subcarrier, respectively.

When payload signals are also used for the phase error estimation the error detector uses slightly modified procedure considering the locations and the involved number of data subcarriers. The combined form of phase error estimation can be represented as follows when both the pilot and the payload data with channel quality information are utilized:

$$\hat{\theta}_i = \frac{\sum_{j \in \{pilot_index\}} \left|\hat{H}_j\right|^2 \phi_{i,j} + \sum_{j \in \{data_index\}} \left|\hat{H}_j\right|^2 \phi'_{i,j}}{N_p \sum_{j \in \{pilot_index\}} \left|\hat{H}_j\right|^2 + N_d \sum_{j \in \{data_index\}} \left|\hat{H}_j\right|^2} \tag{2}$$

where the $data_index$ represents a set of indices where payload data are located. Also N_d is the total number of subcarriers associated with payload data and $\phi'_{i,j}$ is the estimated phase error based on data subcarriers. The estimated channel gain in j^{th} subcarrier is represented as $\hat{H}_j$ which is obtained using the preamble part appeared at the beginning portion of the packet. The estimation performance is expected to be improved when the channel information for each subcarrier is involved in the phase error estimator. In that case the channel with higher gain contributes more in the

estimation process while the channel experiencing more fading less contribution which can provide more reliable results.

2.2 Phase Errors by Sampling Frequency Offset

The presence of sampling frequency offset in the time domain is reflected into the linearly varying phase error in the frequency domain. Here the slope of the phase error in the frequency domain corresponds to the sampling frequency offset. Detection of the slope based on the pilot subcarriers can be summarized as follows when the well known linear regression method is applied:

$$\hat{\alpha}_i = \frac{\sum_{j\in\{pilot_index\}} j \cdot \phi_{i,j}}{\sum_{j\in\{pilot_index\}} j^2} \tag{3}$$

where $\hat{\alpha}_i$ is the estimated slope. Also i and j represent indices for OFDM symbol and subcarrier, respectively. The slope estimation optimized in least mean squared sense utilizing channel quality, payload data and the pilot can be easily derived as follows:

$$\hat{\alpha}_i = \frac{\sum_{j\in\{pilot_index\}} \left|\hat{H}_j\right|^2 \cdot j \cdot \phi_{i,j} + \sum_{j\in\{data_index\}} \left|\hat{H}_j\right|^2 \cdot j \cdot \phi'_{i,j}}{\sum_{j\in\{pilot_index\}} \left|\hat{H}_j\right|^2 \cdot j^2 + \sum_{j\in\{data_index\}} \left|\hat{H}_j\right|^2 \cdot j^2} \tag{4}$$

Here the squared channel gain for the corresponding subcarriers is used for the estimation which is expected to improve the estimation and tracking performances.

3 Simulation Results

In this section four representative tracking schemes are compared based on the BER performances. In all cases the maximum offset 40ppm is assumed which the IEEE 802.11a/g system allows. Also the packet structure follows the above mentioned high speed wireless LAN system with 500 byte payload data per packet modulated with BPSK. The RMS delay spread value is assumed to be 50ns. The pilot based scheme shows 1.9dB degradation and 'Decision + Pilot' case 0.7dB compared to the ideal performance. On the other hand the performance of 'Decision + Pilot + CSI' shows comparable performance to the ideal one. As the simulation results show, the channel state information (CSI) is very effective in the fading channels (additional 1.0dB over the 'Decision + Pilot'). The performance advantage of the payload data based scheme over conventional pilot based one is about 1.0dB gain as shown in Fig. 1.

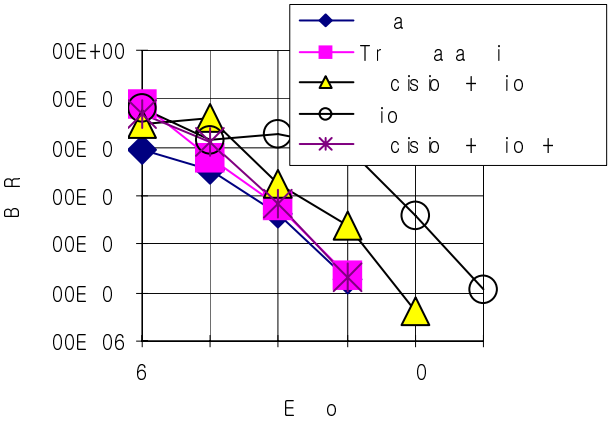


Fig. 1. BER simulation results for 50ns delay spread channel

4 Conclusion

This paper proposes an improved phase tracking algorithm where both carrier frequency offset and the sampling frequency offset are estimated and tracked to reduce the packet error rate which is one of the major QoS factors for high capacity multimedia wireless network. In the proposed method the estimation reliability is improved by adopting channel gain weight or channel information on the phase error estimation process. Computer based simulation results are also discussed to verify the effectiveness of the scheme.

Acknowledgements

This work was supported by Material and Component R&D program funded by Ministry of Commerce, Industry and Energy of Korea (under grant 10027927).

References

1. J. A. C. Bingham: Multicarrier modulation for data transmission: An idea whose time has come. IEEE Communication Magazine, vol. 28 (1990) 4-14
2. B. Yang, K. Letaief, R. Cheng, Z. Cao: Timing Recovery for OFDM Transmission. IEEE JSAC, vol. 8, no. 11 (2000) 2278-2291
3. J. Terry, J. Heiskala: OFDM Wireless LAN, A Theoretical and Practical Guide. Sam Publishing (2002)

EVM's Java Dynamic Memory Manager and Garbage Collector

Sang-Yun Lee[1] and Byung-Uk Choi[2]

[1] Dept. of Electronical Telecommunication Engineering, Hanyang University, Seoul, Korea
syllee@etri.re.kr
[2] Division of Information and Communications, Hanyang University, Seoul, Korea
buchoi@hanyang.ac.kr

Abstract. In this paper, we propose the dynamic memory manager and garbage collector for the embedded Java virtual machine. In order to facilitate to memory allocation and deallocation fast, the memory manager divides a heap into various sizes and thus manages it with the unit of block which is a set of identical cell. And we propose a new 4-color based Mark & Sweep garbage collector in order to support a multi-thread.

Keywords: Garbage Collector, Memory Management, Mark & Sweep, EVM.

1 Introduction

We have developed the *Embedded Java Virtual Machine* (EVM), in which we introduce the dynamic memory manager and the garbage collector implemented in the *EVM*. The proposed memory manager divides the memory into a block and a cell. The cell is a constant sized-memory and thus becomes the unit of the object allocation and garbage collection. A block manages the cells with identical size through a linked-list. Dividing the memory into a block and a cell facilitate allocation of memory fast, because it quickly finds out the cell that fits the size of required object.

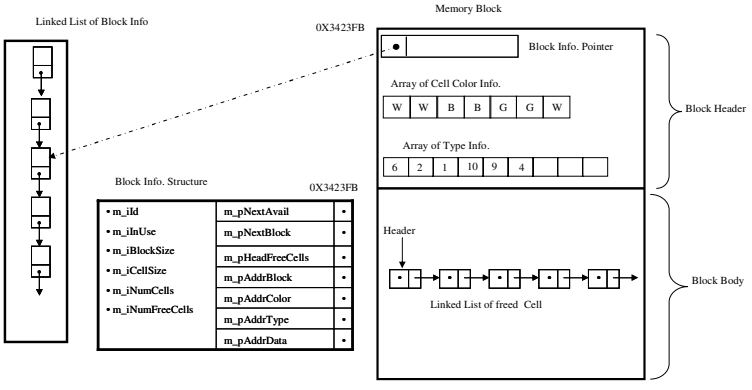
The proposed garbage collector is based on 3-color Mark & Sweep collector [1]. This garbage collector operates well without any problem in a single thread. However, in the Java program operating in a multi-thread, there is a problem that the object is collected as garbage, when a garbage collector operates before the object allocated in the other thread is defined as a root set. In this paper, we propose the 4-color Mark & Sweep garbage collector in order to solve this problem.

2 Design of Dynamic Memory Manager and Garbage Collector

We designed the memory structure considering the following criteria. Firstly, the memory allocation and deallocation speed must be fast. Secondly, we must consider that the programs operating in an embedded system normally operates with small sized objects. Thirdly, the large object must be handled with special care. Fourthly, the memory fragmentation must be small. Lastly, the expansion of memory capacity must be possible.

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 737–740, 2007.
© Springer-Verlag Berlin Heidelberg 2007

We adopted a free-list [2] to satisfy them. Fig. 1 shows the memory block structure designed with the free-list.



(a) Memory Region managed by the System (b) Memory Region managed by the VM

Fig. 1. The memory architecture

The memory area is divided into the system management memory region and the VM management memory region; the former obtains the whole heap from the system and stores the objects for controlling the latter and the latter allocates or deallocates an object dynamically in a virtual machine. And the latter is further divided into a block header and a block body. In the block header, the pointer of the block information structure, the color information of cells belonging to the block, and type information of memory allocation are stored.

The number of blocks is determined by the initial hip size assigned when the Java virtual machine is initiated. And the block memory is created during the Java virtual machine's initialization. The memory is additionally allocated from the system if the requested memory allocation exceeds the maximum heap size.

The Memory Map has three types. The *AvailableBlock* is the block with an available cell. The *EmptyBlock* is the place where all cells are deallocated by the garbage collector. *FilledBlock* is the block with no more available cells. Firstly, the available block is searched in the *AvailableBlock*. and the cell to be allocated is then appointed. If it is unable to find the cell in the *AvailableBlock*, new cell is allocated in the *EmptyBlock*. A new block and a new cell are created if it is unable to more allocate in the *EmptyBlock*. Specially, when there is request for allocation greater than 4,040 bytes it is allocated in the *EmptyBlock*.

The proposed garbage collector is based on the 3-color garbage collector that is easy to implement and its overhead is low in comparison with the other algorithm. The 3-color based *Mark & Sweep* garbage collector runs normally in the single threaded environment, however it can cause serious problems in the multi-threaded environment [3,4].

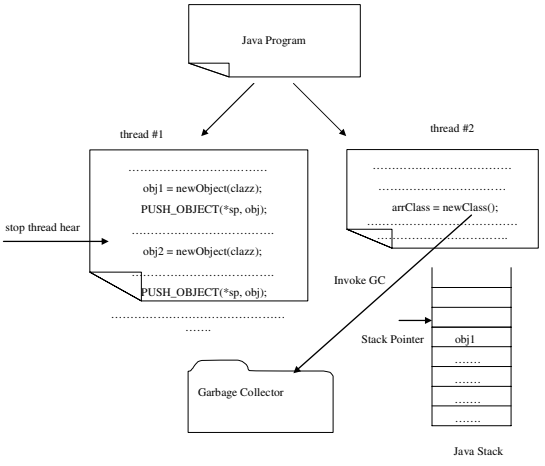


Fig. 2. The Problem of 3-color based Mark & Sweep Garbage Collector

Fig. 2 is an example of programs in which the Java application runs two threads at the same time. The source code of the thread #1 and thread #2 is the internal implementation code of the Java virtual machine. In the thread #1, the objects were stored within an obj1 and an obj2. The obj1 was stored in the Java stack by performing the PUSH OBJECT() function.

By the way, assume that the thread #1 was stopped due to the thread #2's calling the garbage collector, before an obj2 was stored in the Java stack. As shown in the Java stack memory structure, an obj2 was allocated from the memory manager but it is in the state of not going over to the Java stack. Therefore, the garbage collector determines the obj2 as garbage and collects it. Consequently, when the garbage collection is finished, thread #1 runs again the obj2 cannot actually indicate the memory address value.

We propose the 4-color based *Mark & Sweep* garbage collector to solve this problem. The proposed garbage collector has additional color more called *YELLOW* besides 3 colors used in 3-color based *Mark & Sweep* garbage collector. The object with *YELLOW* color represents that it is not yet included in the root set but it is excluded from target object during the garbage collection. When an object is created, the object does not have the *WHITE* color but the *YELLOW* color. And it is changed to the *WHITE* color on appropriate time. We managed the memory for the *YELLOW* color object specially.

3 Experimental Result

We measured how the memory fragmentation changed as the number of object increased. We classified the objects into three classes for an experiment. In group A, we continuously created array of integer of which size is 10. In group B, we continuously created array of double continuously of which size is 10. In group C, we continuously created array of integer, double, and float of which size is 10. Fig. 3 shows the experimental result.

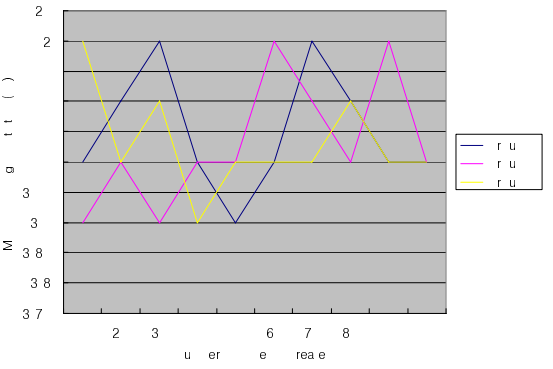


Fig. 3. Fragmentation variation according to creation of object

As shown in Fig. 3, we can know that although the number of objects created increases the memory fragmentation ratio is steadily maintained.

4 Conclusion

In this paper, we proposed the memory manager and garbage collector of the EVM, an embedded Java virtual machine. The memory manager adopted the free-list. This method has the disadvantage of the memory fragmentation but the memory allocation is fast. We designed the memory manage with blocks and cells efficiently. So, the internal and external memory fragmentation was very low (within 10%).

We adopted 3-color based *Mark & Sweep* garbage collector, whereas we proposed 4-color based Mark & Sweep garbage collector to support a multi-threaded environment.

References

1. Narendran Sachindran, J. Eliot B. Moss, Emery D. Berger: MC^2: High-Performance Garbage Collection for Memory-Constrained Environments, In proceedings of 19th annual ACM SIGPLAN conference on OOPSLA (Oct. 2004) 81-98
2. H. Toledano, M. T.: Toward an Analysis of Garbage Collection Techniques for Embedded Real-Time Java systems, 12th IEEE International Conference on Embedded and Real-Time Computing System and Applications (2006) 97-100
3. H. Inoue, D. Stefanovic, S. Forrest: On the prediction of Java object lifetimes, IEEE Transactions on Computers, Vol. 55. Issue 7 (July 2006) 880-892
4. W. Liu, Z. Chen, S. Tu: Research and analysis of garbage collection mechanism for real-time embedded java, In Proceedings of International Conference on Computer Supported Cooperative Work in Design (May 2004) 462-468

An Access Control Model in Lager-Scale P2P File Sharing Systems*

Yue Guang-xue[1,2,3], Yu Fei[3,4], Chen Li-ping[1], and Chen Yi-jun[1]

[1] College of Information Engineering, Jiaxing University, 314000 China
guangxueyue@yahoo.com.cn
[2] Department of Computer Science and Technology, Huaihua University, 418000 China
guangxueyue@126.com
[3] State Key Lab. for Novel Software Technology, Nanjing University, 210093 China
yufei@hunau.edu.cn
[4] Jiangsu Provincial Key Laboratory of Computer Information Processing Technology, Suzhou University ,2150063, China
hunanyufei@126.com

Abstract. In lager-scale P2P file sharing systems, peers often must interact with unknown or unfamiliar peers without the benefit of trusted third parties or authorities to mediate the interactions. The decentralized and anonymous characteristics of P2P environments make the task of controlling access to sharing information more difficult. In this paper, we identify access control requirements and propose a trust based access control framework for P2P file-sharing systems. The model integrates aspects of trust and recommendation, fairness based participation schemes and access control schemes.

1 Introduction

In Peer-to-Peer (P2P) networks, all peers are both consumers and providers of resources and can access each other directly without intermediary peers. Compared with a centralized system, a P2P system provides an easy way to aggregate large amounts of resources residing on the edge of Internet, and it with a low cost of system maintenance[1]. By reason of peers are heterogeneous, some peers might be benevolent in providing services, and some might be even malicious by providing bad services or harming the consumers. There is no centralized node to serve as an authority to monitor and punish the peers that behave badly, malicious peers have an incentive to provide poor quality services for their benefit because they can get away[2]. Some traditional security techniques, such as service providers requiring access authorization, cannot prevent from peers providing variable-quality service. So a major challenge for large-scale P2P file sharing systems is how to establish trust between different peers without the benefit of trusted third parties or authorities. Mechanisms

* The research work was supported by the Natural Science Foundation of hunan provincial (No: 05FJ3018) and Open Science Foundation of State Key Lab. for Novel Software Technology, Nanjing University (No: A2006-06).

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 741–744, 2007.
© Springer-Verlag Berlin Heidelberg 2007

for trust and reputation can be used to help peers distinguish good or bad partners, which are about generation, discovery, and aggregation of rating information in P2P systems.

2 Access Control Model Architecture

There are requirements that an access control model for P2P file sharing systems[3]: Peers in a P2P system are typically loosely coupled, and interacting partners are mostly unknown. Access control mechanisms must provide a mechanism for a host peer to classify users and assign each user different access rights; The incentive for users to join a P2P file-sharing network is the availability and richness of files the system provides. Access control mechanisms needs to give peers the ability to control access to their files it must still encourage them to share their files; The open and unknown characteristics of P2P make it an ideal environment for malicious users to spread unsolicited and harmful content. Access control mechanisms should support mechanisms to limit such malicious spreading, harmful digital content and punish those who are responsible for it.

Peers in P2P file-sharing systems need the autonomy of controlling accesses to their files. Host peer as a stand-alone system where shared files are objects that need to be protected and client peers are subjects who are considered to possess, or gain access rights. Files on a host peer are rated depending on their size and content. They computed from combinations of scores: trust and contribution. The client peer is responsible to collect recommendations that contain the information needed to evaluate its access values for a particular host. After each download transaction, direct trust and direct contribution of both the client peer and host peers are updated accordingly to the satisfaction level of the transaction, which then affect the future evaluation of the access values between these two peers. The architecture sees Figure 1.

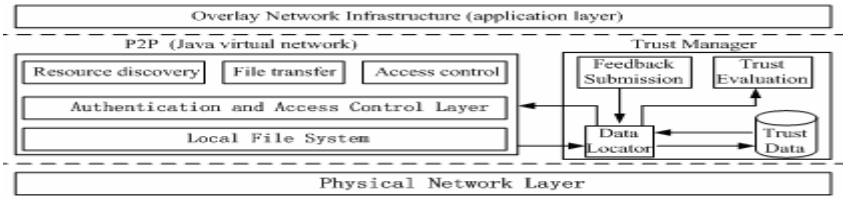


Fig. 1. Overlay architecture of a peer with access control implementation

2.1 Trust and Reputation Metrics

Trust is defined as a peer's belief in attributes such as reliability, honesty of the trusted peer. Trust can be broadly categorized by the relationships between the two involved peers. A peer's belief in another peer's capabilities, honesty and reliability based on its own direct experiences, it measures whether a service provider can provide trustworthy

services; References refer to the peers that make recommendations or share heir trust values. It measures whether a host can provide reliable recommendations.

Reputation of a peer defines an expectation about its behavior, which is based on recommendations received from other peers or information about the peer's past behavior within a specific context at a given time. It can be decentralized, computed independently by each peer after asking other peers for recommendations.

The peer's trust and reputation development cycle see Figure 2[4].

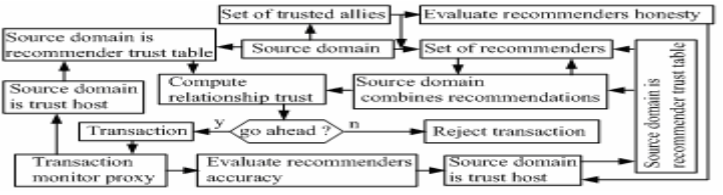


Fig. 2. Trust and reputation development cycle

2.2 Scoring Scheme for Access Control Model

As a host peer needs to classify its client peers in order to provide them different access privileges, then model uses a scoring system to differentiate peers based on their behavior in the P2P network. Hence, after completing the authentication process with the host, a client peer is required to supply its rating certificates for the host to calculate the client's relative access values. They indicate how the host peer perceives the client's trustworthiness and contribution level, which is to ensure the peer is trusted to interact with and to promote fairness in P2P network. There are two sources of information to compute these two values. One source is the host's direct experiences with the client, and other peer's recommendations based on their interactions with the client. Therefore, the access vales are evaluated via combinations of four types of scores: direct trust, indirect trust, direct contribution and indirect contribution. It tries to reduce the problem of unfair trading. Figure 3 represents procedures in a typical interaction between a host and a client in our framework. It can be seen the scheme adopts more of a "push" approach.

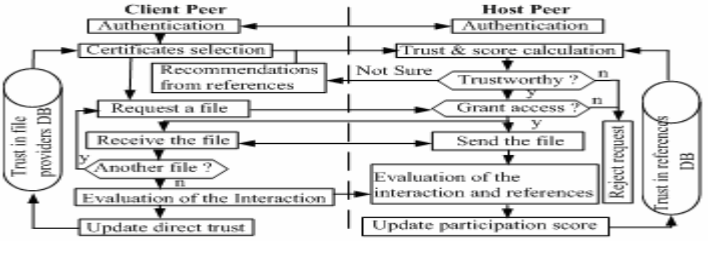


Fig. 3. Flow chart of an interaction between a host peer and a client peer

The Algorithm of peers constructing a trust graph as follows

Step1 suppose peer P_t is evaluating the trustworthiness of peer P_k and P is the set of peers being visited. $\{P_i, P_j\}$ denotes a referral r to peer P_j returned from peer P_i. P is a finite set of peers $\{ P_1, \ldots, P_n \}$, and R is a set of referrals $\{ r_1, \ldots, r_n\}$.

Step2 for ($\forall P_i \notin P$ and P_i has not been queried) do
if (Trust(P_i) < TrustBound) then $\Rightarrow P_t$ queries P_i

Step3 if (P_i is a witness of P_k) then $\Rightarrow P_i$ returns the rating about P_k to P_t, interation, compute *trust;* else

Step4 for ($\forall r = \{P_i, P_j\}$ from P_i) do, computer *recommendation*

Step5 if ($P_j \notin P$) then $\Rightarrow r \quad R, \ P_j \qquad P$; else

Step6 ignore referral r

Step7 evaluation of the interaction, references and update *DBs.*

Therefore, it could effectively solve the follows problems

(1) A client submits other peers' recommendations to a host in the form of rating certificates.

(2) It is reasonable to expect that the greater the number of peers that the client has interacted with, the more recommendations it has and the higher its indirect trust as rated by the host peer.

(3) Avoid specify a required number of recommending peers. In general, this can vary from peer to peer and depending on the type of transaction.

3 Conclusion

The proposed access control model and score systems help to classify both known and unknown visitors according to their trustworthiness and contribution. The implemented contribution scores work effectively as a payment scheme; giving incentive for users to share their resources and safeguarding the fairness of service exchange in a P2P system. The proposed mechanisms for evaluating a transaction not only help to differentiate poorly performing peers from good ones but also ensure that malicious peers are punished and isolated.

References

1. Parameswaran M, Susarla A, Whinston AB. P2P networking: An information-sharing alternative. Computing Practices, 2001,34(7):31-38.
2. Yao Wang, Julita Vassileva. Bayesian Network-Based Trust Model. Proceedings of the IEEE/WIC International Conference on Web Intelligence (WI'03)
3. Huu Tran, Michael Hitchens, Vijay Varadharajan, Paul Watters. A Trust based Access Control Framework for P2P File-Sharing Systems. Proceedings of the 38th Hawaii International Conference on System Sciences -2005.
4. Farag Azzedin, Muthucumaru Maheswaran. Trust Modeling for Peer-to-Peer based Computing Systems. Proceedings of the International Parallel and Distributed Processing Symposium (IPDPS'03).

Sink-Independent Model in Wireless Sensor Networks

Sang-Sik Kim[1], Kwang-Ryul Jung[1], Ki-Il Kim[2,*], and Ae-Soon Park[1]

[1] Electronics and Telecommunications Research Institute,
[2] Department of Information Science, Research Institute of computer and Information Science, GyeongSang National University, 900 Gajwa-dong, Jinju, 660-701, Korea
{pstring,krjung}@etri.re.kr, kikim93@gmail.com, aspark@etri.re.kr

Abstract. Wireless sensor networks generally have three kinds of objects: sensor nodes, sinks, and users that send queries and receive data via the sinks. In addition, the user and the sinks are mostly connected to each other by infrastructure networks. The users, however, should receive the data from the sinks through multi-hop communications between disseminating sensor nodes if such users move into the sensor networks without infrastructure networks. To support mobile users, previous work has studied various user mobility models. Nevertheless, such approaches are not compatible with the existing routing algorithms, and it is difficult for the mobile users to gather data efficiently from sensor nodes due to their mobility. To improve the shortcomings, we propose a view of mobility for wireless sensor networks and propose a model to support a user mobility that is independent of sinks.

Keywords: User Mobility Support, Wireless Sensor Networks.

1 Introduction

Wireless sensor networks typically consist of three objects: user, sink, and sensor node [1]. Firstly, a user is an object that disseminates an interest in the sensor field and collects data about the interest from sensor nodes. Secondly, a sink is an object that collects data. The sink receives an interest from a user and disseminates the interest inside sensor fields. The sink receives sensing data from sensor nodes and forwards the sensing data to the user. Lastly, a sensor node is an object that generates data about the interest and delivers the data to a sink.

The user and the sinks are mostly connected to each other by infrastructure networks. The users, however, should receive the data from the sinks through multi-hop communications between sensor nodes if such users move around the sensor networks without infrastructure networks. Recently, applications transmitting data to moving users inside sensor fields, such as rescue in a disaster area or maneuvers in a war zone, have been on the rise in large-scale sensor networks [5]. (Firefighters and soldiers are users gathering data from sensor networks.)

* Corresponding author.

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 745–752, 2007.
© Springer-Verlag Berlin Heidelberg 2007

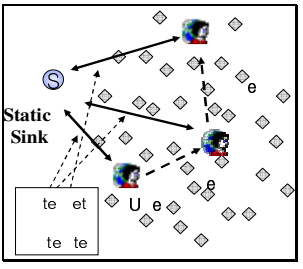


Fig. 1. Direct user-network communication model

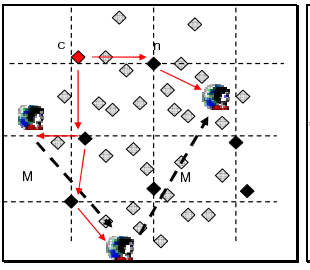


Fig. 2. GPS based user-network communication model

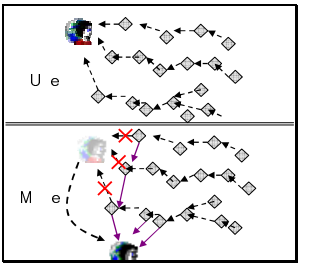


Fig. 3. Topology control based user- network communication model

To support mobile users in wireless sensor network, previous work has studied various user mobility models. But, until now, only three models supported the mobility of users for those applications: the direct user-network communication model, the GPS-based user-network communication model, and the topology-control-based user-network communication model.

The direct user-network communication model (D-COM) is shown in Fig. 1. It supports the mobility of a user on the assumption that the user communicates directly with sinks through infrastructure networks, namely, the Internet, such as communication systems in traditional sensor networks [1]. And, users can communicate directly with the networks via the sinks. But, in applications such as rescues in a disaster area or maneuvers in a war zone, circumstances without infrastructure networks, except sensor networks, are more prevalent. Hence, the assumption that a user and a sink can communicate directly is not entirely accurate.

The GPS-based user-network communication model (G-COM) is seen in Fig. 2. G-COM is source-based topology [5], [6], [7]. In G-COM, sensor nodes proactively construct a GRID system with GPS receivers. G-COM assumes that all sensor nodes have their own GPS receivers and an ability that constructs a GRID. A sensor node, i.e. source, with a stimulus is going to make a GRID in a sensor field. Once a GRID is set up, mobile user floods its interests within a cell only where the user is located. When a sensor node on a GRID receives interests, it sends interests to the source along a GRID path and data from the source are forwarded to the user.

The topology-control-based user-network communication model (T-COM) is seen in Fig. 3. This model supports the mobility of the user by reflecting the movement of the user [8], [9]. In T-COM, the user and sensor nodes construct a tree that is rooted at the user. The user always maintains the tree and gathers data. Intuitively, G-COM and T-COM seem to be suitable for the aforementioned applications. But, these models also have various problems. First of all, they cannot use existing effective data collection algorithms [2], [3], [4] between a sink and sensor nodes because of low protocol compatibility. Accordingly, such algorithms can hardly be exploited if users in sensor networks have mobility. The other problem is that the cost of the overhead to reorganize the network topology and reconstruct dissemination paths is expensive. In G-COM, all sensor nodes make the topology based on location information. Accordingly, each sensor node must have its own GPS receiver. The cost

Table 1. Taxonomy of Mobility Type

Mobility Type	Compatibility with Existing Static Sink Routing Protocols	Feasibility	GPS receivers for sensors	Control Overheads according to user mobility	Control Overheads to support multiple users	Help of infrastructure networks
D-COM	High	Low	Needless	Low	Low	Mandatory
G-COM	Low	Middle	Mandatory	Middle	Low	Needless
T-COM	Low	High	Needless	High	High	Needless
A-COM	High	High	Needless	Low	Low	Needless

of GPS receivers is decreasing, but the overall cost is still high. In T-COM, similarly, user mobility causes topology reconstruction. Users in T-COM have a tree that is rooted at each mobile user. If users move into a new location, then the root of trees must be changed, as seen in Fig. 3. This leads to enormous overhead to sensor nodes.

Hence, this paper proposes **a novel agent-based user-network communication model (A-COM)**. A-COM collects data through a temporary agent and delivers the data to mobile users. In A-COM, the user appoints a sensor node to act as an agent, and the agent forwards interests to the sink. The sink collects data from sensor nodes using the existing data collection algorithm in static sink sensor networks [2], [3], [4]. The collected data are finally forwarded to the user. (If there is no sink, the agent directly disseminates interests and collects data.) A-COM has various advantages, as can be seen in Table 1. First of all, A-COM has the compatibility with existing static sink routing protocols without infrastructure networks. In addition, the users in A-COM do not make a topology (tree or GRID) and communicate only with agents. So, the users are free from topology control. The user's freedom saves energy and enables more users to participate in this model even if the sensors have no GPS receivers.

2 Model Analysis

In our model, if a user intends to obtain data while moving, the user appoints a sensor to act as an agent and forwards an interest to the agent. If there is one or more sink(s), the agent forwards interests to sensor networks via sink(s). The number of sinks, however, depends on the network policy. A network administrator might want to set a single or more sinks in the sensor field, or alternatively the sensor field may be hazardous as he cannot reach the field. Hence, we consider three scenarios according to the number of sinks and describe the scenarios based on following assumptions.

• A user can communicate with static sinks only through sensors because networks within sensor fields are infrastructure-less networks. In addition, It is possible that multiple sinks are deployed in sensor networks and are connected to each other.

• The data which one sink collects is aggregated by the sinks. The aggregated data is shared by every multiple static sinks through the infrastructure network.

• The interest describes how many times the sink forwards the gathered data set.

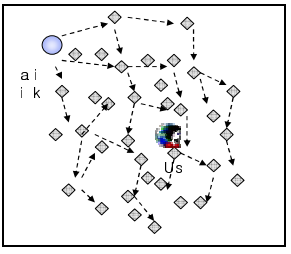


Fig. 4. Dissemination of Sink Announcement Message

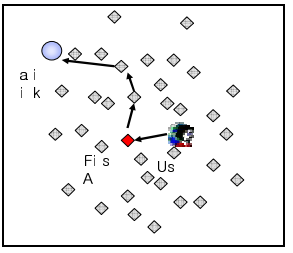


Fig. 5. Interest Dissemination of the user

2.1 Scenario 1: Sensor Fields with Only One Sink

Dissemination of sink announcement message and user interest. In the initial stage of the sensor network, a sink floods a sink announcement message to announce itself inside the whole sensor field (Fig. 4). As a result of the flooding announcement message, every sensor node knows the hop counts and next hop neighbor sensor node. While moving inside the sensor fields, the user selects the nearest node as a first agent, as shown in Fig. 5. The user delivers an interest to the first agent. The first agent forwards the interest to the next hop neighbor node toward the sink. The next hop sensor node, which has had the interest delivered to it, also forwards the interest to the next hop neighbor node toward the sink. This process continues until the sink receives the interest of the user. Also, a route for the interest from the sink to the user has been established through this process. The established route vanishes from the network when the described period in the interest is over.

Data collection. A sensor network with a static sink is a network where sensing data from sensor nodes should be transmitted to the static sink through multi-hop communication. So, existing routing algorithms [2], [3], [4] for a static sink can be used (e.g., routing algorithms collecting data by periods, routing algorithms collecting a minority event, or routing algorithms detecting a moving object.) In Fig. 6, the static sink can forward interests from users to sensors and gather data from sensor networks according to the existing routing protocols. If all data are gathered by routing protocols, the static sink aggregates all data and forwards an aggregated data to the first agent. A user may move to another place after sending an interest to the first agent. In this case, the user selects another agent that can communicate with the first agent. Also, the user makes a new connection between the newly selected agent and the original agent. These agents are used for forwarding the aggregated data.

2.2 Scenario 2: Sensor Fields with Multiple Sinks

Basically, the difference between Scenario 1 and Scenario 2 is only the number of sinks. If there are more than one sink in the sensor field, this means a separation of the sensor fields. As a result of sink announcement message dissemination in this case, all

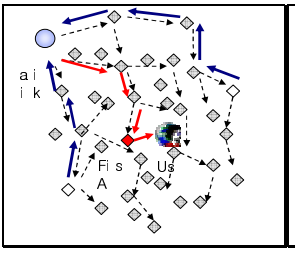


Fig. 6. Data Propagation to the user

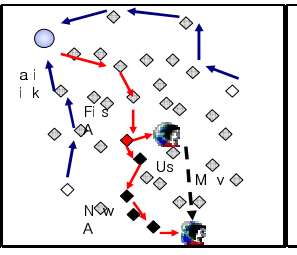


Fig. 7. Mobility support of the user

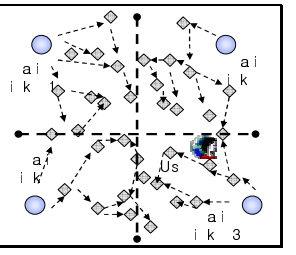


Fig. 8. Seperation of the Sensor Fields

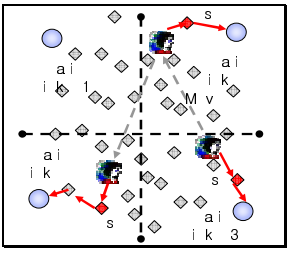


Fig. 9. Interest Dissemination of the user with multiple sinks

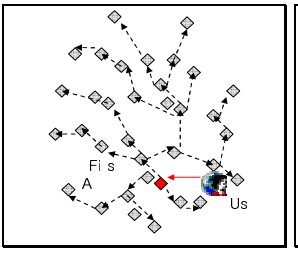


Fig. 10. First Agent Selection and Announcement

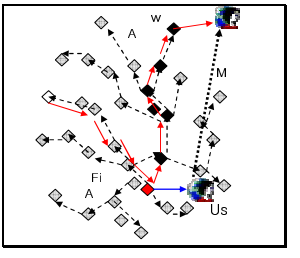


Fig. 11. Mobility support of the user

sensor nodes know the nearest sink according to the hop counts. Accordingly, Interest dissemination of the user targets the nearest sink from the agent, as shown Fig 8. The targeted sinks can be changed whenever the user wants to send its interests (see Fig. 9). Nevertheless, Mobility support of the user and data propagation of the sink is still the same with Scenario 1. In addition, users may not be able to recognize how many static sinks are in the sensor fields. This means that the proposed model is independent of the number of sinks. A user can receive the data from the nearest sink to its position. Short hops communications between a user and a sink are possible. This saves energy, enhances the data delivery ratio, and reduces delay.

2.3 Scenario 3: Sensor Fields with No Sink

The sensor fields without a sink are a special type of the sensor networks. If the sensor field is hazardous as network administrator cannot reach the field (ex. Battlefield), it may not have any sinks. Because there is no sink in the sensor field, the sensor network cannot practice sink announcement message dissemination process for itself. In this case, users must appoint the nearest sensor node as first agent, and the first agent disseminates sink announcement message. As shown Fig 10 and 11, users examine nearby sensor nodes whether there is a sink in the sensor field or not. If not, users appoint the nearest sensor node to first agent. Once a sensor node becomes first agent, it acts like the sink of Scenario 1. Hence, other processes such as sink announcement message dissemination and data propagation of the sink are the same as Scenario 1. The first agents must return to the original state after the described

period. This means that the first agents are appointed whenever users want to send its interest. Then, the first agents are reactively selected and practice all process for user mobility. In the whole network, therefore, the sensor network can remain in an idle state in case there is no user in the sensor field. This is a positive effect because there is no control messages and interests in the network.

3 Performance Evaluation

We evaluate the proposed model in Qualnet, a network simulator [12]. A sensor node's transmitting and receiving power consumption rate are 0.66W and 0.39W. The transceiver in the simulation has a 50m radio range in an outdoor area. The sensor network consists of 100 sensor nodes, which are randomly deployed in a 300m x 300m field. And the user which follows a random waypoint model of 10m/s speed and 10 second pause time moves into the sensor field. The user disseminates an interest at an interval of every 10 seconds. The simulation lasts for 500 seconds.

3.1 Impact of the Number of Static Sinks

Scenarios 1 and 2 of A-COM can be compared with the D-COM because G-COM and T-COM have no static sink. We first study the impact of the number of sinks on A-COM's performance. The number of sinks varies from 1, 2, 3, 4 to 5. And there is only one user in the sensor field. In this part, we compare Scenarios 1 and 2 to D-COM regarding lifetime, delay, and delivery ratio. Fig. 12 shows the number of interest rounds, namely, network lifetime. The number of interest rounds shows little difference between A-COM and D-COM. This means that A-COM can manage sensor fields as well as D-COM without infrastructure. In addition, the lifetime is increased according to the number of sinks. This is a side effect of multiple sinks. Sinks separate the sensor field, and besides, users only use the nearest sink to send interests and receive replies. Users can use the shortest path to communicate with multiple sinks. As a result of the shortest communication, the lifetime in A-COM is enhanced according to the number of sinks. The delay is also enhanced by this side effect of multiple sinks. A-COM basically has some delay due to multi-hop communication between users and sinks. However, the delay is diminished according to the number of sinks, as shown in Fig. 13. Nevertheless, the data delivery ratio of A-COM is comparable with D-COM, as shown in Fig. 14. This also proves that the proposed model can manage sensor fields without infrastructure.

3.2 Impact of the Number of Users

The number of users only results in path increase between users and sinks. D-COM uses direct communication between users and sinks, and A-COM uses multi-hop communication. A-COM has more paths and consumes more energy. (e.g., five users in A-COM consumes five times of the energy that is consumed by one user.) However, it is a tradeoff between energy and infrastructure. Although A-COM has more energy consumption and delays than D-COM, the merit of A-COM is infrastructure-less communication systems.

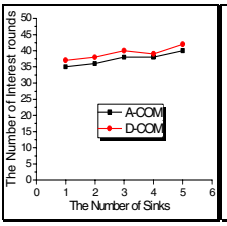


Fig. 12. Network lifetime for the Number of Sinks

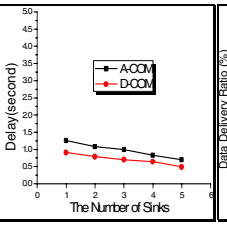


Fig. 13. Delay for the Number of Sinks

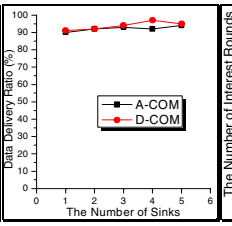


Fig. 14. Data Delivery Ratio for the Number of Sinks

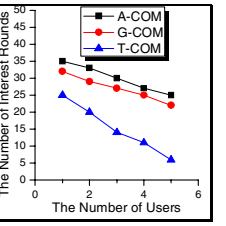


Fig. 15. Network Lifetime for the Number of Users

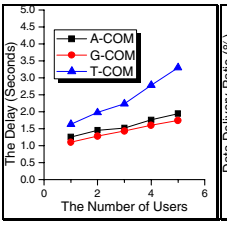


Fig. 16. Delay for the number of Users

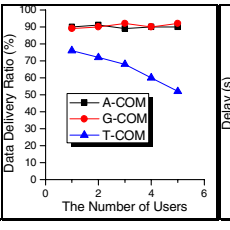


Fig. 17. Data Delivery Ratio for the Number of Users

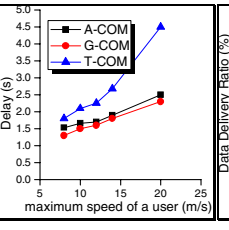


Fig. 18. Delay for User Speed

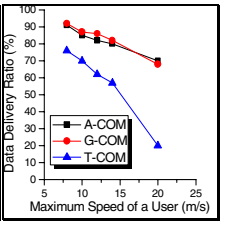


Fig. 19. Data Delivery Ratio for User Speed

Scenario 3 of A-COM can be compared with G-COM and T-COM because Scenario 3 of A-COM, G-COM, and T-COM have no static sinks. There are no sinks, and the number of users varies from 1, 2, 3, 4 to 5. In this part, we compare Scenario 3 of A-COM to G-COM and T-COM regarding lifetime, delay, and delivery ratio. G-COM and T-COM make and change the topology proactively, but Scenario 3 of A-COM reactively makes and shares it among users. Generally, users move about the sensor field only and generate its interest occasionally. Hence, sensors in Scenario 3 can save considerable energy. Alternatively, sensors in G-COM and T-COM maintain a topology continuously. Fig. 15 shows each lifetime of these sensor networks. As shown in Fig. 15, the lifetime of T-COM is considerably low due to frequent topology change and that of G-COM is relatively low due to GRID maintenance.

In Fig. 16, G-COM has little delay due to proactive GRID topology by the GPS receiver. T-COM proactively creates the topology, but frequent topology changes of T-COM delay data delivery considerably. The delay of Scenario 3, as shown in Fig. 16, however, is only a little high due to the reactive first agent selection and topology construction. In the case of the data delivery ratio, A-COM and G-COM in Fig. 17 are similar except for T-COM. The reason is frequent topology change. Topology change messages disturb the data delivery ratio.

3.3 Impact of the User Mobility

We lastly evaluate the impact of user speed on A-COM. We vary the maximum speed of a user from 8, 10, 12, 14 to 20m/s. We assume that there is one user in the sensor field. In this part, we compare Scenario 3 to G-COM and T-COM because D-COM is independent of user speed. Fig. 18 shows the delay in data delivery, which slightly

increases as the user moves faster. The delay depends on a movement operation that is processed by the user. The faster a user moves, the more the time is needed to establish a connection between the user and the network. Nevertheless, the delay of A-COM is comparable with G-COM because A-COM creates only one communication path between the user and its first agent. The delay of T-COM, on the other hand, is relatively higher than the others due to frequent topology changes.

And, Fig. 19 shows the data delivery ratio when the user's moving speed changes. The data delivery ratio of A-COM is slightly decreased according to the delay. But the data delivery ratio remains around 0.7 - 0.9; nevertheless, the user moves faster. Besides, the data delivery ratio of G-COM remains high because the GPS receiver may help the user with geographical routing. On the other hand, the data delivery ratio of T-COM is relatively lower than the others because it has too many topology changes when moving. The results in Fig. 18 and Fig 19 mean that A-COM is fast and stable without GPS receiver.

4 Conclusion

In this paper, we propose a novel agent-based user-network communication model to support the mobility of users in wireless sensor networks. In the proposed network model, the user can receive data with a higher data delivery ratio and in a faster time without infrastructure. We verified that the lifetime of sensor networks is prolonged because the reactive path construction decreases the energy consumption of sensor nodes. Also, we verified that performance of the data delivery ratio and the delay never falls; nevertheless, communication between the user and the network for guaranteeing movement of the user is supported by only sensor nodes.

References

1. I.F. Akyildiz, *et al.,* "A survey on sensor networks," Communications Magazine, Aug. 2002.
2. C. Intanagonwiwat, *et al.*, "Directed diffusion: A scalable and robust communication paradigm for sensor networks," ACM Mobicom, 2000.
3. C. Schurgers, *et al.*, "Energy efficient routing in wireless sensor networks," MILCOM 2001.
4. W.R. Heinzelman, *et al.,* "Adaptive Protocols for Information Dissemination in Wireless Sensor Networks," ACM Mobicom, 1999.
5. F. Ye, *et al.*, "A Two-Tier Data Dissemination Model for Large-scale Wireless Sensor Networks," ACM MobiCOM, Sept. 2002.
6. S. Kim, *et al.,* "SAFE: A Data Dissemination Protocol for Periodic Updates in Sensor Networks," Distributed Computing Systems Workshops 2003.
7. H. L. Xuan and S. Lee, "A Coordination-based Data Dissemination Protocol for Wireless Sensor Networks," IEEE ISSNIP, Dec. 2004.
8. K. Hwang, *et al.,* "Dynamic sink oriented tree algorithm for efficient target tracking of multiple mobile sink users in wide sensor field," IEEE VTC, Sep. 2004.
9. S. R. Gandham, *et al.,* "Energy Efficient Schemes for Wireless Sensor Networks with Multiple Mobile Base Stations," IEEE GLOBECOM, Dec. 2003.
10. Scalable Network Technologies, Qualnet, available: http://www.scalable-networks.com.

An Update Propagation Algorithm for P2P File Sharing over Wireless Mobile Networks*

Haengrae Cho

Department of Computer Engineering, Yeungnam University
Gyungsan, Gyungbuk 712-749, Republic of Korea

Abstract. Peer-to-peer (P2P) file sharing systems often replicate files to multiple nodes. File replication is beneficial in the sense that it can achieve good query latency, load balance, and reliability. However, it introduces another problem of maintaining mutual consistency among replica when a file is updated. The new file has to be propagated to all of its replica. In this paper, we propose an update propagation algorithm for P2P file sharing over wireless mobile networks (MONET). Compared to the previous algorithms proposed in wired P2P file sharing systems, our algorithm has low communication overhead. It also guarantees safe delivery of updates even when the underlying network is unreliable. This means that our algorithm is well matched to the characteristics of MONET such as limited battery power of mobile nodes, lower network reliability, and frequent disconnection of mobile nodes.

Keywords: Mobile network, P2P, file sharing, update propagation.

1 Introduction

Peer-to-peer (P2P) computing is becoming a very popular computing paradigm due to the wide diffusion of file sharing applications [1]. Representative P2P systems such as Gnutella [6] and Kazaa [10] have millions of nodes sharing petabytes of files over Internet. The complexity of a file discovery in P2P systems would be very high since it may require scanning the entire network to find a required file. Efficient file replication can reduce the complexity [3,7]. However, it introduces another problem of maintaining mutual consistency among replica when a file is updated. The new file has to be propagated to all of its replica.

The growing number of mobile devices, together with the proliferation of wireless and pervasive communication technologies, also demand for the adoption of P2P paradigms in wireless mobile networks (MONET) [4,5,9]. An update propagation algorithm for MONET based P2P file sharing systems has to deal with the following issues arising from constraints specific to MONET.

- **Limited battery power**: A node in MONET runs using its battery power. Then the node might lose its power rapidly if it transmits information heavily. This means that the update propagation algorithm should be able to reduce message traffic between nodes.

* This research was supported by University IT Research Center Project.

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 753–760, 2007.
© Springer-Verlag Berlin Heidelberg 2007

- **Frequent disconnection of mobile nodes**: Mobile nodes often get disconnected from the network due to power failure or due to their mobility. In addition, some mobile users switch their units on and off regularly to save power, causing more network disconnections. So the update propagation algorithm should be able to deliver missing updates to mobile nodes when they are reconnected to the system.
- **Unreliable network**: An wireless connection is more unreliable than an wired one. This means that the update propagation algorithm has to be resilient to the message lost.

The P2P file sharing over MONET is still in its early stage [4,5,9]. Most of previous work in this area investigate only the file discovery algorithm. They do not consider file replication or do not describe how replicated files get same values if any of them is updated. In this paper, we propose a new update propagation algorithm for P2P file sharing over MONET. The novel features of our algorithm are as follows.

- The algorithm is *purely decentralized*. All updates must be eventually propagated to their replica, but update propagation between nodes is performed in asynchronous manner. This is a great advantage in MONET based P2P systems that may experience transient failures and network partitioning.
- The algorithm has *lower communication overhead*. It tries to reduce duplicate delivery of an update message between nodes. The lower communication overhead can contribute power savings of mobile nodes and efficient usage of limited network bandwidth.
- The algorithm is *adaptive* to the dynamic behavior of MONET. When most neighbor nodes leave the network, a node reconstructs new neighborship autonomously with other nodes. The movement of mobile nodes also causes the neighborship reconstruction. This contributes to fault-tolerance and fast update propagation.

The rest of this paper is organized as follows. Section 2 introduces previous work in wired P2P systems. Section 3 describes the proposed algorithm in detail. Section 4 presents the experiment results. Finally, Section 5 concludes the paper.

2 Update Propagation in Wired P2P Systems

There are two types of update propagation algorithms in wired P2P systems: *push* and *pull*. In the push-based algorithm [6,8], a new update is pushed by an initiator to its neighbor nodes, which in turn propagate it to their neighbor nodes. In the pull-based algorithm [2], a node polls periodically to one of neighbors if there are any new updates. The push-based algorithm can provide good consistency guarantees for nodes that are online and reachable from the initiator. However, it suffers from high communication overhead due to duplicate message delivery. Figure 1.(a) illustrates a case of duplicate message delivery. Both node C and node D propagate a same update message to each other. The pull-based

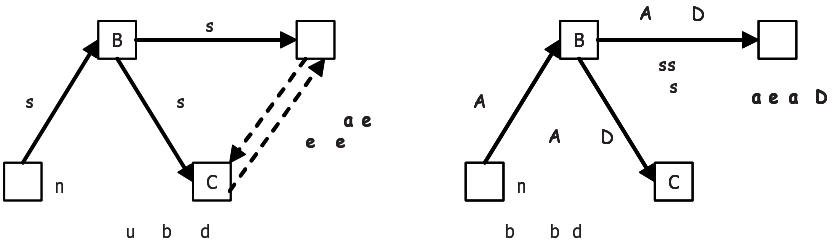


Fig. 1. Problems of previous algorithms

algorithm can reduce the communication overhead but causes longer propagation delay and weaker consistency guarantees. The consistency guarantees in pull are critically dependent on the effectiveness of polling.

To combine the best features of push and pull, a *hybrid push-pull* algorithm [3] is proposed. An updating node initiates a push phase by propagates the new value to its neighbors. A push message includes a receiver list. The receiver list includes identifiers of nodes to which the same message has been sent. The algorithm can avoid duplicate message delivery by restricting an update message to be sent to any node in the receiver list. However, a node could not receive an update at the push phase if messages are lost. Figure 1.(b) illustrates the case of lost update. A receiver list sent to C is $[A, B, C, D]$. Then node C does not propagate an update message to node D because the receiver list includes D. When a node joins the network, it enters a pull phase to synchronize and reconcile. The pull phase also introduces another problem. At the pull phase, a node inquires for missed updates based on version vector to one of its neighbors. However, a version vector of entire replicated data items is a huge unit for messaging and consumes lots of CPU cycles of recipient node for version comparison. A selective pull on part of replicated files is strongly required.

3 Update Propagation over MONET

In this section, we propose a new hybrid push-pull algorithm, named UPM (*Update Propagation over MONET*) for MONET based P2P file sharing systems. UPM consists of four sub-algorithms for push, selective pull, entire pull, and reconstruction of neighborship.

3.1 Data Structures

A node creates new neighborship with other nodes when it joins a P2P system. Gnutella's ping-pong protocol [1] is an example to make such neighborship. A node N_i maintains the identifiers of its neighbor nodes at $neighbor(i)$.

We assume that each file has an owner. The owner is a node that has a role to synchronize updates on the file. Initial owner is set to a node that creates the file. When a non-owner node tries to update the file, it first contacts the owner.

If the owner grants the update operation, the ownership is changed to the new node. The old owner keeps the information of new owner so that it can forward following update requests to the new owner.

Each node N_i maintains an *update counter* $UC(i)$. N_i increments $UC(i)$ whenever it updates any file. N_i also maintains an *update counter vector* UV_i for other nodes. $UV_i[k]$ stores a value of $UC(k)$ that N_i knows. Finally, N_i maintains an *update history table* H_i. If $H_i[k, UC(k)] = f$, then N_i knows that another node N_k updates a file f at $UC(k)$. The initial settings of the data structures are $UC(i) = 0$, $UV_i[*] = 0$, and $H_i[*, *] = \emptyset$. The contents of UV and H are updated at push phase and pull phase.

3.2 Push Phase

When an owner updates a file, it initiates update propagation to its neighbor nodes. The update message consists of six attributes: <sender, owner, file id, new file, update counter of initiator, sender list>. The sender list is a set of nodes that have sent the same message. Let us suppose that node N_i updates a file f. Then the following steps are performed.

(1) $UC(i) = UC(i) + 1$
(2) $H_i[i, UC(i)] = f$, $UV_i[i] = UC(i)$
(3) To every node in *neighbor*(i), send an update message $<i$, i, f, value of f, $UC(i)$, $\{i\}>$.

Since N_i is both an owner and a sender, the update message includes i twice. The sender list includes only i with the same reason. Now let us suppose that node N_k receives an update message $<s$, i, f, value of f, $UC(i)$, $\{i, s\}>$ from node N_s. Then N_k performs the following steps to process the message.

(1) If $UC(i) == UV_k[i] + 1$, N_k has received every update message of N_i that was sent before $UC(i)$. So N_k needs to process this message only as follows.
   - $H_k[i, UC(i)] = f$, $UV_k[i] = UC(i)$
   - Replace N_k's local copy of f by the value of f at the message.
   - To every node in *neighbor*$(k) - \{i, s\}$, send an update message $<k$, i, f, value of f, $UC(i)$, $\{i, s, k\}>$.
(2) Else if $UC(i) == UV_k[i]$, this is a duplicate message. N_k just ignores it.
(3) Else if $UC(i) > UV_k[i] + 1$, there are some updates that N_k did not receive. Then N_k inquires for the missed updates to N_s at the pull phase.

If a node is in the sender list, it actually received the message. So other nodes do not need to send the message to any nodes in the sender list. As a result, compared with sending the receiver list [3], UPM can increase the probability of receiving the update message. Figure 2.(a) illustrates a scenario that UPM can propagate an update when there is a message lost. Node D can receive the update from other link of node C. Furthermore, unlike the pure push-based algorithm, D need not propagate the update to node B again since the sender list includes B. This contributes to reduce the duplicate message delivery.

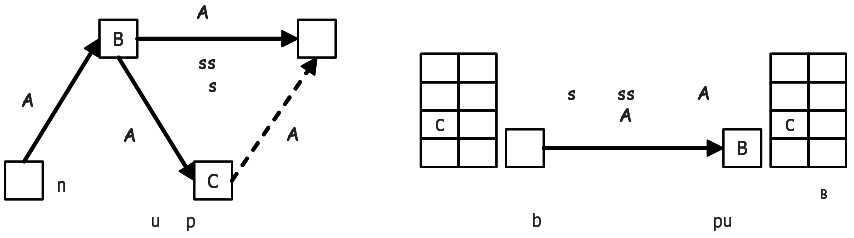


Fig. 2. Scenarios of push phase and selective pull of UPM

3.3 Pull Phase

A node performs a pull phase at two cases: when it detects any missed updates at the push phase (*selective pull*), or just after it joins the system (*entire pull*).

Selective Pull. Let us suppose that a node N_k receives an update message $<s, i, f,$ value of $f, UC(i), \{i, s\}>$ from node N_s at a push phase. If $UC(i) > UV_k[i] + 1$, N_k performs a selective pull as follows.

(1) N_k sends a selective pull message $<k, i, UV_k[i]>$ to N_s.
(2) N_s makes a response message PR as follows.
  – $PR = \{\}$
  – For $(t = UV_k[i] + 1; t \leq UV_s[i]; t = t + 1)$
    $PR = PR \cup <i, H_s[i, t],$value of $H_s[i, t], t>$.
(3) N_s returns PR to N_k. Then N_k updates its data structures and local copies as follows.
  – For each tuple $<i, f, v, t> \in PR$
    Set $H_k[i, t] = f$ and replace N_k's local copy of f by v.
  – $UV_k[i]$ is set to the maximum value of t for all tuples in PR.

Figure 2.(b) illustrates a case when a selective pull happens. Node A is an initiator of push phase and propagates an update message with $UC(A) = 3$. Node B missed an update of A with $UC(A) = 2$, and thus $UV_B[A]$ is set to 1. Since $UC(A) > UV_B[A] + 1$, node B can detect that it missed some updates. At the selective pull phase, node B first sends $UV_B[A]$ to node A. Then node A returns all of its updates after $UV_B[A]$ to node B.

Entire Pull. When a node joins a P2P system, it needs to validate the currency of its local copies by performing an entire pull. Let us suppose that node N_k joins the P2P system. The entire pull consists of the following steps.

(1) N_k selects a node $N_s \in neighbor(k)$ in a random manner. Then N_k sends an entire pull message $<k, UV_k>$ to N_s.
(2) N_s makes a response message PR as follows.
  – $PR = \{\}$

- For each node n where $UV_k[n] < UV_s[n]$, perform the following steps.
  For $(t = UV_k[n] + 1; t \leq UV_s[n]; t = t + 1)$
  $PR = PR \cup <n, H_s[n,t]$,value of $H_s[n,t], t>$.

(3) N_s returns PR to N_k. Then N_k updates its data structures and local copies as follows.
- For each tuple $<i, f, v, t> \in PR$
  Set $H_k[i,t] = f$ and replace N_k's local copy of f by v.
- $UV_k[i]$ is set to the maximum value of t for all tuples of i in PR.

Note that the random selection of target node would not be optimal. This is especially true when the target node has missed some updates also. However, both the push phase and the selective pull can guarantee to get the missed updates eventually. If a node contacts multiple nodes at the entire pull, it may reduce the amount of missed updates. This in turn increases the communication overhead and consumes CPU cycles of more neighbor nodes.

3.4 Reconstruction of Neighborship

The correctness of UPM depends on the connectivity of a node with its neighbor nodes. If most neighbors leave the P2P system or the node moves to other network, it would not receive any push messages. To detect such condition, each node sends periodically an `are_you_alive` message to every neighbor node. The living nodes reply to the message. If the number of living nodes is under the threshold value, the node has to reconstruct new neighborship with other nodes by Gnutella's ping-pong protocol [1]. Furthermore, if the node has been isolated completely (i.e. the number of living nodes is zero), it has to perform an entire pull to one of the new neighbors.

4 Experiments

We develop an experiment model of a MONET based P2P system using CSIM [11] simulation package. We compare UPM with two algorithms, PO (push only) and LRA (List of Receivers Algorithm). PO is a pure push algorithm [6]. LRA is a hybrid push-pull algorithm [3]. Table 1 summarizes simulation parameters. We model a P2P system with a limited number of data items and high update ratio. This setting helps us investigate the differences between algorithms. There are three types of nodes: online, join, and leaving. Join nodes are off-line initially and then join the network with a probability of *OnlineRate*. Online nodes leave the network with a probability of *LeaveRate*. *MsgLossRate* is a probability of a message being lost. The last two parameters are used to model the mobility and reliability of the P2P system over MONET.

We first compare the message overhead of three algorithms. Figure 3.(a) show the experiment results when *LeaveRate* varies from 0 to 0.5. *MsgLossRate* is set to 0.3. PO suffers from the heavy communication traffic due to large number of messages. Note that PO does not filter any duplicate messages. Every node just propagates an update message to all of its neighbors. UPM reduces the number of

Table 1. Simulation Parameters

Parameter	Description	Setting
CPUSpeed	Speed of nodes' CPU	500 Mips
NetBandwidth	Wireless network bandwidth	1 Mbit/s
NumNode	Number of nodes	500
NumNeighbor	Number of neighbor nodes	8
DiskTime	Disk access time	10 ms $\sim$ 30 ms
MsgInst	CPU instructions to process a message	22000
PerIOInst	CPU instructions for a disk I/O	5000
NumDataItem	Number of data items	1000
UpdateRate	Probability of a data item being updated	1.0
OnlineRate	Probability of a node being online	0.2
LeaveRate	Probability of an online node leaving network	0.0 $\sim$ 0.5
MsgLossRate	Probability of a message being lost	0.0 $\sim$ 0.3
MsgDelay	Delay to transfer a message	1 ms $\sim$ 5 ms

messages about 35% compared with PO. This is because UPM attaches a sender list at the update message so that it can reduce the duplicate message delivery significantly. As expected, LRA reduces the number of messages the most. It was about half of PO. The receiver list can prevent the duplicate message delivery completely, even though lost updates would happen.

Then we evaluate the number of lost updates for each algorithm. Figure 3.(b) shows the experiment results by varying *LeaveRate* while *MsgLossRate* is set to either 0.15 or 0.3. PO performs worse as *LeaveRate* increases due to the lack of pull phase. If a node is off-line during a push phase of an update, it should lose the update. On the other hand, the performance of UPM and LRA are nearly constant under the change of *LeaveRate*. They can allow a new joining node to receive any missed updates at the pull phase.

The performance of LRA degrades when *MsgLossRate* is high. The receiver list approach of LRA restricts a node to receive an update message from only

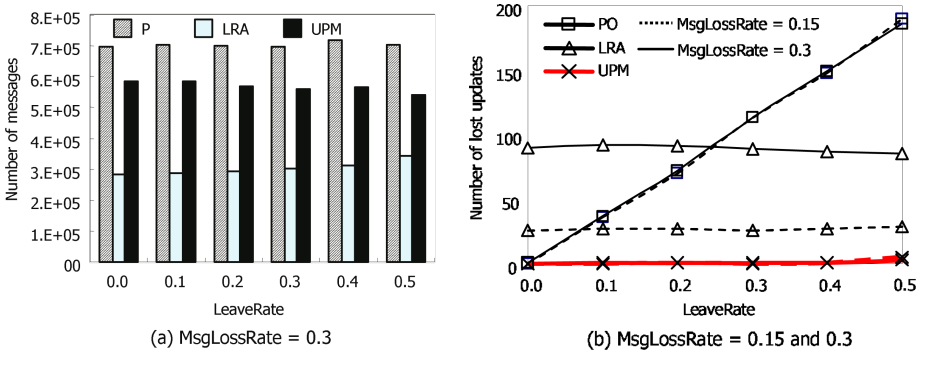


Fig. 3. Experiment Results

one of its neighbor nodes. If the message is lost, the node cannot receive it again from any other nodes. UPM does not cause any lost updates. The duplicate message delivery is beneficial at this setting. As a result, UPM can guarantee safe update propagation at unstable network when messages could be lost and many nodes leave or join the network.

5 Concluding Remarks

In this paper, we describe UPM (*Update Propagation over MONET*), a new update propagation algorithm for P2P file sharing over MONET. UPM is novel in the sense that it is purely decentralized and has lower communication overhead. UPM also reconstructs the neighborship autonomously and thus is resilient to node failures and can achieve fast update propagation.

We have demonstrated the efficacy of UPM using a number of different experiments. In the experiments, UPM propagates most update messages even when messages could be lost or nodes are disconnected frequently. This means that UPM can perform well at P2P systems on highly dynamic network such as MONET. Furthermore, UPM also reduces the amount of duplicate message delivery significantly compared with the push only algorithm. This contributes to the saving of network bandwidth and battery power of mobile nodes.

References

1. Androutsellis-Theotokis, S., Spinellis, D.: A Survey of Peer-to-Peer Content Distribution Technologies. ACM Computing Surveys **36** (2004) 335-371
2. Cetintemel, U., Keleher, P., Bhattacharjee, B., Franklin, M.: Deno: A Decentralized, Peer-to-Peer Object-Replication System for Weakly Connected Environments. IEEE Trans. Computers **52** (2003) 943-959
3. Datta, A., Hauswirth, M., Aberer, K.: Updates in Highly Unreliable, Replicated Peer-to-Peer Systems. Proc. 23rd ICDCS (2003)
4. Ding, G., Bhargava, B.: Peer-to-peer File Sharing over Mobile Ad hoc Networks. Proc. 2nd IEEE Conf. Pervasive Computing and Comm. Workshops (2004)
5. Duran, A., Shen, C.: Mobile Ad hoc P2P File Sharing. Proc. Wireless Comm. and Networking (2004)
6. Gnutella, http://www.gnutelliums.com/
7. Gopalakrishnan, V., Silaghi, B., Bhattacharjee, B., Keleher, P.: Adaptive Replication in Peer-to-Peer Systems. Proc. 24th ICDCS (2004)
8. Holliday, J., Steinke, R., Agrawal, D., Abbadi, A.: Epidemic Algorithms for Replicated Databases. IEEE Trans. Knowledge and Data Eng. **15** (2003) 1218-1238
9. Huang, C-M., Hsu, T-H., Hsu, M-F.: A File Discovery Control Scheme for P2P File Sharing Applications in Wireless Mobile Environments. Proc. 28th Australian Computer Science Conf. (2005)
10. Kazaa, http://www.kazaa.com/.
11. Schwetmann, H.: User's Guide of CSIM18 Simulation Engine. Mesquite Software, Inc. (1996)

P2P Mobile Multimedia Group Conferencing: Combining SIP, SSM and Scalable Adaptive Coding for Heterogeneous Networks*

Thomas C. Schmidt[1], Matthias Wählisch[1,2],
Hans L. Cycon[3], and Mark Palkow[4]

[1] HAW Hamburg, Dep. Informatik, Berliner Tor 7, D–20099 Hamburg, Germany
[2] link-lab, Hönower Str. 35, D–10318 Berlin, Germany
[3] FHTW Berlin, FB I, Allee der Kosmonauten 20–22, D–10315 Berlin, Germany
[4] daViKo GmbH, Am Borsigturm 40, D–13507 Berlin, Germany
{t.schmidt,waehlisch}@ieee.org
hcycon@fhtw-berlin.de, palkow@daviko.com

Abstract. In this paper we present work in progress on extending multimedia conferencing standards to scalable, mobile multimedia group support based on SIP initiated Source Specific IP Multicast. We propose extensions of SIP for negotiating SSM sessions. SIP protocol specifications and semantics are compatibly extended without adding new SIP methods. We will introduce a multimedia communication software with distributed architecture as implementation reference. The software is built on a scalable video codec adaptive to heterogeneous network capacities.

1 Introduction

Today's rapidly rising capabilities of devices and infrastructure, as well as an increasing social focus on ubiquitous infotainment will most likely establish mobile multimedia communication as a day-to-day companion. Real–time video conferencing and group communication thereby are likely to face acceptance not only as a particular, but a common service, raising questions of ease and quality of the underlying Internet service layer. Here IP multicasting will be of particular importance to mobile environments, where users commonly share frequency bands of limited and heterogeneous capacities.

In the present paper we address the issue of mobile multimedia group conferencing, exemplarily taking perspective on a VCoIP (Video Conferencing over IP) software with a distributed, peer–to–peer architecture and its applications [1]. The software is build around a scalable video codec [2] which can serve heterogeneous network capacities. Such lightweight solution should receive support from network layer multicast, restricting service to Source Specific Multicast communication for the sake of deployment simplicity. Source Specific Multicast (SSM) [3], just released as an initial standard, is considered a promising improvement of group distribution techniques.

* Supported by the German BmBF within the Project *Moviecast*.

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 761–764, 2007.
© Springer-Verlag Berlin Heidelberg 2007

However, up until now the session signalling standard SIP [4] is not prepared to negotiate SSM group sessions. We therefore present a straight forward extension of session initiation handshakes, suitable to establish SSM conferencing within an uncoordinated peer–to–peer model. Compliant with standard unicast and ASM transactions, we propose to solely add on protocol semantics without introducing new SIP methods. Our video conferencing system serves as a platform for reference implementation. We further reference to work on session mobility with the special focus on real-time multicast group communication. Conferencing parties request seamless real–time performance of a mobility aware group communication service, thereby attaining the simultaneous roles of mobile multicast listener and source. Up until now only limited work has been dedicated to multicast source mobility, which poses the more delicate problem [5,6].

In this paper we first briefly discuss the mobile multimedia group conferencing problem and related work. In section 3 we present our SIP signalling scheme for SSM group initiation. Section 4 is dedicated to a conclusion and an outlook.

2 The Mobile Multimedia Group Conferencing Problem and Related Work

Multimedia session–based communication requires a number of initial negotiations to be accomplished. At first, a caller requesting contact to one or several partners will expect to address a personal identifier, but establish the corresponding conference session with the devices currently in use by the callees. At second, media and service data need to be identified as to meet capabilities of all session members. Once established, sessions need to persist, while mobile devices roam. Intermediate handovers thereby should unnoticeably comply with quality of service measures for real-time communication.

SIP forms a multi–layered application protocol that interacts between components in a transactional way. Each (asynchronous) request initiates an open transaction state and requires completion by at least one response. Group communication complicates this process significantly. The basic SIP RFC only defines a minimal message exchange for Any Source IP layer multicast: A client wishing to initiate or join into a multiparty conference sends its INVITE request to a multicast group by employing the maddr attribute in the SIP VIA header. Group members subsequently respond to the same group. The transactional nature of SIP is preserved in the sense that the inviting party interprets the first arriving OK as the regular completion, while interpreting further messages as iterates. Consequently, SDP Negotiations on media parameters are displaced out–of–band, since they are not achievable within one transaction. While Any Source Multicast accounts for an open common addressing, SSM presupposes source specific subscriptions. Hence it requires a distribution of newly joining session member's addresses, which otherwise remain unnoticeable to the group.

Group conferencing must be considered a generic term for a wide variety of meanings in different contexts [7]. Three relevant types are essentially considered: In *loosely coupled* conferences, no signalling relationship is maintained

by conference parties, while meta information are pre–shared out of band or gradually learned by RTCP control streams. *Tightly coupled* conferences rely on central managing for all participants. Finally, the *fully distributed*, infrastructureless multiparty model is built upon peer-to-peer signalling. Its scalability can be largely enhanced by the use of multicast. Further on we will concentrate on this fully distributed approach, including SSM at the signalling level.

Without optional proxy servers the SIP protocol architecture draws near the general peer-to-peer model. Recent IETF work aims on extending the SIP framework to include fully distributed infrastructureless approaches [8]. Additional work is needed to develop peer-to-peer group support within a SIP control plane. Keeping in mind the routing complexity inherent to ASM, it is desirable to rigorously restrict all signalling to unicast or SSM communication.

3 SIP Initiated SSM Group Conferences

Instantaneous establishment of a fully distributed peer–to–peer conference commonly follows an incremental setup: Some party will initiate a conference by contacting one or several peers via unicast. Following an initial contact, signalling will then be turned to scalable multicast. Further on new parties will join the conference by either calling or being called by an existing member. Such group initiation scheme is neither covered by the current status of SIP, nor is the employment of Source Specific Multicast for signalling. In order to enable SSM, all dialogs must carefully provision addresses of newly arriving senders to all current group members to allow for appropriate source specific subscriptions.

In detail protocol operations of the suggested extensions proceed as follows. A caller, wishing to establish an SSM signalling session with a single peer, will initiate a regular INVITE request to callee's unicast address. Eventually, after the call setup has completed, either party will decide to turn the established session to group communication. Heading for SSM, it will submit a re–INVITE, i.e., an INVITE carrying the previously established session identifier, announcing its desired multicast group address in the CONTACT field of the SIP header. Concomitantly, the SIP protocol stack will submit a multicast source specific JOIN to its underlying IGMP/MLD stack for subscribing to the group and peer's source address, which it had both learned from the previous SIP message exchange. Any peer will identify the multicast address in the CONTACT field, designed to specify contacts for subsequent requests, and proceed along the new protocol semantic for SSM. As ASM multicast address announcement will distinctively appear in a separate VIA header, it will identify the presence of SSM, answer with a regular unicast reply, but will submit a multicast JOIN to the announced group and caller's source address. This two–step process purposefully decouples application layer session establishment and underlying multicast routing operations. Temporal progress in IP–layer multicast routing and SIP transactional timers thereby remain independent for the sake of robustly layered protocol operations. Appropriate media session descriptions for SSM distribution of media streams may be negotiated along the re–INVITE request.

In multiparty environments, a newly arriving party will contact some member of the established session via regular unicast INVITE. The callee might decide to accept this request and forward it to its partner, thereby initiating unicast sessions among the three. Eventually the callee will decide to select multicast for the conference signalling and will submit the corresponding re–INVITE procedure.

If a new source S_{new} contacts an established SSM group conference, it will do so by inviting some member S. If S decides to accept the caller, it will redistribute its INVITE to the SSM group and acknowledge the initial call by placing the group address in the CONTACT header field. All group members will immediately add S_{new} to their source specific multicast filters. S_{new} subsequently will learn about all group members from (unicast) OK–messages as needed for its own multicast subscriptions. Note that call redistribution will remain a point–to–point request of S_{new} at the application layer, but a transmission of source S at the network layer compliant with previous SSM group establishment.

4 Conclusions and Outlook

In this paper we addressed the essential issues of multimedia group conferences, negotiated by the Session Initiation Protocol and relying on Source Specific Multicast distribution among equal peers. We introduced peer-to-peer signalling extensions of SIP. Proceeding along the proposed incremental way, a callee will never be required to communicate messages to more than one party and one group. This scheme thus remains fully scalable and fairly transparent to group sizes. Multicast initiation of media sessions may be led correspondingly. We also reported on a peer-to-peer reference multimedia communication software with temporal scalable video codec as core module and referred to routing methods to adapt SSM to source mobility. This enables conferencing in networks with heterogeneous capabilities. Implementations for real deployment are under way.

References

1. Palkow, M.: The daViKo homepage (2006) http://www.daviko.com.
2. Schwarz, H., Marpe, D., Wiegand, T.: Overview of the Scalable H.264/MPEG4–AVC Extension. In: IEEE International Conference on Image Processing (ICIP 2006), Atlanta, GA, USA (2006)
3. Holbrook, H., Cain, B.: Source-Specific Multicast for IP. RFC 4607, IETF (2006)
4. Rosenberg, J., Schulzrinne, H., Camarillo, G., Johnston, A., Peterson, J., Sparks, R., Handley, M., Schooler, E.: SIP: Session Initiation Protocol. RFC 3261, (2002)
5. Schmidt, T.C., Wählisch, M.: Multicast Mobility in MIPv6: Problem Statement. IRTF Internet Draft – work in progress 01, MobOpts (2006)
6. Schmidt, T.C., Wählisch, M.: Morphing Distribution Trees – On the Evolution of Multicast States under Mobility and an Adaptive Routing Scheme for Mobile SSM Sources. Telecommunication Systems **33** (2006) 131–154
7. Rosenberg, J.: A Framework for Conferencing with the Session Initiation Protocol (SIP). RFC 4353, IETF (2006)
8. Baset, S., Schulzrinne, H., Shim, E., Dhara, K.: Requirements for SIP–based Peer-to-Peer Internet Telephony. Internet Draft - work in progress 00, IETF (2006)

Federation Based Solution for Peer-to-Peer Network Management

Jilong Wang and Jing Zhang

Network Research Center, Tsinghua University, Beijing 100084, China
wjl@cernet.edu.cn, zhj_021@tom.com

Abstract. Recently, Peer-to-Peer (P2P) technology has become one of the hottest topics in the research area of Internet. With a variety of P2P applications, especially those applications sharing large-size file resources among a large scale of user community, P2P brought us an even more exciting Internet life. However, P2P also made lots of troubles to network managers because of consuming too much network bandwidth sometimes. Lacking of effective management solution, some ISPs plan to block all P2P services on their network boundaries. In this paper, we propose a federation-based solution for peer-to-peer network management. By setting up a P2P federation, ISPs and P2P service providers can work together for P2P network management. From P2P federation service, ISPs can get P2P nodes information of their own network to make some control to those that disturb normal network services. At the same time, service providers of P2P may get routing information of specific ISPs to optimize the routing of P2P network itself by joining in federation. Under such scenario, it will save much cost for ISPs to detect and control P2P traffic and give up the idea of killing P2P services. Save and help to the development of P2P are the most important target of this solution.

Keywords: Peer to Peer, Federation, Management.

1 Introduction

Currently, Peer-to-Peer (P2P) technology is widely applied in Internet application systems, especially those applications sharing large-size video files among a large scale of user community.

P2P changes the traditional Client/Server communication model into a point-to-point model. P2P traffic is no longer converged at a few computers that play as server hosts, but is distributed among each node in the network, which makes the distribution of traffic more reasonable and helps avoid congestion of network.

Although P2P technology has lots of advantages, it brings troubles to network operation sometimes. On account of being frequently used to transfer large size files such as audio and video clips, P2P applications may consume much network bandwidth. Furthermore, P2P application is always apt to use the network bandwidth as much as it possible. In a certain study [2], it has estimated that the total P2P volume goes about 80% of the total downloading traffic in the Internet, and the ratio has been constantly on the rise, nearly twice a year.

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 765–772, 2007.
© Springer-Verlag Berlin Heidelberg 2007

Long-time congestion brings the network management great difficulty. Let's take account of the case of BitTorrent (BT) [1], one P2P application for file sharing. Providing there are numerous clients using BT to download and upload files at the same time, this application will take up lots of network bandwidth and influence other services in the Internet [3]. On the other hand, lack of proper management of P2P applications also causes a series of social problems, relating to copyright, privacy, security and so on. Owing to these referring matters, Internet Service Providers (ISPs) urgently desire to find an effective way to monitor and control the P2P applications. But it is not an easy subject. Current P2P applications tend to disguise their protocol specifications, operate on the random port numbers, and even use the custom port number, such as port 80 of HTTP, intentionally. These characteristics of new-style P2P applications make it even more difficult to monitor and control P2P traffic.

In this paper, we propose a federation-based solution for peer-to-peer network management. Not by measurement, but by exchanging information among ISPs and P2P providers, which is named PSPs under federation model. By setting up a P2P federation, ISPs and PSPs can work together for P2P network management. From P2P federation service, ISPs can get P2P nodes information of their own network to make some control to those that disturb normal network services. Meanwhile, PSPs may get routing information of specific ISPs to optimize the routing of P2P network itself by joining in federation. Under such circumstance, it will save much cost for ISPs to detect and control P2P traffic and give up the idea of killing P2P. Save and help to the development of P2P are the most important target of this solution.

The rest of this paper is structured as follows:Section2 analyzes the previous work in the areas of P2P traffic measurement and management, and then indicates some problems. Section 3 presents the framework of the solution. Section 4 describes some details, including considerations of implementation. Finally, we conclude the paper in Section 5.

2 Previous Works

In order to monitor and manage the P2P applications, we need to distinguish P2P traffic from other network load.

At present, there are basically two ways to measure P2P traffic. One is active measurement, and the other is passive measurement which contains two methods: payload analysis and nonpayload analysis.

2.1 The Active Measurement

In this method we collect P2P nodes information by setting some P2P crawlers. P2P crawlers are some special nodes deployed in the P2P network, connected with other live hosts through TCP. Communicating with known hosts, crawler establishes a peer-list and adds hosts that were freshly found into the list. In this way, crawler host acquires the information of other peers in the P2P network, from which we could know the nodes' (peers') actual distribution and operating status.

However, there are two serious limitations using this methodology. Firstly, it does not adapt to the management of large scale P2P network since this method depends on

the number of connections that crawlers keep. And one significant character of P2P applications is just the enormous user community. Secondly, this method is concerned with specific protocols. Therefore, it is not apt to P2P systems having several types of protocols, especially when some protocols are not open.

2.2 The Passive Measurement—Payload Analysis

Payload analysis of P2P traffic is based on identifying characteristic bit strings in packet payload. It analyses the potential character (usually using the IP packet head and the sixteen bits of the data packets) in the packet payload to identify the type of P2P traffic. E.g. the character sign of BitTorrent protocol is "0x13Bit" [1]. We can distinguish BT traffic by analyzing whether the packet payload contains "0x13Bit". In the same way, eDonkey packets can be identified by "0xe319010000" [7].

The disadvantage of this method is the inefficient application layer identification. Due to the demand of dealing all the data packets, the speed of analysis is slow, which means that it is not adapt to the real-time analysis. In addition, this method is not able to deal with the unknown protocol either.

2.3 The Passive Measurement—NonPayload Analysis

Using nonpayload analysis methodology, we can identify the P2P traffic without inspecting the user payload. The analysis needs only some packet header information such as the connection patterns of source and destination IPs and ports.

From some related references, at present there are six out of nine popular P2P applications using both TCP and UDP as the transport layer's protocols. These P2P applications usually use TCP to transfer actual data and use UDP to control traffic [2]. Due to this characteristic, we can identify P2P hosts by looking for pairs of source-destination hosts that use both transport protocols (TCP and UDP). Whereas the usage of both TCP and UDP is not only for the P2P protocols, it is also used for some other application protocols such as DNS to transport. Therefore, it is hard to distinguish the P2P application from other applications which have the same characteristic, especially in the case of some P2P applications intentionally using the custom port number.

For the descriptions presented above, we strongly commit to the notion that although the management of P2P applications is very important, at the present there is still not an adaptive and effective methodology to analyze so many kinds of P2P applications.

3 Federation Based Solution for P2P Network Management

Because of the diversity and variability of P2P protocols, it is quite hard for ISPs to monitor and manage P2P applications just by active and passive measurement. In order to deal with the matters in the network caused by P2P applications, ISPs attempt several means (e.g. ban the port numbers P2P protocols usually using), and do their best to forbid the P2P applications. At the same time, PSPs create a series of improved protocols and renew-edition client software to confront the forbiddance from ISP, which makes the management of P2P applications more difficult. Thus, both sides of ISP and P2P are plunging into a hostile competition.

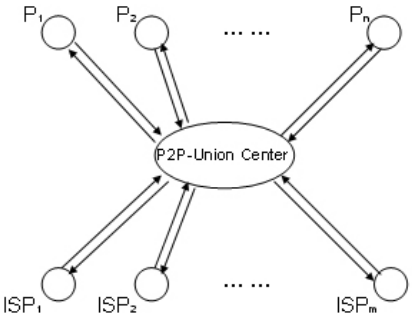


Fig. 1. Architecture design of P2P federation. P1…Pn represent various of P2P service providers, ISP1…ISPn represent Internet Service Providers.

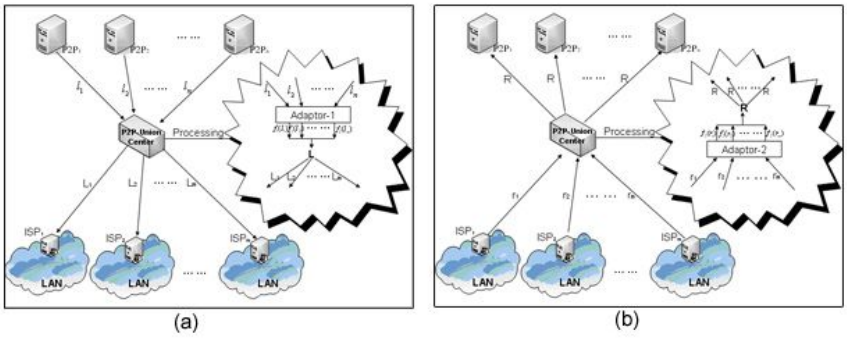


Fig. 2. (a) ISPs learn P2P network information from federation. (b) P2Ps learn ISP's routing information from federation.

To solve above problems, we propose to establish a service called P2P-Federation that helps to exchange P2P client and communication information between ISPs and PSPs. Using this methodology, we can easily deal with the current problems existing in P2P management such as multiple protocols, disguising port numbers and so on.

The basic idea of P2P federation service is cooperation. Instead of the traditional method that made a contrastive relationship between ISPs and PSPs, P2P federation solution aims to found collaboration between ISPs and PSPs.

Commonly, a P2P federation system is composed of three parts. They are ISP units (a series of ISPs), P2P units (a series of P2P providers) and a P2P-Federation Center. Figure 1 shows the architecture design of the P2P federation. Figure 2(a) and (b) show the scenarios of exchanging nodes and routing information between ISPs and PSPs.

The P2P-Federation Center works as a backbone to connect ISPs and P2Ps together, and provides services of information exchange. The federation center gets nodes lists from P2P service providers and provides nodes information to ISPs. On the other hand, the federation center may also gains the IP routing information from

ISPs and provides such information to P2P service providers. The communication among ISP, federation center and PSP should be based on the reliable and secure protocols.

4 The Working Procedure and the Experiment

This section describes how the P2P Federation works in detail. The basic Working Procedure of P2P Federation Service is shown as Figure 3.

Step 0: Register. The P2P-Federation Center provides the register service. The acceptable users contain all kinds of ISPs and P2P service providers. For users whose identity is ISP, the personal information they should submit to register includes: userName, userPwd (that is user's password), userIP (that is the IP address used by ISP's server), IPSection (that is the IP sections that is managed by this ISP), routingFormat (that is the format of the routing table this ISP uses). In the same way, as to users whose identity is PSP (P2P service provider), they should also submit their personal information including userName, userPwd, userIP (that is the IP address that the server providing P2P service uses), p2pType (that is the type of the P2P application, such as Bittorrent, eMule, and so on), p2pFormat (that is the format of the P2P's routing information). Then, the P2P-Federation Center provides each user a unique userID to identify. The P2P-Federation Center maintains a database for preserving the users' identity information.

Step 1: P2P servers that have registered in the P2P-Federation Center will transfer Peers lists (l_i) they maintained to the Center in real-time. The content of Peers list (l_i) is the routing information of all the P2P clients which use the service provided by the corresponding P2P server. The formats of these Peers lists ($l_1, l_2 ....., l_n$) from different P2P servers may also be different. It lies on whether the p2pIDs of the different P2P application are the same.

Step 2: The P2P-Federation Center formats the Peers lists (l_i, i=1,…,n) by an adaptor (signed as Adaptor-1). In this way, all the various Peers lists have been formatted in the same mode, that is (IP, p2pID). We can sign the new lists as $f_1(l_1), f_1(l_2), ..., f_1(l_n)$. The function f_1 () represented the process by Adaptor-1.

Step 3: The P2P-Federation Center then adds up all the new lists ($f_1(l_1), f_1(l_2), ..., f_1(l_n)$) and acquires an entire Peers list which is called L. $L = f_1(l_1) + f_1(l_2) + ... + f_1(l_n)$.

Step 4: According to the IPSections given by ISP users who have registered in the Center, the P2P-Union Center decomposes the L into several child lists which is noted as $L_1, L_2 ....., L_m$.

Step 5: The P2P-Federation Center transfers the child lists L_j (j=1,…,m) to corresponding ISP users in real-time.

As shown in Figure 4, we can easily and directly comprehend the process of how to transfer the P2P routing information to ISPs.

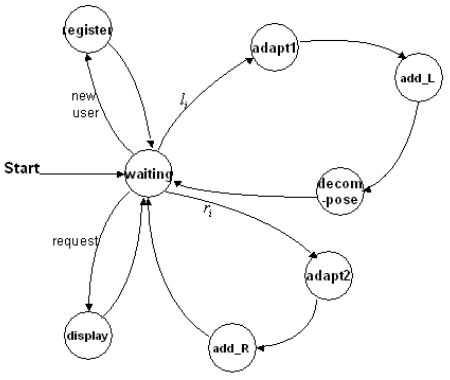


Fig. 3. The Working Procedure of P2P Federation Service

Step6: ISP users that have registered in the P2P-Federation Center will transfer routing tables (r_i) maintained of their own to the Center in time. The content of routing table (r_i) is the routing information of the IPSection managed by the corresponding ISP. The formats of these routing tables ($r_1, r_2 ..... , r_m$) from different ISPs may be different too.

Step 7: The P2P-Federation Center formats the routing tables (r_i, i=1,…,m) by an adaptor (signed as Adaptor-2). In this way, all the various routing tables have been formatted in the same mode, that is (sourceIP, nextRoute, target). We can sign the new lists as $f_2(r_1), f_2(r_2), ..., f_2(r_m)$. The function f_2 () represented the process by Adaptor-2.

Step 8: The P2P-Federation Center then adds up all the new lists ($f_2(r_1), f_2(r_2), ..., f_2(r_m)$) and acquires an entire Peers list which is called R. $R = f_2(r_1) + f_2(r_2) + ... + f_2(r_m)$.

Step 9: The P2P-Federation Center transfer the child lists R to each PSPs in real-time.

We can see the process of how the ISPs transfer the IP routing information to PSPs in the Figure 5.

Step 10: ISP uses the Peers lists to manage P2P traffic. Taking into account of ISP_i , it gets the Peers list L_i from the P2P-Federation Center, in which there is all the P2P clients' routing information (including IP and p2pType). Then ISP_i can confine the hosts whose IP is in the list L_i to a low speed network through infusing specific routing information into the network. Notes that this is just one of possible ways to use the Peers lists.

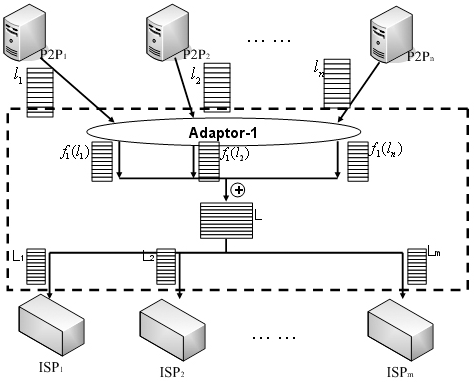


Fig. 4. The Scenario of Step 1 to Step 5

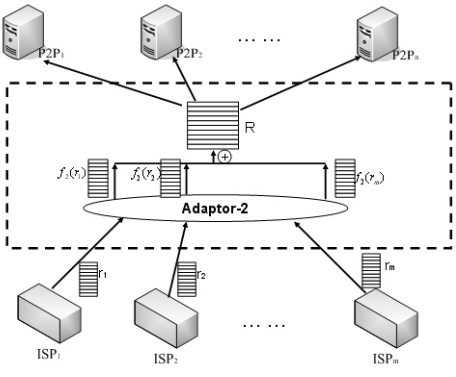


Fig. 5. The Scenario of Step 6 to Step 9

We set up a web server to simulate the P2P-Federation Center which provides some web pages used for registering, logging in and showing the information and maintains a database keeping the peers information. Then we set up two BitTorrent's tracker server to simulate the P2P units, and we use two users registered in the center as ISP units. In this way, the ISP users can acquire the current P2P users' IP and port.

5 Conclusions

This paper discusses P2P management in a new vision. Different from traditional measurement based method, we propose a federation-based solution for P2P management. By exchanging information from a trustable federation center, ISPs can easily acquire P2P peers' information. In this way, we avoid measuring the P2P traffic accurately which is difficult and inextricable to some extent.

At the same time, the P2P service providers who have joined in the federation may get the IP routing information of specific ISPs from the P2P-Federation Center, which help much to improve the quality and efficiency of P2P services.

Therefore, the P2P federation solution described in this paper can achieve a win-win consequence to both ISPs and PSPs. It will advance the P2P technology and improve the management of P2P network.

References

1. Bram Cohen. BitTorrent Protocol Specification v1.0 Identification. http://www.bitconjurer.org/BitTorrent/protocol.html
2. Subhabrata Sen, Jia Wang. Analyzing Peer-to-Peer Traffic Across Large Networks. In: IEEE/ACM TRANSACTIONS ON NETWORKING, VOL. 12, NO.2, (APRIL 2004)
3. Bram Cohen. Incentives Build Robustness in BitTorrent. http://bittorrent.com/bittorrentecon.pdf
4. T. Karagiannis, A. Broido, M. Faloutsos, and K. Claffy. Transport Layer Identification of P2P Traffic. In: Proceedings of 2004 ACM SIGCOMM Internet Measurement Conference, IMC 2004, Taormina, Italy, 2004.
5. T. Hamada, K. Chujo, T. Chujo, and X. Yang. Peer-to-Peer Traffic in Metro Networks: Analysis, Modeling, and Policies. In: Proceedings of IEEE Symposium Record on Network Operations and Management Symposium, Seoul, South Korea, (2004).
6. Ashwin R. Bharambe, Cormac Herley, Venkata N. Padmanabhan. Some Observations on BitTorrent Performance. Proceedings of the 2005 ACM SIGMETRICS international conference on Measurement and modeling of computer systems, (Jun. 2005)
7. eDonkey2000. http://www.edonkey2000.com

A Feedback Based Adaptive Marking Algorithm for Assured Service

Fanjun Su, Chunxue Wu, and Guoqiang Sun

College of Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China
sufanjun@163.com, sufanjun@usst.edu.cn

Abstract. We propose an algorithm to realize proportional bandwidth allocation in DiffServ networks. According to the feedback information, ingress routers adjust Committed Information Rates (CIRs) of different aggregates using AIMD way to let the network reach an exactly subscribed state. Based on CIRs, packets are marked. Simulations with ns2 show that proportional bandwidth allocation under different network conditions can be achieved.

Keywords: DiffServ, Proportional bandwidth allocation, AIMD.

1 Introduction

To provide assured service in DiffServ [1] networks, the markers at ingress routers (e.g. srTCM [2]) mark packets that obey the profile to high priority (e.g. IN), and mark packets that beyond the profile to low priority (e.g. OUT). The queue management mechanism operated at the core routers (e.g. RIO [3]) will preferentially treat high priority packets.

Ideally, proportional bandwidth allocation should be realized between different aggregates. Suppose a bottleneck link whose capacity is C, which is shared by N aggregates. Let CIR_i denote Committed Information Rate (CIR) of aggregate i, and R_i denote the allocated bandwidth, where $1 <= i <= N$. Then we say a network is under-subscribed when $\sum_i CIR_i < C$, over-subscribed when $\sum_i CIR_i > C$, and exactly subscribed when $\sum_i CIR_i = C$. If $R_i = CIR_i * C / \sum_j CIR_j$, we say it is proportional bandwidth allocation. However, theoretic and simulation research in [4, 5, 6] reveal that in DiffServ networks bandwidth are allocated unfairly between different aggregates unless the network is exactly subscribed. Hongjun Su [7] proposed an improved time sliding window based three color marker, but this algorithm can get better fairness only for low to middle level of provision. The proposal in [6] needs the modification of TCP in senders and receivers, and requires the support of ECN mechanism. When different aggregates start at different time, it cannot work well.

Therefore, in this paper a new algorithm named Feedback Based Adaptive Marking algorithm (FBAM) is proposed to achieve proportional bandwidth allocation between different aggregates.

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 773–776, 2007.
© Springer-Verlag Berlin Heidelberg 2007

2 Feedback Based Adaptive Marking Algorithm

FBAM algorithm includes three parts:

(1) Subscription information feedback mechanism

Non-overlapped RIO [3] is wildly used in core routers. The calculation of q_{out} includes IN and OUT packets, while the calculation of q_{in} only includes IN packets. Therefore, only after all OUT packets have been dropped does IN packets begin to be dropped.

We set a variable q_{cong}, which corresponds to an exactly subscribed state. When $q_{in} < q_{cong}$, it means that the network is in an under-subscribed state. When $q_{in} > q_{cong}$, the network is in an over-subscribed state. According to the mechanism of RIO, we set $q_{cong} \cong q_{in}^{\min}$. Feedback module monitors the value of q_{in}. When $q_{in} > q_{cong}$, feedback module will send a two-value feedback information to the ingress routers. We choose one unused bit in the packet head, and we call it subscription-bit. When a packet arrives or leaves the DiffServ domain, edge routers set the value of the subscription-bit to 0. When detecting $q_{in} > q_{cong}$, feedback module located in core routers will set the value of the subscription-bit to 1.

To fully utilize the resource of the network, we use the reverse forward packets to transfer such feedback information. For DiffServ network, the topology of core networks usually is simple, so a static configuration can be set. In addition, most of the traffic in the Internet is based on TCP, so in most cases, there exists reverse ACK packets which can be used to transfer feedback information.

(2) Adjustment of CIRs at Ingress Routers

When a packet coming from core routers arrives, ingress routers will read the value of subscription-bit in the packet head. If the value of the subscription-bit is 0, ingress routers increase CIRs of the aggregates at an interval δ as follows:

$$CIR_{i,t} = CIR_{i,t-1} + \alpha * CIR_{i,0} \tag{1}$$

Where $CIR_{i,0}$ is the original CIR of aggregate flow i and $CIR_{i,t}$ is the CIR of aggregate i at time t after adjustment, and α is the increment factor.

If the value of subscription-bit is 1, the ingress routers decrease CIRs as follows:

$$CIR_{i,t} = (1-\beta)\ CIR_{i,t-1} \tag{2}$$

Where β is the decrement factor. Therefore, we can find the mechanism we mentioned has an idea of AIMD (additive increase multiplicative decrease). The convergence, stability, and fairness of AIMD have been analyzed in [8]. It should be emphasized that different ingress routers implement AIMD adjustment independently without the communication between them.

(3) Packet marking

We can use a rate meter such as TSW [2] to measure the sending rate of aggregate flows. Let avg_rate_i denotes the average rate of aggregate flow i. When $avg_rate_i <= CIR_i$, all packets are marked as IN. When $avg_rate_i > CIR_i$, let $P_0 = CIR_i / avg_rate_i$, and $P_1 = 1 - P_0$. Packets are marked as IN with probability P_0, and OUT with probability P_1. It can be inferred that such marking policy ensures that throughput of IN packets of the aggregate flow i equals CIR_i approximately.

3 Experimental Evaluations

(1) Simulation Topology and Configuration

We adopt ns2 [9] as the simulation tool. As shown in Fig.1, IE1, IE2, IE3 are ingress edge routers, and EE1, EE2, EE3 are egress edge routers. R1 and R2 are core routers. C_1 is the bandwidth of the bottleneck link, and C_2 is the bandwidth of link IE3-R1. Every aggregate flow contains 9 TCP connections and 1 UDP connection respectively. The rate of UDP is 0.1* $CIR_{i,0}$. The edge router queues employ drop-tail policy, and core router queues are managed by non-overlapping RIO algorithm. The parameters of RIO are set as follows: $(q_{out}^{min}, q_{out}^{max}, p_{out}^{max})$ =(20, 40, 0.1), $(q_{in}^{min}, q_{in}^{max}, p_{in}^{max})$ = (40, 80, 0.02), q_{cong} = 42. The CIR adjustment parameters are set as follows: α = 0.002, β = 0.002, and δ =20ms. In the simulation, we use TSW [2] as rate meter. The Win_length of TSW is set to 0.1 second.

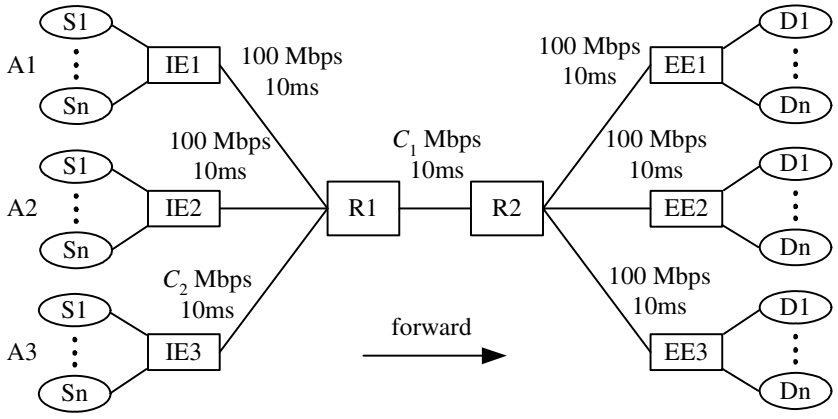


Fig. 1. Simulation topology and configuration

(2) Simulation results

Our first simulations are used to valid the effect of adjustment of CIR using AIMD way. Aggregates A1, A2, A3 are used. We set C_2 =100Mbps. The simulations are carried out in both the under-subscribed case and the over-subscribed case. Here we only list one result in over-subscribed case. $\{CIR_{i,0}\}$ is set to (2, 5, 10) Mbps, and C_1 = 10Mbps. Therefore, CIRs should be adjusted to (1.18, 2.94, 5.88) Mbps, which corresponds to an exactly subscribed state. As shown in Fig. 2 (1), after adjustment, CIRs of different aggregates oscillate around ideal values, so proportional bandwidth allocation has been achieved as shown in Fig. 2 (2). In under-subscribed case, we get the same conclusions from the simulations.

We also carry out other simulations, such as the simulation in the case of aggregates start or close at different time and the case of aggregates have different bottleneck link. All the simulation results show that our algorithm can work well in different cases.

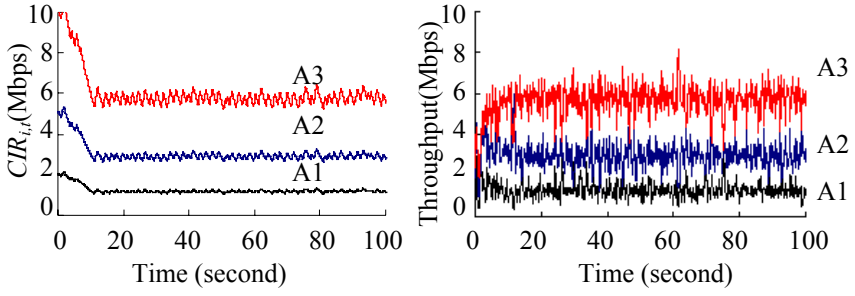


Fig. 2. Adjustment of CIRs and the bandwidth allocation in over-subscribed case

4 Conclusions and the Future Work

FBAM algorithm can detect the subscription state of the network, and send feedback information via reverse forward packets. According to the feedback information, ingress routers will adjust CIRs using AIMD way to reach an exactly subscripted state. Based on the adjusted CIR, FBAM marks the packets. Simulation results show that our algorithm can realize proportional bandwidth allocation under different network conditions. FBAM algorithm doesn't need communication between different ingress routers. Feedback information doesn't require special message packets. More simulations are needed for the optimal choice of parameters, such as α, β.

References

1. Blake, S., Black, D., Carlson, M., et al: Architecture for differentiated services. RFC2475 (1998)
2. Heinanen, J., Guerin, R.: A single rate three color marker. IETF RFC2697 (1999)
3. Clark, D., Fang, W.: Explicit allocation of best effort packet delivery service. ACM Transactions on Networking (1998) 6(4): 362-373
4. Seddigh, N., Nandy, B., Pieda, P.: Study of TCP and UDP interaction for the AF PHB. http://www.watersprings.org/pub/id/draft-nsbnpp-diffserv-udptcpaf-00.txt (1999)
5. Seddigh, N., Nandy, B., Pieda, P.: Bandwidth assurance issues for TCP flows in a differentiated services network. Proceedings of Globecom, Rio De Janeiro (1999) 3: 1792 - 1798
6. Park, E.C., Ho, C.: Proportional bandwidth allocation in DiffServ networks. IEEE INFOCOM (2004)
7. Su, H., Atiquzzaman, M.: ItswTCM: A new aggregate marker to improve fairness in DiffServ. IEEE GLOBECOM (2001) 25-29
8. Kelly, F., Maulloo, A., Tan, D.: Rate control for communication networks: shadow prices proportional fairness and stability. Journal of the Operational Research Society (1998) 49(3): 237-252
9. Network simulator-ns2. http://www.isi.edu/nsnam/ns

QoS-Aware MAP Selection Scheme Based on Average Handover Delay for Multimedia Services in Multi-level HMIPv6 Networks

Y.-X. Lei and Z.-M. Zeng

Beijing University of Posts and Telecommunications
leiyixuebupt@hotmail.com

Abstract. In multi-level HMIPv6 (M-HMIPv6) networks the MN has to select a proper MAP using MAP selection schemes when migrating to the visited ARs. Currently, distance-based MAP Selection (DMS), Mobility-based MAP Selection (MMS) and Adaptive MAP Selection (AMS) schemes have been proposed, and AMS outperforms the other ones in several aspects. However, none of these schemes consider the handover Quality-of-Service (QoS) of mobile multimedia services as the metric for MAP selection. In this paper, we propose a novel QoS-aware MAP selection (QMS) scheme. In this scheme, the optimal MAP is determined according to the average handover delay (AHD) which characterizes the handover QoS for mobile multimedia services. We derive a lemma through theoretical analysis to estimate the AHD and design the procedure of QMS scheme based on this lemma. Through extensive NS-2 simulations, we verify the effectiveness of the proposed scheme and its advantage in AHD than AMS.

Keywords: Hierarchical Mobile IPv6 (HMIPv6), Mobility Anchor Point (MAP), handover delay, Quality of Service (QoS).

1 Introduction

Due to the limitation of Mobile IPv6 [1] for micro mobility management, Hierarchical Mobile IPv6 (HMIPv6) [2] [3] [10] was proposed by IETF to reduce the signaling cost and handover delay by adopting a Mobility Anchor Point (MAP) which can handle the localized mobility [3]. But in HMIPv6 networks with single MAP, the capacity and reliability of the networks need to be improved [5] [6] [10]. To meet this requirement, Multi-level HMIPv6 (M-HMIPv6) networks [5-9] were proposed inside which there are multiple MAPs. And, these MAPs forms different levels according to the number of hops from them to the Access Router (AR). In such circumstances, there exist more than one MAPs above the ARs, and it is essential for the MN to select an optimal MAP when it visits the ARs using MAP selection schemes. Several schemes have been proposed for MAP selection in M-HMIPv6 networks such as Distance-based MAP Selection (DMS) [2], Mobility-based MAP Selection (MMS) [8] [9], Adaptive MAP Selection (AMS) [6], mobile history-based MAP selection [12] and abstraction node based MAP selection schemes [7], etc.

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 777–784, 2007.
© Springer-Verlag Berlin Heidelberg 2007

All the existing MAP selection schemes adopt different metrics including distance, mobile history, load balance, location update cost and packet delivery cost, etc. However, they don't consider QoS of mobile multimedia services, such as handover delay, as the metric for MAP selection [16]. We believe that handover QoS of mobile multimedia services are crucial issues that should be considered during MAP selection because the handover performance is strongly dependant upon the selection of MAPs.

In this paper, we propose a novel QoS-aware MAP selection scheme (QMS) which determines the optimal MAP based on average handover delay (AHD). We analyze the handover process theoretically, and derive a lemma to characterize the impact of MAP selection on AHD. Then, we conduct NS-2 simulations to verify our design. The results show that the analyses and the simulations are consistent. Finally, we compare our scheme with AMS, which is the most convincing scheme in the literature. The results show our scheme outperforms AMS in AHD.

The rest of the paper is organized as follows. Section 2 reviews the architecture of M-HMIPv6 networks, and gives a lemma to estimate the AHD in M-HMIPv6 networks which is the key design for QMS scheme. Then, the procedure of QMS scheme is given. In Section 3, numerical analyses and NS-2 simulations are conducted to verify our design. In Section 4, the performance of QMS and AMS is compared through NS-2 simulations. And, the conclusions are given in Section 5.

2 QMS Scheme

2.1 Handover Process Analyses

Fig. 1 shows the architecture of M-HMIPv6 networks. We can observe that there are multiple MAPs locating at different levels according to the number of hops from them

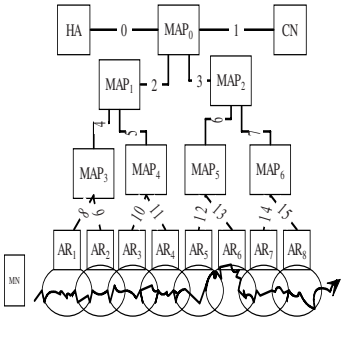


Fig. 1. An M-HMIPv6 network

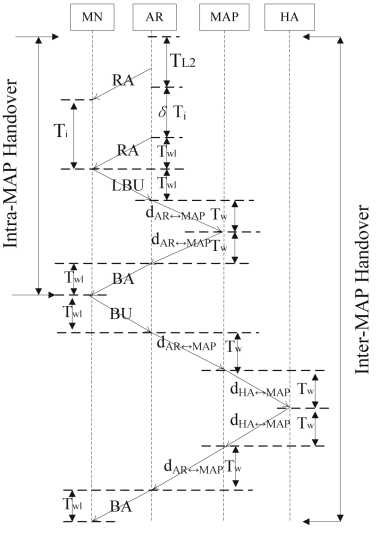


Fig. 2. Inter-MAP and Intra-MAP Handover Procedure

to the ARs. For example, MAP_0 is the Level-1 MAP and there are three hops from it to an AR, e.g., AR_1. MAP_1 and MAP_2 are Level-2 MAPs since there are two hops away from AR_1. To select a proper MAP for the MN, dynamic MAP discovery process [3] is conducted periodically. With this process, the information of available MAPs is collected and delivered to the ARs. When the MN migrates to an AR, e.g., AR_2, it receives the Router Advertisement (RA) [1] [2] messages which contain the MAP options regarding to MAP_0, MAP_1 and MAP_3. Then, an MN selects an MAP with certain MAP selection schemes according to the information within these MAP options. The key design of QMS scheme is Lemma I, which gives a lemma for estimating the AHD. The AHD is determined by two types of handovers, i.e., inter-MAP handover and intra-MAP handover which are shown in Fig. 2. Some abbreviations and definitions are listed in Table 1 to facilitate the descriptions in the following parts of the paper.

Table 1. ABBREVIATION AND DEFINITIONS

CN	Corespondent Node
(L)BU	(Local) Binding Update
BA	Acknowledgement for the BU
T_{L2}	layer two handover delay
PDF	Probability Density Function
T_i	the interval of Router Advertisement of the ARs
$d_{A\leftrightarrow B}$	The number of hops in wired link connecting node A and B
T_{wl}	the transmission delay in wireless link
T_w	the one hop transmission delay in wired link
p_a	the probability of MN to perform intra-MAP handover
p_m	the probability of MN to perform inter-MAP handover
d_1	intra-MAP handover delay
d_2	inter-MAP handover delay
δ	a positive variable between 0 and 1 which follows uniform distribution
t_s	inter session time of the MN
t_A	AR residence time
t_M	MAP residence time
λ_s	session arrival rate
k_A	The shape parameter of probability density function (PDF) for t_A
λ_A	AR crossing rate
k_M	The shape parameter of PDF for t_M
λ_M	MAP crossing rate

2.2 The Lemma for AHD Estimation

Lemma I: Given an M-HMIPv6 network like the one shown in Fig. 1 and handover process in Fig. 2, assume (1) the session arrival of the MN follows Poisson distribution, (2) the AR residence time and MAP residence time follow Gamma distribution with the parameters listed in Table 1, the AHDs of the MAP can be estimated as (1). In (1), ρ stands for the ratio of λ_s and λ_A which is coined Session-Mobility-Ratio (SMR) [7], and n is the number of ARs covered by an MAP.

$$AHD=(\frac{k_A}{K_A+\rho})^{k_A}(T_{L2}+3T_{wl}+\frac{1}{2}T_i+2T_w d_{AR\leftrightarrow MAP})+(\frac{k_M}{k_M+\rho\sqrt{n}})^{k_M}(T_{L2}+5T_{wl}+\frac{1}{2}T_i+4T_w d_{AR\leftrightarrow MAP}+2T_w d_{MAP\leftrightarrow HA}) \quad (1)$$

We now prove Lemma I as follows. The PDFs of t_A and t_M are given in (2) and (3) respectively.

$$f_{t_A}(t)=\frac{(k_A\lambda_A)^{k_A}t^{k_A-1}}{\Gamma(k_A)}e^{-k_A\lambda_A t} \quad (2)$$

$$f_{t_M}(t) = \frac{(k_M \lambda_M)^{k_M} t^{k_M - 1}}{\Gamma(k_M)} e^{-k_M \lambda_M t} \tag{3}$$

According to Fig. 2, we derive d_1 and d_2 as follows:

$$d_1 = T_{L2} + 3T_{wl} + \delta T_i + 2T_w d_{AR \leftrightarrow MAP} \tag{4}$$

$$d_2 = T_{L2} + 5T_{wl} + \delta T_i + 4T_w d_{AR \leftrightarrow MAP} + 2T_w d_{MAP \leftrightarrow HA} \tag{5}$$

According to (4) and (5), we obtain AHD as (6).

$$AHD = E(p_a d_1 + p_m d_2) \tag{6}$$

In (6), we can assume that δ is independent with p_a and p_m, then we rewrite (6) as (7).

$$AHD = p_a(T_{L2} + 3T_{wl} + \frac{1}{2}T_i + 2T_w d_{AR \leftrightarrow MAP}) + p_m(T_{L2} + 5T_{wl} + \frac{1}{2}T_i + 4T_w d_{AR \leftrightarrow MAP} + 2T_w d_{MAP \leftrightarrow HA}) \tag{7}$$

According to (2) and (3), we deduce p_a in (8).

$$\begin{aligned} P_a &= \int_0^\infty P(t > t_A) f_{t_A}(t) dt = \int_0^\infty e^{-\lambda_s t} \frac{(k_A \lambda_M)^{k_A} t^{k_A - 1}}{\Gamma(k_A)} e^{-k_A \lambda_M t} dt \\ &= (\frac{k_A \lambda_A}{k_A \lambda_A + \lambda_S})^{k_A} \int_0^\infty \frac{(k_A \lambda_A + \lambda_s)^{k_A} t^{k_A - 1}}{\Gamma(k_A)} e^{-(k_A \lambda_A + \lambda_s) t} dt = (\frac{k_A \lambda_A}{K_A \lambda_A + \lambda_S})^{k_A} = (\frac{k_A}{K_A + \frac{\lambda_S}{\lambda_A}})^{k_A} \end{aligned} \tag{8}$$

Similarly, p_m can be obtained as (9).

$$P_m = (\frac{k_M}{k_M + \lambda_s / \lambda_M})^{k_M} \tag{9}$$

Then, take (8) and (9) into (7) and we get (10):

$$AHD = (\frac{k_A}{K_A + \frac{\lambda_S}{\lambda_A}})^{k_A}(T_{L2} + 3T_{wl} + \frac{1}{2}T_i + 2T_w d_{AR \leftrightarrow MAP}) + (\frac{k_M}{k_M + \frac{\lambda_S}{\lambda_A}})^{k_M}(T_{L2} + 5T_{wl} + \frac{1}{2}T_i + 4T_w d_{AR \leftrightarrow MAP} + 2T_w d_{MAP \leftrightarrow HA}) \tag{10}$$

According to [6], we can assume $\lambda_M = \frac{\lambda_A}{\sqrt{n}}$ and $k_M = k_A$. Then 10 can be rewritten as (11).

$$AHD = (\frac{k_A}{K_A + \rho})^{k_A}(T_{L2} + 3T_{wl} + \frac{1}{2}T_i + 2T_w d_{AR \leftrightarrow MAP}) + (\frac{k_M}{k_M + \rho\sqrt{n}})^{k_M}(T_{L2} + 5T_{wl} + \frac{1}{2}T_i + 4T_w d_{AR \leftrightarrow MAP} + 2T_w d_{MAP \leftrightarrow HA}) \tag{11}$$

Up to now, we have proved Lemma I. Base on this lemma, we designed the procedure of QMS scheme as follows.

2.3 Procedure of QMS Scheme

The procedure of QMS scheme is as follows. (1) To obtain the MAP options. When an MN migrates to an AR, it obtains the MAPs above it from the MAP options within

the RA messages. (2) To estimate SMR. The MN estimates SMR with the approach in [7]. (3) To extract the distances and the number of ARs an MAP covers. The MN obtains the hop number between AR and the available MAPs ($d_{AR\leftrightarrow MAP}$), the distance from home agent (HA) to MAP ($d_{HA\leftrightarrow MAP}$) and the number of ARs from the MAP option. To achieve this goal, the format of MAP option should be extended to contain these fields. The hops between HA and MAP can be obtained from routing protocols, and the number of ARs within one MAP can be obtained by configuration. (4) To estimate the AHD of each level MAP and select the optimal MAP. The AHDs of the MAPs in different levels are calculated according to Lemma I and the MAP which achieves the minimal AHD is selected by the MN. (5) Adaptive load control. If the optimal MAP is overloaded, then choose the suboptimal one until no MAP is available. In case of no available MAPs exist, the MN should register to the HA.

3 Numerical Analyses and NS-2 Simulations for QMS Scheme

In this section, we obtain some numerical results based on Lemma I, and conduct NS-2 simulations to verify the effectiveness of QMS scheme. The topology we used is shown in Fig. 1.

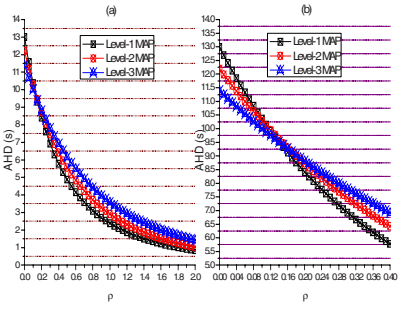


Fig. 3. AHD of different Levels of MAPs

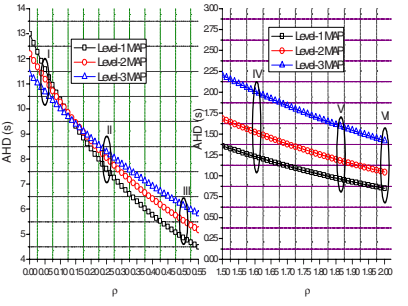


Fig. 4. AHD of Sampled ρ

3.1 Numerical Results

The parameters for numerical analysis are shown in Table 2.

Table 2. PARAMETERS FOR NUMERICAL ANALYSIS

T_{L2}	T_{wl}	$T_{w\ (Link\ 2\sim15)}$	$T_{w\ (Link\ 0\sim1)}$	$k_A(k_M)$	T_i	$d_{HA\text{-}MAP0}$
0.2s	0.5s	0.2s	2s	5	1s	10

In (11), $d_{HA\leftrightarrow MAP}$ and $d_{AR\leftrightarrow MAP}$ are determined by which MAP is selected by the MN. We denote the mean MAP crossing rate of three levels of MAPs as $\lambda^1_M \sim \lambda^3_M$. According to [7], we obtain (12).

$$\lambda_M^{1} = \frac{\lambda_M^{2}}{\sqrt{2}} = \frac{\lambda_M^{3}}{2} = \frac{\lambda_A}{2\sqrt{2}} \tag{12}$$

We plot AHD of different levels of MAPs in Fig. 3 (a) and (b). Fig. 3 (a) is the whole graph, and Fig. 3 (b) enlarges the part with ρ ranging from 0 to 0.4. We observe that the optimal AHD is obtained by different levels of MAPs with the growing of ρ. For example, when ρ equals 0.0125, level-3 MAPs achieve the lowest AHD. And, when ρ is 0.5, level-1 obtains the optimal AHD.

3.2 NS-2 Simulations

The simulation runs with NS-2.1b6a [15] installed in Redhat Linux 9.0. Robert Hsieh [4] previously contributed an extension to NS-2 concerning the Fast Handover for Hierarchical Mobile IPv6 (F-HMIPv6) [13] without implementing the MAP discovery and MAP selection functions in HMIPv6 protocol [2]. We have added the missed functions according the latest protocol standard [2], and then implement MAP selection schemes for performance evaluation. To verify the effectiveness of QMS, we adopt the same network topology which is used for numerical analysis. The coordinates of the 8 ARs are as follows: (250,250), (375,250), (500,250), (625,250), (250,375), (375,375), (500,375), (625,375). The radius of the circular area covered by each AR is 75 m. According to the radius, we set the receive threshold as 3.41828e-8 w (-49.466 dBm) and the transmission power of AR as 0.28183815 w. To approximate the numerical analysis in the previous section, we set the delay of the wired links as Table 2. And, the bandwidth of all the wired links is 100 Mbps.

We adopt CBR traffics based on UDP. The CBR traffic is unidirectional and originated from CN to MN with the bandwidth and packet size as 128 kbps and 512 bytes respectively. We obtain the AHD from UDP traffic which is defined as the interval from the instant of receiving the last packet at previous AR to the instant of receiving the first packet at new AR. The call arrival is follows Poisson distribution and the mobility of the MN simulated by Random Waypoint model [12] generated by a Java-based scenario program [14]. The simulation time is 10 hours, and the results are obtained by averaging the statistics over 10 runs.

Table 3. SIX SCENARIOS FOR SIMULATIONS

Scenario Index	speed range	call arrival rate	samples of ρ
I	40-60m/s	0.0375	0.05
II	40-60m/s	0.1875	0.25
III	5-15m/s	1.8750	0.5
IV	5-15m/s	6.0	1.6
V	5-15m/s	6.9375	1.85
VI	5-15m/s	7.3125	2.00

During the simulation, the MN chooses different level of MAPs in each scenario, shown in Table 3, and each scenario refers to a sample of ρ. The numerical results of the samples of ρ are plotted in Fig. 4, and the simulation results of the samples of ρ are provided in Fig. 5. From Fig. 4 and Fig. 5, we observe that the orders of AHDs incurred by different MAP level from numerical analysis are consistent with the orders in simulation results for all the scenarios. For example, when ρ equals 0.50, $AHD_1 < AHD_2 < AHD_3$ holds in both simulations and analyses results.

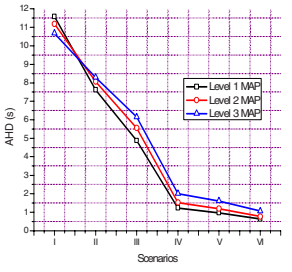


Fig. 5. AHD of Simulations

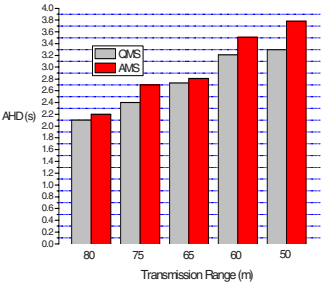


Fig. 6. Comparison of AHD between AMS and QMS

4 Comparison with AMS Through NS-2 Simulations

Since AMS [6] is the most convincing scheme in the literature, we compare the performance of QMS with AMS in AHD through NS-2 simulations. The topology and traffic scenarios are the same with that in Section 3.2.

We simulated 120 MNs with different session arrival rates and mobility patterns. The capacity of the MAPs, i.e., the number of concurrent MNs that can be served by an MAP, is larger than the number of MNs we simulated. And the AHD in this section is the mean of handover delay for the sessions of all the MNs, not for only one MN as defined in previous sections. The simulation results are shown in Fig. 6. We change the transmission range of the ARs in different scenarios, and the results show that QMS can effectively reduce the AHD for all the senarios compared with AMS. The main reason is that it considers the handover delay directly as the criteria for MAP selection.

5 Conclusion

This paper proposes a novel QMS scheme in M-HMIPv6 networks. Firstly, we analyze the handover process theoretically and derive a lemma to estimate AHD, which is the key design of QMS scheme. Then, we verify the analyses by comparing the numerical results with the statistics from NS-2 simulations. Moreover, the performance of QMS is compared with AMS, which is considered as the most convincing MAP selection scheme in the literature. The results show that QMS outperforms AMS in AHD. The proposed scheme can be applied for handover optimization for multimedia services in M-HMIPv6 networks. In future works, we will extend the theoretical analyses to cover more generic mobility models, and conduct more comprehensive simulations to evaluate the performance of QMS scheme on other factors, such as location update cost and packet delivery cost. We will also study how to apply QMS scheme in F-HMIPv6 [13] networks to achieve better handover performance for multimedia services in M-HMIPv6 networks.

Acknowledgments. This research is supported by the Education doctoral special fund, No. 20040013010.

References

1. D. Johnson, C. Perkins and J. Arkko: Mobility Support in IPv6, IETF RFC 3775, (June 2004)
2. H Soliman, C. Castelluccia, K. El Malki and L. Bellier, "Hierarchical Mobile IPv6 Mobility Management," IETF RFC 4140, (August 2005)
3. M. H. Habaebi: Macro/micro-mobility Fast Handover in Hierarchical Mobile IPv6, Computer Communications, Vol. 29, (2006), 611-617
4. Robert Hsieh, Aruna Seneviratne, Hesham Soliman and Karim El-Malki: Performance Analysis on Hierarchical Mobile IPv6 with Fast-handoff over End-to-end TCP, Proc. of EEE Globecom 2002, Vol. 21, No. 1, (November 2002), 2500-2504
5. Kawano K., Kinoshita K. and Murakami K., Multilevel Hierarchical Mobility Management Scheme in Complicated Structured Networks, Proc. of 29th Annual IEEE International Conference on Local Computer Networks, (November 2004), 34-41
6. Sangheon Pack, Minji Nam, Taeyoung Kwon and Yanghee Choi, An Adaptive Mobility Anchor Point Selection Scheme in Hierarchical Mobile IPv6 Networks, Elsevier Computer Communications, Vol. 29, No. 16, (2006), 3065-3078
7. Takaashi Kumagai, Takuya Asaka and Tatsuro Takahashi: "Location Management Using Mobile History for Hierarchical Mobile IPv6 Networks," Proc. of IEEE Globecom 2004, Vol 3, (December 2004), 1585-1589
8. E. Natalizio, A. Scicchitano and S. Marano: Mobility Anchor Point Selection Based on User Mobility in HMIPv6 Integrated with Fast Handover Mechanism, Proc. of IEEE WCNC 2005, Vol. 3, (March 2005), 1434-1439
9. Taleb. T, Suzuki T., Nei Kato and Nemoto Y.: A Dynamic and Efficient MAP Selection Scheme for Mobile IPv6 Networks, Proc. of IEEE Globecom 2005, Vol. 5, (December 2005), 2891-2895
10. Youngjunce Gwon, James Kempf and Alper Yegin: Scalability and Robustness Analysis of Mobile IPv6, Fast Mobile IPv6, Hierarchical Mobile IPv6 and Hybrid IPv6 Mobility Protocols Using a Large-scale Simulation, Proc. of IEEE ICC 2004, Vol. 7, (June 2004), 4087-4091
11. Y.-W. Chen and M.-J. Huang: A Novel MAP Selection Scheme by Using Abstraction Node in Hierarchical MIPv6," to be pubished in Proc. of IEEE ICC 2006, (2006)
12. T. Camp, J. Boleng, and V. Davies: A Survey of Mobility Models for Ad Hoc Network Research, Wireless Communication & Mobile Computing (WCMC): Special Issue on Mobile Ad Hoc Networking: Research, Trends and Applications, Vol. 2, No. 5, (2002), 483-502
13. HeeYoung Jung, Gesham Soliman, Seok Joo Koh and Jae Yong Lee: Fast Handover for Hierarchical MIPv6, IETF Internet Draft, (2005)
14. http://web.informatik.uni-bonn.de/IV/Mitarbeiter/dewaal/BonnMotion/
15. http://www.isi.edu/nsnam/ns
16. Yi-Xue Lei, et. al.: Impact of MAP Selection on Handover Performance for Multimedia Services in Multi-level HMIPv6 Networks, to be published in Proc. of IEEE WCNC 2007, (2007)

On Composite Service Optimization Across Distributed QoS Registries*

Fei Li, Fangchun Yang, Kai Shuang, and Sen Su

State Key Lab. of Networking and Switching, Beijing University of Posts and Telecommunications
187#,10 Xi Tu Cheng Rd.,Beijing,100876, P.R. China
pathos.lf@gmail.com, {shuangk,fcyang,susen}@bupt.edu.cn

Abstract. Web service composition is a promising technology to effectively integrate distributed autonomous services in service oriented paradigm. When providing composite services, ensuring user experienced QoS (Quality of Service) in dynamic environment poses a great challenge. In this paper, we present a distributed service selection approach for optimizing composite service with complex structures. The approach does not require a centralized QoS registry to have complete composition logic, but runs on our distributed QoS architecture iteratively. Experimental results show that the approach is efficient and effective for our problem.

1 Introduction

Web service, as an implementation of service oriented architecture, is gaining more and more acceptance in both academia and industries. In web service framework, a set of XML(eXtensible Markup Language) based standards[1], greatly improves the interoperability of business applications. Services (In this paper, we use the term service and web service interchangeably) could publish their functional and non-functional attributes. Service users could automatically discover services. Service provider could integrate other provider's services to fulfill complex user requirements. The integration process is known as *service composition* and the integrated service is *composite service*.

Automatic service composition is a hot topic in web service research which aims easily reusing of business applications and fast provisioning of new services. S. Dustar and W. Schreiner[2] surveyed some of the important works on different aspects of service composition. Most of the early works are function related, as automatic generation of composition logic, or coordination of composite service. But user requirements are not only functional, also non-functional. Thus the "QoS driven service composition" problem was presented afterwards[3]. As far as we know, all

* This work is supported by the National Basic Researchand Development Program (973 program) of China under Grant No.2003CB314806; the Program for New Century Excellent Talents in University (No:NCET-05-0114); the Program for Changjiang Scholars and Innovative Research Team in University (PCSIRT); the Hi-Tech Research and Development Program (863 Program) of China under Grant No.2006AA01Z164.

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 785–792, 2007.
© Springer-Verlag Berlin Heidelberg 2007

published works was focusing on improving the selection algorithm performance and can only execute on a centralized entity called QoS center or registry. However, with the growing deployment of service oriented applications, centralized architectures may not satisfy the requirements of scalability and flexibility, and suffer from single-point failure. More importantly, for the requirements of global business, each centralized entity could only serve for a specific business region like a corporation or an organization, but not support global scale B2B(business-to-business) applications.

In this paper, we propose a distributed service selection approach based on our distributed QoS registry architecture[4]. The distributed QoS registry is a limited number of QoS registries which could communicate and cooperate with each other. Each one of them maintains QoS information of at least one set of functional identical services. These services could be selected based on their real-time QoS to optimize composite service in a dynamic environment. Figure 1 illustrates the scenario of a composite service logic and related distributed QoS registries. Every Ti is a set of services which could accomplish a specific task.

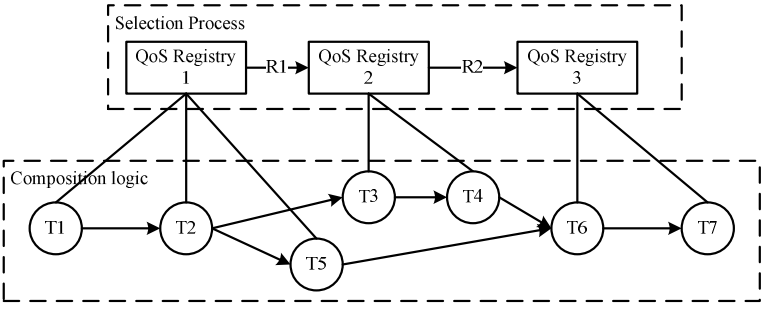


Fig. 1. Distributed QoS registries and a composite service

Our approach to solve the distributed service selection problem has two steps: first, we use an algorithm which could iteratively select services to optimize sequence structure. The algorithm is improved from our previous algorithm by using a more mature heuristic approach. Second, we divide a composite service to several connected sequence structures. The selection results for these structures are aggregated iteratively to optimize the whole composite service.

The rest of the paper is organized as follows: Section 2 reviews some related works. Section 3 describes the distributed selection algorithm for sequence structure in detail. Section 4 presents the approach to apply the algorithm on complex composite structure in distributed environment. Section 5 discusses some experimental results. Finally, the paper concludes in Section 6.

2 Related Works

Some pioneer works have been done on distributed service architecture. Distributed orchestration of web services has been discussed in Benatallah and Sheng et al.[5] for

the first time. They designed the SELF-SERV (compoSing wEb accessibLe inFormation & buSiness sERVices) platform for dynamic and peer-to-peer provisioning of web services. Composite service execution on the platform was in peer-to-peer manner without centralized coordinator. Chafle and Nanda et al.[6] researched some critical problems in automating decentralized execution of composite service. They provided a method for automatic BPEL code partition deriving from program partition method for multiprocessor execution.

A lot of works have been done on QoS based service selection for composite web service. Zeng et al.[3] has presented a basic QoS model for service composition which solved by Integer Programming. For the problem is NP-hard[7], many heuristic approaches have been carried out to improve efficiency. For example: Canfora et al.[7] proposed a genetic algorithm approach to optimize the selection process; Yu et al.[8] model the problem as a Multi-choice, Multi-dimension 0-1 Knapsack Problem (MMKP) and used a modified HEU algorithm to solve it.

In our previous work[4], we have presented a distributed QoS registry architecture and a QoS model with network condition. Based on our architecture and QoS model, we designed a distributed heuristic algorithm to optimize critical task path.

3 Service Selection Algorithm for Sequence Structure

For service selection problem in centralized environment, where QoS information of all candidate services is stored in one registry, the selection processes are not different for sequence structure and other structures. But in decentralized environment, we have to deal with them respectively because no registry has the whole composition logic. This section describes the selection algorithm for sequence structure.

3.1 Problem Definition

A sequence structure with *l* tasks is $p = \langle t_1, t_2, ..., t_l \rangle$, where t_i is the *i*th task in topological order. Each task has a set of candidate services, $S(t) = \{s_1, s_2, ..., s_m\}$, one of them will be selected for corresponding task. Each service *s* has a set of QoS parameters $Q(s) = \langle q_1, q_2, ..., q_n \rangle$ monitored by QoS registry. The composite service *cs* also has a set of QoS parameters $Q(cs) = \langle q_1^c, q_2^c, ..., q_n^c \rangle$, where $q_i^c = f_i\left(q_i^1, q_i^2, ..., q_i^m\right), 1 \le i \le n$, f_i computes the compositional effects of the *i*th QoS parameter. User could give constraints for composite service on any parameter $C = \langle c_1, c_2, ..., c_n \rangle$. Our target is to optimize *Q(cs)* and make sure no QoS parameter exceeds the constraints *C*. Every possible composition of services is called a *plan*. We use the common accepted weighted average to evaluate a plan. In this paper, we assume the better plan is the one with smaller score. w_i is the significance of the *i*th QoS parameter:

$$Score(p) = \sum_{1 \le i \le z} w_i q_i^p, \left(0 \le w_i \le 1, \sum_{1 \le i \le z} w_i = 1 \right) \quad (1)$$

This is a common optimization problem but no existing approaches are applicable to our case, because in our selection model, any registry may only have a part of the

sequence. We need an iterative algorithm which could be applied to tasks one by one and finally get global optimized QoS. For space reasons, we only describe idea of the algorithm and a further optimization in this paper. Interested reader can refer to our previous publication[4] for more details.

3.2 Iterative Selection Algorithm

The basic ISA is enlightened by Extended Bellman-Ford Algorithm (EBFA)[9], but we modified it for optimizing nodes QoS rather than link QoS. For each task, basic ISA computes scores of all possible plans with previous result and records new plans in each service node for next computing iteration. If currently computing task is $t_i \ (1 < i \le l)$, the algorithm should have all possible plans from t_1 to t_{i-1} and carry out plans from t_1 to t_i. When the last task is computed, the best plan is selected.

Basic ISA has a significant problem. With the growing of candidate service number, the recorded plan number grows exponentially. For a task path with *l* tasks and *m* services for each task, the candidate plans would be m^l. In practice, services with better scores have higher possibility to be selected in the final composition plan. We have proposed a heuristic algorithm in our previous work which keeps *K* plans with the best scores in each service node, called ISA-Heu here. The algorithm can be further optimized.

The best plan for a part of the sequence structure may not always be the best for the whole sequence. In computation, some of the plans' QoS parameter may be too close to constraints, even if they have excellent score, they are highly likely to exceed constraints in next iteration. Task optimization should take user constraints into account and predict which plan is better for overall composite service. We achieve this prediction by adjust the scoring function applied in middle task computation. The new scoring function could magnify the effects of QoS parameters which are approaching constraints. We call this algorithm ISA-HeuPred. The adjusted scoring function with prediction is:

$$Score\,\mathrm{Pr}ed(p) = \sum_{1 \le i \le n} w_i \frac{q_i^s}{c_i - q_i^s}, \left(0 \le w_i \le 1, \sum_{1 \le i \le n} w_i = 1, q_i^s < c_i \right) \tag{2}$$

Suppose there are *l* tasks in critical task path and each task has *m* candidate services, in the worst case, the time complexity of ISA is $O(m^l)$, otherwise, the time complexity of ISA-Heu and ISA-HeuPred is $O(K^2 l m^2)$. The space complexity of ISA-Heu and ISA-HeuPred is $O(Km)$.

4 Iterative Computing for Composition Structures

Currently, mainstream composition logic description approach as WSBPEL (Web Services Business Process Execution Language) [10] is developed from traditional

business process modeling works. Sequence, switch, parallel, loop and pick are familiar basic structures for constructing composite service. In our distributed QoS architecture, no QoS registry has an overall view of the whole composition logic, even these basic structures may not exist on a single registry, so the QoS optimization of whole composite service have to be carried out in distributed and iterative manner.

Although registries have no knowledge of the overall process, they have the information that a specific task has how many predecessors and successors. For a task node has more than one successor, as a parallel, switch or pick begins from the node, we call it a *branch node*. For a task node has more than one predecessors, we call it an *aggregation node*. Both branch node and aggregation node are *structure nodes*. Other nodes are all called *sequence nodes*. For example, in Fig.1, T2 is a branch node, T6 is an aggregation node. For the node classification, 2 issues should be noticed: 1, loop structure could be transformed to limited number of switch structures[3][11]; 2, when partitioning, it does not matter what kind of branch it is, the branch type only affects QoS aggregation.

We define a *task path* as a set of task nodes begin at the start task node or a structure node, and end at the final task node, or before the next structure node in topology order. By this definition, a composition logic could be divided into several connected task path where every path is a sequence structure. In Fig.1, there are 4 task paths: $p_1 =< t_1 >$, $p_2 =< t_2, t_3, t_4 >$, $p_3 =< t_2, t_5 >$, $p_4 =< t_6, t_7 >$. These task paths are computed iteratively in topological order. When the branch node sends result to next task paths, the branch type (parallel, switch or pick) and node identification is sent with the result, so that corresponding aggregation node could compute aggregated QoS of previous paths. As presented in [3], every QoS parameter has an aggregation function. For example, price is the summary of all the selected services' price, availability is the product of all the selected services' availability. For service selection before composite service running, the aggregation functions are different for different structures. The detailed aggregation method of QoS parameter is not in the scope of this paper, but it is obvious that the optimal result of task paths aggregated together is still optimal.

The overall selection process for composite service in Fig.1 by our approach is: compute p_1 at registry 1 first. Based on p_1's result, compute p_2 at registry1 and p_3 at registry2 respectively. In fact, we optimize $p_1 + p_2 =< t_1, t_2, t_3, t_4 >$ and $p_1 + p_3 =< t_1, t_2, t_5 >$ as 2 independent sequences. p_1, p_2, and p_3 are aggregated at registry 3. Here, registry 3 identifies results of p_2 and p_3, aggregates them based on branch type, then combines with p_1. From the viewpoint of t_7, the topology before t_6 has nothing to do with its computation. Then, registry 3 finish computation of t_7 based on aggregated result of t_6.

The aggregation operation would bring additional execution time to selection algorithm but it could be ignored. If n task nodes are aggregation nodes, the time complexity of using ISA-Heu for task path would be $O(K^2 lm^2) + O(n) = O(K^2 lm^2)$.

5 Experiments

We studied the performance and effectiveness of our approach in different cases by a series of experiments. The experiments were conducted on a LINUX server with configuration of a Xeon 3.20GHz CPU and 2GB RAM, running Red Hat LINUX.

5.1 Evaluation Methodology

At first, we compare the performance of 2 types of ISA: non-heuristic and heuristic. In both cases, each service has 6 QoS parameters and each parameter is a randomly generated integer between 1 and 100. No constraints on any QoS parameter because we would like to test the worst case, but constraints could contribute to delete some plans and improve the execution time. For non-heuristic ISA, task number in composite service ranges from 1 to 10 and each task has 4, 5 or 6 candidate services respectively. In heuristic case, for the scalability of ISA-Heu, task number in composite service ranges from 10 to 100 with a step of 10 and each task has 20, 40 or 60 candidate services. Composite service is constructed by different structures, but these structures do not affect the overall selection time significantly, so we generate test composition logic by randomly repeating nodes in Fig.1. Aggregation function of every QoS parameter at t_6 is product. *K* is set at 5. The performance of ISA-Heu and ISA-HeuPred is the same so we only conduct our test on ISA-Heu. We run each case 100 times.

The heuristic algorithm may discard some "better" plans in computation. We study the effectiveness of 2 heuristic approaches by 2 criteria which has been used in our previous work: *success ratio* and *approximation*. For both heuristic approaches, we fix the task number at 8 and candidate service number for each task at 6. The constraints is adjusted to make 90% cases could have at least one feasible plan. The heuristic parameter *K* ranges from 1 to 8.

5.2 Result and Analysis

Figure 2 shows the performance comparison of ISA and ISA-Heu. Because the heuristic approach limit plan number kept in each service node, the plan search time in each iteration is greatly decreased. When task number is 9 and 6 services for each task, the computation time of ISA approaches 10 seconds and grows exponentially. Otherwise, the ISA-Heu could easily scale up to 5000 candidate services and computation time is under 2 seconds. In practical composite service scale, ISA-Heu could complete in several milliseconds.

The effectiveness of ISA-Heu and ISA-HeuPred are illustrated in Fig. 3. When K<5, the success ratio of ISA-HeuPred is much higher than ISA-Heu. Especially when K=1, the heuristic with prediction could get a very impressive ratio at about 92%. However, the approximation of ISA-HeuPred is lower than ISA-Heu by about 1

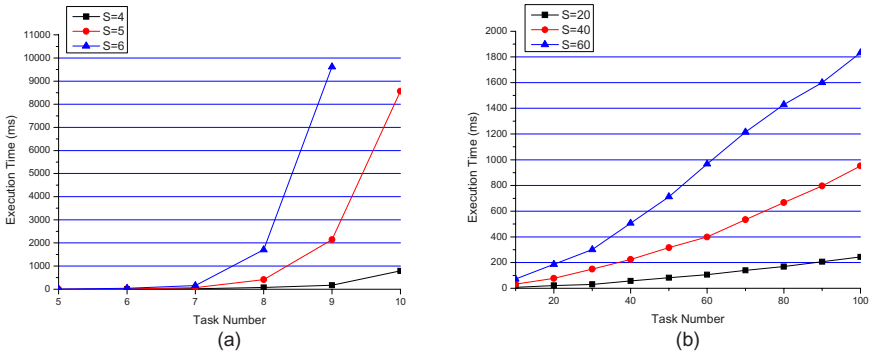


Fig. 2. (a) Execution time of ISA. (b) Execution time of ISA-Heu/ISA-HeuPred.

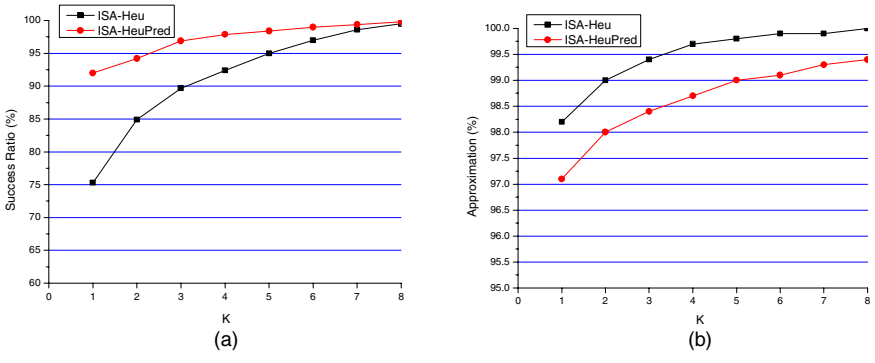


Fig. 3. (a) Success ratio of ISA-Heu and ISA-HeuPred. (b) Approximation of ISA-Heu and ISA-HeuPred.

percent in all tested values of *K*. But considering the approximation when K=1 could exceed 97%, this loss of approximation is highly acceptable.

6 Conclusion

In this paper, we presented a distributed service selection approach running on decentralized QoS registry architecture. The approach is combined by a service selection algorithm for simple sequence structure, and a method to divide complex structure to several connected sequence structures. The selection algorithm is modified from EBFA. By applying heuristic improvements which record the plans most likely to optimize overall path, the algorithm achieved excellent performance. The result of service selection for sequence structure could be aggregated at run-time to optimize the overall QoS of composite service.

Reference

1. Tsalgatidou, A., Pilioura, T.: An Overview of Standards and Related Technology in Web Services. Distributed and Parallel Databases 12(2) (2002) 135–162
2. Dustdar, S., Schreiner, W.: A survey on web services composition. International Journal of Web and Grid Services 1(1) (2005) 1–30
3. Zeng, L., Benatallah, B., Ngu, A., Dumas, M., Kalagnanam, J., Chang, H.: QoS-Aware Middleware for Web Services Composition. IEEE Transactions on Software Engineering 30(5) (2004) 311–327
4. Li, F. Su, S., Yang, F.C.: On Distributed Service Selection for QoS Driven Service Composition. Proceedings of the 7th International Conference on Electronic Commerce and Web Technologies,EC-Web'06,LNCS 4082 (2006)
5. Benatallah, B., Dumas, M., Sheng, Q., Ngu, A.: Declarative composition and peer-to-peer provisioning of dynamic Web services. Proceedings of the 18th International Conference on Data Engineering, ICDE'02 (2002) 297–308
6. Nanda, M., Chandra, S., Sarkar, V.: Decentralizing execution of composite web services. Proceedings of the 19th annual ACM SIGPLAN conference on Object-oriented programming, systems, languages, and applications, OOPSLA'04 (2004) 170–187
7. Canfora, G., Di Penta, M., Esposito, R., Villani, M.: An approach for QoS-aware service composition based on genetic algorithms. Proceedings of the Genetic and Evolutionary Computation Conference, GECOO'05 (2005) 1069–1075
8. Yu, T., Lin, K.: Service Selection Algorithms for Composing Complex Services with Multiple QoS Constraints. Proceedings of the 3rd International Conference on Service Oriented Computing, ICSOC'05, LNCS (3826) 130–143
9. Yuan, X.: On the extended Bellman-Ford algorithm to solve two-constrained quality of service routing problems. Proceedings of the 8th International Conference on Computer Communications and Networks, ICCCN'99 (1999) 304–310
10. OASIS, Web Services Business Process Execution Language Version 2.0 Public Review Draft, http://docs.oasis-open.org/wsbpel/2.0/, 23rd August, 2006
11. M. Gillmann, G. Weikum, and W. Wonner.: Workflow Management with Service Quality Guarantees. Proc. ACM SIGMOD Int'l Conf. Management of Data, pp. 228-239, June 2002.

Estimating Flow Length Distributions Using Least Square Method and Maximum Likelihood Estimation*

Weijiang Liu

School of Computer Science and Technology,
Dalian Maritime University, 116026, Dalian, China
wjliu@newmail.dlmu.edu.cn

Abstract. Traffic sampling technology has been widely deployed in front of many high-speed network applications to alleviate the great pressure on packet capturing. However, knowing the number and length of the original flows is necessary for some applications. This paper provides a novel method that uses flow statistics formed from sampled packet stream to infer the absolute frequencies of lengths of flows in the unsampled stream. First, flows are classified as small (S) or large (L) based on their probability that no packet is sampled. For large flows we use maximum likelihood estimation to infer their length distribution, and for short flows we apply least square method. The theoretical analysis shows that the computational complexity of this method is well under control, and the experiment results demonstrate the inferred distributions are as accurate as EM algorithm.

1 Introduction

With the rapid increase of network link speed, packet sampling has become an attractive and scalable means to measure flow data. However, knowing the number and lengths of the unsampled flows is required for some applications. Sampling entails an inherent loss of information. We expect use statistic inference to recover information as much as possible. However, more detailed characteristics of the original traffic are not so easily estimated. Quantities of interest include the number of packets in the flow–we shall refer to this as the flow length–and the number of flows with fixed length.

In [1], the authors studied the statistical properties of packet-level sampling using real-world Internet traffic traces. Scaling method and EM algorithm were given to infer the flow distribution from the sampled statistics in [2]. Scaling method is simple, but it exploits the sampling properties of SYN flows to estimate TCP flow frequencies; EM algorithm does not rely on the properties of SYN flows

* Thanks go to Jian Gong of Southeast University, for his assistance and suggestions. This work is supported in part by 973 Program of China under Grant No.2003CB314804.

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 793–796, 2007.
© Springer-Verlag Berlin Heidelberg 2007

and hence is not restricted to TCP traffic, but its versatility comes at the cost of computational complexity.

A flow is defined as a stream of packets subject to flow specification and timeout. In this paper, we will use the term original flow to describe the above flow. A sampled flow is defined as a stream of packets that are sampled at probability $p = 1/N$ from an original flow.

2 Probability Distribution of Original Flow Length

For a specific original flow F, let X_F denote the number of packets in F, Y_F denote the number of packets in the sampled flow from F. The conditional distribution of Y_F, given that $X_F = l$, follows a binomial distribution $Pr[Y_F = k|X_F = l] = B_p(l,k) = \binom{l}{k} p^k(1-p)^{l-k}$. By the conditional probability formula,

$$Pr[X_F = x|Y_F = y] = \frac{Pr[Y_F = y|X_F = x]Pr[X_F = x]}{Pr[Y_F = y]} \tag{1}$$

where $Pr[Y_F = y] = \sum_{i=y}^{\infty} B_p(i,y)Pr[X_F = i]$ by the complete probability formula.

Assume that original flow lengths satisfy Pareto distribution. Its probability mass function is

$$Pr[X_F = x] = \beta\alpha^\beta/x^{\beta+1}, \ \ \alpha, \beta > 0, \ \ x \geq \alpha \tag{2}$$

where β is called Pareto parameter. By Equations (1)and (2), we have:

Lemma 1. *Under the assumption that original flow lengths satisfy Pareto distribution, the probability that a sampled flow of length $y(\geq \alpha)$ is sampled from an original flow of length x is*

$$Pr[X_F = x|Y_F = y] = \frac{B_p(x,y)/x^{\beta+1}}{\sum_{i=y}^{\infty} B_p(i,y)\beta\alpha^\beta/i^{\beta+1}}.$$

Using similar method in [3], we obtain

Lemma 2. *Under the assumption of Lemma 1, for fixed $p = 1/N, \beta$ and $y(\geq \alpha)$, the probability $Pr[X_F = x|Y_F = y]$ is maximized at $x = Ny - n(p,\beta)$. It is increasing as x increases for $x < Ny - n(p,\beta)$ and decreasing as x increases for $x > Ny - n(p,\beta)$.*

The properties of function $n(p,\beta)$ can be seen in [3].

3 Estimation Method of Flow Length Distributions

Let $g = \{g_j : j = 1, 2, \cdots, n\}$, where g_j is sampled flow frequencies of length j, be a set of sampled flow length frequencies, $f = \{f_i : i = 1, 2, \cdots, n, \cdots\}$ a set of original flow length frequencies. Consider sampling the packets of an original flow of length Nj independently with probability $1/N$, the probability that no packet is sampled is $(1 - 1/N)^{Nj} = ((1 - 1/N)^N)^j$. $\{(1 - 1/N)^N\}$ is increasing in N and $\lim_{N\to\infty}(1 - 1/N)^N = 1/e < 0.37$. Thus for a given error ε,we require $(1 - 1/N)^{Nj} < (1/e)^j < \varepsilon$ and choose $j_{bord} \geq max(j(\varepsilon) = \lceil log(1/\varepsilon) \rceil, \alpha)$. For example, $j(0.01) = 5, j(0.001) = 7$. We classify two types of flows based on their probability that no packet is sampled. A flow is labeled as small (S) when it's probability that no packet is sampled is more than ε and large (L) otherwise.

3.1 Maximum Likelihood Estimation for Large Flows

For a sampled flow of length $j > j_{bord}$, by Lemma 2, the original flow length values of the $2N$ relatively large probabilities are $N(j-1)-n(p,\beta)+1, \cdots, N(j+1) - n(p, \beta)$ where $\beta = 1.0$. We estimate the sampled flow is sampled from one of the $2N$ original flows. Then there are $\frac{g_j}{2N}$ sampled flows that are sampled from one of original flows of the above lengths in $g_j (j > j_{bord})$ sampled flows. Therefore, for all large flows of length $i > Nj_{bord}$, we have

$$f_i = \frac{1}{2N}(g_j + g_{j+1}), where\, j = \lfloor (i + n(p, \beta) - 1)/N \rfloor. \tag{3}$$

3.2 Least Square Method for Small Flows

For all small flows of length $i \leq Nj_{bord}$, we estimate as follows:

$$g_j = \sum_{i=j}^{m} B_p(i, j) f_i, \quad where \quad m = \max\{i : f_i \neq 0\}, \quad j = 1, \cdots, Nj_{bord}. \tag{4}$$

For $i, l \in [\alpha, Nj_{bord}], i > l$, by Equation (2) we have:

$$f_i = (l/i)^{\beta+1} f_l \tag{5}$$

Substituting (3) and (5) into Equations (4):

$$\overline{g}_j = \sum_{i=j}^{l} B_p(i, j) f_i + \sum_{i=l+1}^{Nj_{bord}} B_p(i, j)(l/i)^{\beta+1} f_l, \quad j = 1, \cdots, Nj_{bord}. \tag{6}$$

Now, we let the values of β increase from 0 to 4.0 by increment 0.1. Applying each concrete β of the above values to compute Equations (6), we obtain:

$$y_j^{(\beta)} = \sum_{k=1}^{l} x_{jk}^{(\beta)} f_k \quad j = 1, \cdots, Nj_{bord}. \tag{7}$$

we use least square method to solve Equations (7) and get the solutions $f_k^{(\beta)}$, $k = 1, \cdots, l$. If each $f_k^{(\beta)} > 0, k = 1, \cdots, l$, then let

$$m_\beta = \sum_{k=1}^{Nj_{bord}} (y_j^{(\beta)} - \sum_{k=1}^{l} x_{jk}^{(\beta)} f_k^{(\beta)})^2.$$

We find the value of β such that m_β is minimized in all positive solutions. Denoting the found value as $\overline{\beta}$, we substitute the corresponding $f_l^{(\overline{\beta})}$ into Equation (5). Finally, we obtain $f_k^{(\overline{\beta})}, k = 1, \cdots, Nj_{bord}$. We write our estimation of original small flows as $f_k^{(\overline{\beta})}, k = 1, \cdots, Nj_{bord}$.

4 Evaluations and Comparison

Computational complexity. Let i_{max} denote the maximum original flow length. The computation for binomial coefficients of Equations (4) is $O(Nj_{bord}i_{max})$. The computation for least square method needs little time. We compare the computational complexity of our method against the best known EM algorithm in [2] for estimating flow distribution from sampled traffic. In [2] for all ϕ_i completing an EM iteration is $O(i_{max}^2 j_{size})$,where j_{size} denote the number of non-zero sampled flow length frequencies g_j.

Estimation accuracy: We use some traces from [4,5] in our comparison experiments. Experimental results show the estimation accuracy of our algorithm is close enough to that of EM algorithm. In most cases, our algorithm is much more accurate.

5 Conclusions

In this paper we present a novel method for estimation of flow length distributions from sampled flow statistics. The main advantage is that it could significantly reduce the computational complexity. The theoretical analysis shows that the computational complexity of our method is well under control. The experimental results demonstrate that our method achieves an accurate estimation for flow distribution.

References

1. Duffield, N.G., Lund, C. , Thorup, M.: Properties and Prediction of Flow Statistics from Sampled Packet Streams. ACM SIGCOMM Internet Measurement Workshop 2002,159-171, November 2002.
2. Duffield, N.G., Lund, C. , Thorup, M.: Estimating Flow Distributions from Sampled Flow Statistics. IEEE/ACM Transation on Networking, **13**(2005) 933-945.
3. Liu W.J., Gong J., Ding W.,and Peng Y.B.: Estimating Original Flow Length from Sampled Flow Statistics. Lecture Notes in Computer Science, **3994**(2006)120-127.
4. NLANR: Abilene-III data set, hppt://pma.nlanr.net/Special/ipls3.html.
5. NLANR:Abilene-I data set, http://pma.nlanr.net/Traces/long/bell1.html.

Local Link Protection Scheme in IP Networks

Hui-Kai Su[1] and Cheng-Shong Wu[2]

[1] Dept. of Computer Science and Information Engineering
Nanhua University
No. 32, Chung Keng Li, Dalin, Chia-Yi, 622 Taiwan
hksu@mail.nhu.edu.tw
[2] Department of Electrical Engineering
National Chung-Cheng University
No. 160, San-Hsing, Min-Hsiung, Chia-Yi, 621 Taiwan
ieecsw@ccu.edu.tw

Abstract. In this paper, we proposed an IP Local Link-Protection (IPLLP) scheme based on the characteristic of shortest-path routing in IP networks. Our scheme working in an intra-area routing domain provides a simple and efficient solution to improve IP network survivability without extra control protocols and enhanced routing protocols. Because the backup next hops are predetermined in advance, the service interrupted time can be limited to a few milliseconds. In the simulation results, we observe that IP Local Link-Protection scheme can efficiently improve network survivability in a small-scale and high-degree network.

Keyword: IP network survivability, link protection, fast reroute.

1 Introduction

Network availability becomes a more and more important QoS (Quality of Service) parameter in IP networks. Certain services should not be interrupted regardless of the scale, duration and type of failures. IP network has the ability of routing restoration since the ARPANET was built, i.e. IP restoration. In addition, many protection schemes have been implemented at lower layers in IP networks, e.g. SONET APS (Automatic Protection Switching), MPLS Fast-Reroute, etc. Since their backup paths are decided and set up for network failures in advance, the service interrupted time can be limited to a few milliseconds.

Recently, the IP protection issue has been discussed since 2002. The precomputation scheme of second shortest paths is introduced in [1]. However, in the practice, how to decide feasible backup routes efficiently, provide an efficient fast reroute service and avoid routing loops was not discussed. The drafts of IP Fast-Reroute (IPFRR) framework [2] and LFAP (Loop Free Alternate Paths) scheme[3] were proposed by *IETF Routing Area Working Group*. Equal Cost Multipath (ECMP) and LFAP offer the simplest repair paths, and it is anticipated that around 80% of failures can be repaired using these alone. However, the ECMP scheme needs extra control protocols to negotiate which equal cost

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 797–800, 2007.
© Springer-Verlag Berlin Heidelberg 2007

path is failed after a node or a link fails. Additionally, multi-hop repair paths are considerably more complex, and extra control protocols or enhanced routing protocols should be needed. It is anticipated that around 98% of failures can be repaired.

In this paper, we proposed an *IP Local Lode-Protection (IPLLP)* scheme for IP networks in an intra-area routing domain. Our scheme provides a simple and efficient solution for IP network protection without extra control protocols and enhanced routing protocols. According to characteristic of destination tree with shortest-path routing, our scheme can prevent service disruption and packet loss caused by the loops which normally occur during the re-convergence of the network following a failure. The packets through the failure link can be locally rerouted to the backup next hop as soon as the upstream adjacent nodes of the failure node detect the failure. Because the backup next hop is predetermined, the service interrupted time can be limited to within a few milliseconds.

2 Local Link Protection Scheme

A network topology $G(N, L)$, where N and L denote the router set and the link set respectively, is given. Note that $G(N, L)$ can be deduced from the database of link-state routing protocols. After calculating the primary next hop, every router assumes that the connected link with their primary next hop to the destination node d is failed. First, they calculate the destination tree T_d to node d according to $G(N, L)$. Second, they remove the connected link with their primary next hop from the destination tree T_d, divide the tree into two subtrees and then try to repair the incomplete destination subtree to node d with the leaved tree as a partial shortest-path destination tree $\check{T}_d$. Because only the connected nodes of the failed link can sense this failure in the first time, the traffic flows are delivered along the partial shortest-path destination tree $\check{T}_d$ before IGP converges. Therefore, if the adjacent nodes can cooperatively construct the partial shortest-path destination tree based on the divided subtrees, their loop-free backup next hops to the destination node d will be decided.

An example of IPLLP scheme is shown in Fig. 1. If a bidirectional link $\{N_3, N_2\}$ fails, N_3 is responsible to predecide a feasible backup next hop for this link failure to the destination node N_0. The $\{N_3, N_2\}$ is removed from the topology, and T_{N_0} is divided into T'_{N_0} and T'_{N_3}. If N_3 can select a shortest-path and feasible next hop on T'_{N_0} to construct a partial shortest-path destination tree $\check{T}_{N_0}$, its backup next hop is existent; otherwise, the link failure cannot be protected. N_4 is selected for link $\{N_3, N_2\}$ failure to N_0 by N_3 in our example, packet loss caused by the loops doesn't arise during the IGP routing recovery.

3 Simulation and Numerical Result

The goal of our simulations is to justify our IPLLP scheme. We observe the performance in protectability. Protectability is defined as the ratio of the protectable O-D pairs to the recoverable O-D pairs. For example, after IGP converges, the 6

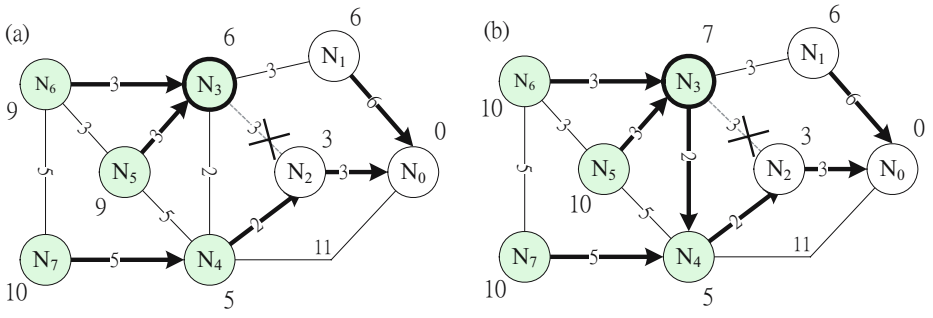


Fig. 1. (a) The divided subtrees after link $\{N_3, N_2\}$ failure occurs, and (b) the traffic flows along the partial shortest-path destination tree to N_0 with IPLLP scheme

failure paths can be repaired. However, by a protection scheme, if only 3 paths can be protected, the protectability is equal to 0.5.

In our simulations, topologies are given. The shortest path algorithm (i.e., Dijkstra's algorithm), our IPLLP and the LFAP are implemented in our program. First, the routing tables of each node that contain primary next hops and backup next hops are built. Second, all scenarios of each link failure are simulated, and all O-D pairs are tested according to the present routing tables. Third, the shortest path algorithm is performed to repair all scenarios of each link failure. Finally, the numbers of the protected paths and the recoverable paths are collected, and the average protectability can be computed statistically.

Random flat topologies in our simulation are generated by BRITE topology generator[4]. A new node connects to a candidate neighbor node using Waxman's probability function ($\alpha = 0.19$ and $\beta = 0.2$), and the total node number $|N|$ and the average connectivity degree $\overline{d}$ are given. The connectivity degree means the number of connected links in a node, and then a topology $G(N, L)$ can be generated. Additionally, all link costs in our topologies are constant and equal, i.e., the same link bandwidth.

Fig. 2 shows the relationships between average network protectability and network scalability while a link failure occurs in the random flat topologies whose average connectivity degrees are equal to 4 and 6. We observe the performances of our IPLLP scheme are better than the LFAP scheme, especially in high connectivity-degree networks. Some ECMP next hops may be feasible loop-free backup next hops, but they don't satisfy the criteria of LFAP scheme. Additionally, with the increase of network scalability, the protectabilities of both schemes are decreased. Although many available paths can be found in a large IP network, the IP network still limits the traffic to go through the few shortest paths because of the destination-based routing. Thus, the performance of them cannot achieve to that of other protection schemes in connection-oriented networks, e.g., MPLS network, SONET and optical network; however, the IPLLP scheme can provide a simple and effective protection solution.

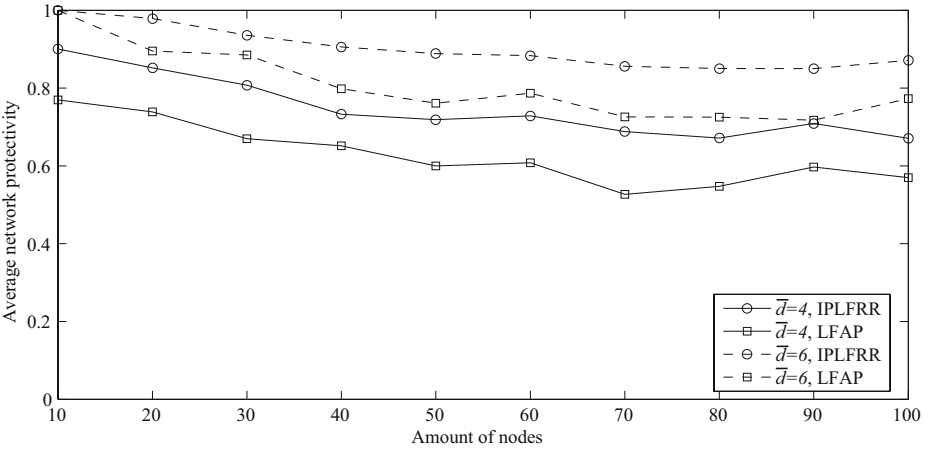


Fig. 2. The performance of link protection in random flat topologies

4 Conclusion

In this paper, we propose an *IP Local Link-Protection (IPLLP)* scheme in an intra-area routing domain, which provides a simple and efficient solution for IP network protection. Extra control protocols and enhanced routing protocols are not needed if a conventional link-state routing protocol is used. In our simulation, all of IPLLP scheme and LFAP scheme work well, and they are suitable for a small-scale and high-degree intra-area network. In a small network, the computational complexity would not be the major factor to impact the performance of IPLLP scheme. Thus, the alternative of computational complexity and network scale may be considered between IPLLP scheme and LFAP scheme. Additionally, based on the Destination SPT concept, our work may extend to protect multiple failures by grouping failures and defining failure events. Its affected area and feasible next hop can be decided in advance. Finally, we believe that IPLLP scheme can give a good solution to IP network protection technology.

References

1. Alaettinoglu, C., Zinin, A.: IGP fast reroute. In: IETF Routing Mtg., Atlanta, GA, USA (2002)
2. Shand, M., Bryant, S.: IP fast reroute framework. IETF Draft (2006) draft-ietf-rtgwg-ipfrr-framework-05.txt.
3. Atlas, A., Zinin, A.: Basic specification for IP fast-reroute: Loop-free alternates. IETF Draft (2006) draft-ietf-rtgwg-ipfrr-spec-base-05.txt.
4. Medina, A., Lakhina, A., Matta, I., Byers, J.: BRITE: an approach to universal topology generation. In: Proc. IEEE Modeling, Analysis and Simulation of Computer and Telecommunication Systems. (2001)

An Authentication Based Source Address Spoofing Prevention Method Deployed in IPv6 Edge Network

Lizhong Xie, Jun Bi, and Jianpin Wu

Network Research Center, Tsinghua University,
Beijing, 100084, China
junbi@tsinghua.edu.cn

Abstract. In today's Internet routing architecture, the router doesn't validate the correctness of the source address carried in the packet, nor keep the state information when forwarding the packet. Thus the DDoS attacks with spoofed IP source address can cause security problems. In this paper, we aim to prevent the attackers from attacking somewhere outside the IPv6 edge network with forged source address in the fine granularity. The proposed methods include source address authentication by using session key and hash digest algorithm, and replay attack prevention by combining the sequence number method and the timestamp method. This paper presents the algorithm design and evaluates its feasibility and correctness by simulation experiments.

Keywords: Source Address Spoofing, IPv6, Edge Network.

1 Introduction

Today's Internet is vulnerable to a lot of security threats, such as DDoS attack TCP SYN attack, Smurf attack, etc. These attacks greatly rely on the IP source address spoofing. According to the statistics of US CERT, the increase rate of the Internet security attacks is much quicker than the development of the Internet itself.

There are several approaches to tackle the IP source address spoofing, including:

1. The end-to-end cryptographic authentication based approaches, such as IPSec, SPM [1]. IPSec supports host-to-host cryptographic authentication; and SPM is a kind of AS-to-AS authentication.
2. The tracing back based approaches. In these methods, some useful information is recorded by the routers, the ICMP messages or the packets themselves. When the receiver finds some packets whose IP source addresses are forged, it can trace back to the forger by using the recorded information. SPIE [2], iTrace [3], PPM [4], APPM [5], PPPM [6], and DPM [7] all belong to this kind of methods.
3. The filtering based approaches. In these methods, the routers filter the forwarding packets according to some filtering rules (such as the IP prefixes). These methods include Ingress filtering [8], DPF [9], and SAVE [10].

However, all of these approaches have some drawbacks:

1. In the end-to-end authentication methods, the main problem of IPSec lies in: the routers can't validate the sender's IP address, which is, in some extent, under some

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 801–808, 2007.
© Springer-Verlag Berlin Heidelberg 2007

possible security threats on the routers. SPM has a rough granularity of source spoofing prevention because it just supports authentication in the AS granularity.

2. The tracing back approaches have three main flaws. Firstly, they are passive methods because they don't take action until attack has occurred. Secondly, these algorithms of the tracing back are too complicated to deploy. Thirdly, the effectiveness of this method always relies on the sensitivity of the intrusion detection.

3. The current filtering based approaches are the most effective because they can proactively prevent the source address spoofing attacks. However these methods can only filter the packets based on the IP address without validation (which means an attacker that use a legal source address won't be filtered). Therefore, it can only prevent source address spoofing in a rough granularity.

The emerging IPv6 network grants us an opportunity to redesign the trustworthy network infrastructure. In this paper, we aim to prevent the attackers from attacking somewhere outside the IPv6 edge network with forged source address in the fine granularity. If an attacker forges the source address of another host in the same edge network, the ingress filtering method won't work. We also need to prevent the replay attacks, because attackers can send the detected victim's packets to the outside of the edge network since its source address is valid. In our approach, if an attacker sends packets to somewhere outside the edge network by forging the IP address of another host or replays the victim's packets, these malicious packets will be filtered out and can't damage the outside of the edge network. Moreover, the method can work with other effective but rough-granularity methods such as ingress filing and SPM to form multi-fence defense architecture for the next generation Internet.

The rest of this paper is organized as follows: Section 2 describes the algorithm, including: the mechanism of source address validation and the mechanism of replay attacks prevention; Section 3 presents simulation results on the feasibility and the correctness of our approach; Section 4 discusses the future work and concludes the paper.

2 The Algorithms

2.1 Method Description

Our approach of source address spoofing prevention in the IPv6 edge network is based on filtering and authentication. As shown in Figure 1, we deploy a security gateway to carry out the authentication algorithm. The authentication algorithm includes two main mechanisms: the source address validation mechanism which verifies the source address by using the session key and hash digest algorithm (such as MD5 [11], SHA-1 [12], etc.), and the anti-replay mechanism which combines the sequence number method and the timestamp method.

Initially, each host in the edge network needs to be authenticated by the security gateway. This action also binds the host's IP address and its session key which is shared by the host and security gateway. If the access authentication succeeds, each packet sent to somewhere outside the edge network will carry a signature. Then, the packet passes through the source address validation and the replay check of security gateway. If both the authentications succeed, the packet will be forwarded to the edge router of the edge network; otherwise, the packet will be dropped.

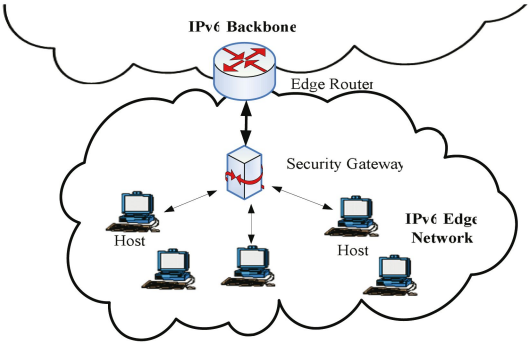


Fig. 1. The deployment scenario

Our approach mainly includes the following steps:

1. When a host wants to access the Internet, it firstly accesses the security gateway for authentication. This process can use the existing access authentication mechanism such as RADIUS [13], Kerberos [14].
2. The host generates a session key and sends it to the security gateway via some key exchange mechanisms such as IKE [15], IKE2 [16]. The security gateway binds the session key and the host's IP address.
3. When the host sends packets to somewhere outside the edge network, it needs to generate one signature for each packet by using the hash digest algorithm. The signature is carried in a new IPv6 extension header, named as "source address validation header".
4. The security gateway authenticates the signature carried in the packet to validate the source address.
5. The security gateway identifies the replay packets by checking whether the sequence number of the packet is increasing within the life time of the session key.

In addition, the session key will be changed frequently for the security purpose.

2.2 Source Address Validation Algorithm

Because every packet needs to be authenticated when they are sent to somewhere outside the edge network, the primary design requirement of source address validation algorithm is the high performance. We evaluate several authentication algorithms and finally choose the algorithm which verifies the source address by using the session key and hash digest algorithm (such as MD5, SHA-1, etc.). In this algorithm, the host certifies its ownership of a certain IP address via showing the security gateway its secret session key which is shared with the security gateway. The session key is a random number that is at least 12 bytes long. The host generates a signature by using the hash digest function with the session key and some certain data as the input. The security gateway checks the signature to validate the source address. Since the secret session key itself is not transferred in the plaintext form, the attacker is thus impossible to modify or forge another host's packets. Therefore this mechanism can support the authentication effectively.

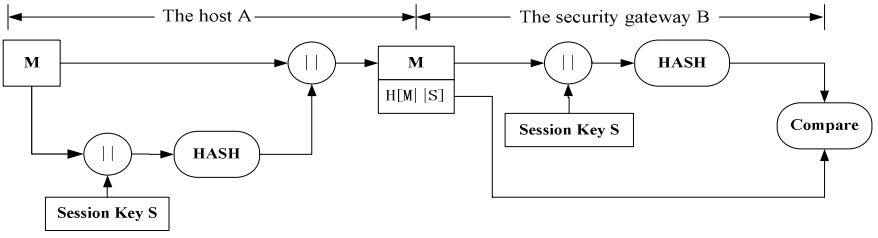


Fig. 2. The authentication process of source address validation

The authentication process shown in Figure 2 is as follows:

1. Host *A* sends a packet *M* to the security gateway *B*. *M* carries a signature *H*[*M*||*S*] which is computed by the hash digest function by using the session key *S* and the certain part of the packet *M* (source address, destination address, sequence number, etc.) as the input.

When the security gateway *B* receives the packet *M*, *B* can re-compute the hash value *HB* according to the packet *M*, since *B* also knows the session key *S*. If *HB* is equal to the signature *H* [*M*||*S*] carried in the packet *M*, the security gateway *B* can confirm that the packet has the valid source address; Otherwise, *B* can conclude that the packet's source address is forged and then drops it.

2.3 The Anti-replay Algorithm

Our anti-replay algorithm combines the timestamp method and the sequence number method. Both the timestamp method and the sequence number method are the prevailing anti-replay algorithms. The timestamp method works as follows: when the host *A* sends a packet *M* to the security gateway, the packet *M* is marked with a timestamp T_a , which represents the sending time of the packet *M*. Once the security gateway receives the packet *M*, it reads its local time T_b. If $|T_b\text{-}T_a|>\triangle T$, where $\triangle T$ is the admission time window, the security gateway can conclude that the packet *M* is a replay one then drops it. However, it's hard to synchronize the clocks of the host and the gateway exactly. Moreover, the transfer time of the packet in the network is also uncertain. Therefore, the admission time window $\triangle T$ is always larger than the real transfer time of the packet. This feature makes the timestamp method unfaithful for anti-replay. When $|T_b\text{-}T_a|<\triangle T$, the packet should be a non-replay one. But afterwards, if the replay packet is received in the margin time ($\triangle T\text{-}|T_b\text{-}T_a|$), the security gateway will wrongfully regard it as a normal packet.

The main idea of the sequence number method is: when the host *A* sends packets to the security gateway, each packet carries an incremental sequence number. If the latest packet's sequence number is greater than the previous one, the packet is normal; otherwise, the packet is a replay one. However, this method may not identify some replay packets when the sequence number is used in a cycle way. For example, assuming the length of the sequence number is 16 bits, once the sequence number reaches the maximum 65535, it will return to 0 and increase as the previous cycle. In this case, if the attacker keeps a packet of the n^{th} cycle and replays it in the $(n+1)^{th}$ cycle, the security gateway can't identify the replay packet.

In order to overcome the drawbacks of the timestamp method and sequence method, we combine these two methods to prevent the replay attack. The timestamp method can use the sequence number mechanism to identify the replay packets in the admission time window $\triangle T$; And the sequence method can avoid the confusion between the normal and the replay packet by limiting the period of the sequence number cycle within the admission time window $\triangle T$.

However, in our approach it is not necessary to mark a real timestamp in the packet. For the convenience of updating the session key, the packet carries the session key version when it is sent to somewhere outside the edge network. We can regard the session key version as the timestamp and the life time of the session key as the admission time window $\triangle T$. So what we need to do is just setting the life time of the session key less than the period of the sequence number cycle.

2.4 IPv6 Source Address Validation Header

In our approach, we design a new IPv6 extension header to carry the signature, the sequence number and other useful information. We call this new extension header "source address validation header".

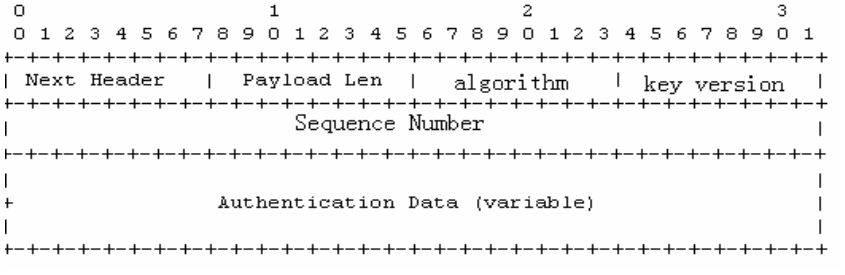


Fig. 3. The format of the source address validation header

The format of the extension header is shown in figure 3:

- Next Header: 8-bit. Indicate either the type of the next extension header or the protocol type of the payload (TCP/UDP).
- Payload Len: 8-bit. Length of the source address validation header in 8-octet units, not including the first 8 octets.
- Algorithm: 8-bit. Point out the hash digest algorithm. For example, MD5 is set to 1.
- Key Version: 8-bit. Key version refers to the version of the current using session key, for the convenience of updating the session key. We also regard it as a timestamp.
- Sequence Number: 32-bit. It is used to anti-replay as described above.
- Authentication Data: 128 bits if using MD5 as the hash digest algorithm. The authentication data is computed by the hash digest algorithm. The input of the hash digest algorithm includes: IPv6 source address, IPv6 destination address, sequence number, session key and the session key version.

The usage of IPv6 source address validation header:

1. Each packet which is sent to somewhere outside the edge network needs to carry a source address validation header. The header is inserted after the IPv6 header and before all the other extension headers.
2. The security gateway does the following validation when receiving the packet:

– Drop the packet directly if there is no source address validation header in the packet.
– Check the authentication data in the source address validation header as described above. Drop the packet if this process is failed.
– Affirm that the sequence number is greater than the previous one within the life time of the session key. If that is not true, drop the packet.

If the whole process in the step 2 passes, the security gateway needs to remove the source address validation header from the packet. Considering the partial deployment, if we don't remove the source address validation header, the hosts in other edge network may drop the packet due to misunderstanding of the new extension header. This step is unnecessary if our approach has been global deployed.

3 Experiment Evaluation

The performance and the correctness of our approach have been evaluated by simulation.

3.1 The Performance Evaluation

As described above, the performance is the primary requirement of source address validation algorithm. In our approach, this requirement means whether the performance of the hash digest algorithm is high enough. We do some experiments to testify it. Table 1 shows our experimental results, which are evaluated in the platform of Intel P4 2.0G CPU and 512M memory.

Table 1. The performance compare of two main hash digest algorithms

HASH Digest algorithm	The capacity per second (MB/S)
MD5	204.346
SHA-1	65.963

The results show that the performance of MD5 is about 1.63 Gbps (204.346MB/s 8). This performance can satisfy the requirements of most edge networks. We should note that this result is gotten from the MD5 algorithm implemented in the software. If we implement the MD5 algorithm by using hardware, we will get a higher performance. Therefore, the source address validation algorithm in our approach is completely feasible.

3.2 The Correctness Evaluate

In order to test the correctness and the effectiveness of the source address validation mechanism and the anti-replay mechanism, we performed a simulation experiment.

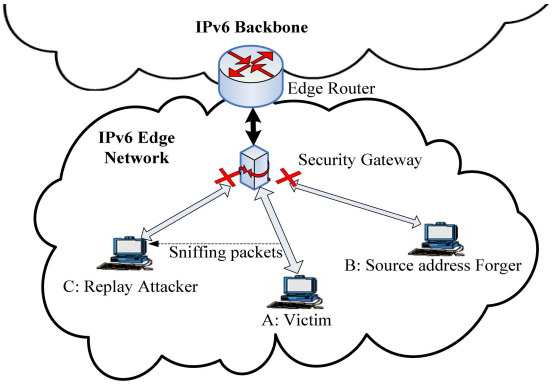


Fig. 4. The scenario of the simulation experiment

As shown in Figure 4, host *A* is the victim; host *B* sends packets to somewhere outside the edge network using the host *A*'s IP address as its source address; host *C* sniffs the packets sent by *A* and replays them. Initially, the functions of the security gateway are turned off. All the forged packets produced by *B* and the replay packets produced by *C* can be sent outwards. Once we turn on the functions of the security gateway, all these malicious packets are filtered by the security gateway (We build 10,000,000 malicious packets during the experiment.).

The experiment results show that both the source address validation mechanism and the anti-replay mechanism are effective.

4 Conclusion and Future Work

In this paper, we aim to prevent the attackers from attacking somewhere outside the IPv6 edge network with forged source address in the fine granularity. The proposed methods include source address authentication and the mechanism of anti-replay. The authentication algorithm uses the signature generated by the hash digest function (such as MD5, SHA-1, etc.) with the session key and the certain part of the packet. The anti-replay mechanism combines the sequence number method and the timestamp method to prevent the replay attack more reliably.

We evaluate our approach by using the simulation experiments. The experiment results show that our approach can prevent the source address spoofing and the replay attack effectively; and the performance of our approach is high enough to satisfy the requirement of the most edge networks. Moreover, our approach supports partial deployment.

The proposed fine-granularity method can work with other effective but rough-granularity methods such as ingress filing and SPM to form multi-fence defense architecture for the next generation Internet. We have implemented the prototype system and are deploying those mechanisms in CERNET2 (China Education and Research Network) IPv6 network.

The proposed method doesn't consider the multihoming situation yet, which refers to that an edge network obtains two or more simultaneous IP connectivity. In the

multihoming environment, a host in the edge network may have several IP addresses and the edge network may have several outbound links to different ISPs. In this case, we are studying whether we should deploy security gateway for each outbound link or just use one security gateway for all outbound traffic. We are also considering a gateway backup mechanism to protect the security gateway from attacking and becoming a bottleneck of the user traffic.

References

1. Bremler-Barr, A. and Levy, H.: Spoofing Prevention Method, INFOCOM (2005)
2. Snoeren, C. and Luis, A.: Hash-based IP traceback, SIGCOMM (2001)
3. Bellovin, S.: ICMP Traceback messages, IETF Internet Draft draft-ietf-itrace -03.txt (2003)
4. Savage, S., Wetherall, D., Karlin, A. and Anderson, T.: Pratical network support for IP traceback, SIGCOMM (2000)
5. Rizvi, B.: Analysis of Adjusted Probabilistic Packet Marking, IPOM 2003
6. Al-Duwairi, B. and Manimaran, G.: A Novel Packet Marking Scheme for IP Traceback, ICPADS (2004)
7. Belenky, A. and Ansari, N.: Tracing multiple attackers with deterministic packet marking (DPM),PACRIM (2003)
8. Ferguson, P. and Senie, D.: Network Ingress Filtering: Defeating Denial of Service Attacks which employ IP Source Address Spoofing , RFC2827, (2000)
9. Park, K. and Lee, H.: On the effectiveness of route-based packet filtering for distributed DoS attack prevention in power-law internets, SIGCOMM (2001)
10. Li, J., Mirkovic, J., Wang, M., Reiher, P., and Zhang, L.: SAVE: Source Address Validity Enforcement Protocol, INFOCOM (2002)
11. Rivest, R.: The MD5 Message-Digest Algorithm, RFC1321, (1992)
12. Eastlake, D. and Jones, P.: US Secure Hash Algorithm 1 (SHA1), RFC 3174, (2001)
13. Rigney, C., Willens, S., Rubens, A. and Simpson, W.: Remote Authentication Dial In User Service (RADIUS), RFC2865, (2000)
14. Kohl, J., and Neuman, C.: The Kerberos Network Authentication Service (V5),RFC 1510, September (1993)
15. Harkins, D. and Carrel, D.: The Internet Key Exchange (IKE), RFC 2409, (1998)
16. Kaufman, C.: Internet Key Exchange (IKEv2) Protocol, RFC 4306, (2005)

An Intrusion Plan Recognition Algorithm Based on Max-1-Connected Causal Networks

Zhuo Ning and Jian Gong

Department of Computer Science and Engineering, Southeast Univ.
Nanjing, Jiangsu, 210096 China
{zhning,gjian}@njnet.edu.cn

Abstract. Intrusion plan prediction and recognition is a critical and challenging task for NIDS. Among several approaches proposed so far, probability inference using causal network seems to be one of the most promising mechanisms. Our analysis shows that the polytree is limited in its expressiveness, and belief updating in max-k-connected networks is hard for all $k \geq 2$ [12]. To find a tradeoff between expressive power and inference efficiency, this paper extends the structure of causal network from polytree to max-1-connected Bayesian network, and proposes a new intrusion plan prediction algorithm IPR on it. We evaluate the approach using LLOS1.0, and the results demonstrate that IPR can predict the occurrence probability of DDOS when Sandmind attack occurs to gain root privilege, and then confirm the prediction in the beginning of Syn flooding.

Keywords: NIDS, Max_1_connected Causal Network, belief updating, MAP.

1 Introduction

The wide spread of use of NIDS and the increasing complain about its high volume alerts have led to more and more intensive exploratory works on it. People begin to realize that the high false positive ratio is due to short of understanding of inner logical relations in the attack flows. So it is not reasonable to detect intrusion only based on single-packet signature and to view the attack flow separately. Among the recent researches focusing on how to accumulate the alert logical relations from the context [1], inference using causal network is one of the most promising approach for its powerful expression of causality and belief propagation consistent with ongoing evidence. [2] designed a two-level Bayesian tree model to discover novel attack strategies by correlating alerts. The expensive cost of this method leads to its poor behavior on practice. [3] proposed an abnormal IDS called eBayes TCP to detect some TCP abnormal behaviors, but the false positive ratio is also high. Comparatively more mature work was done by [4]. [4] proposed an abuse detecting approach based on polytree. Relying on the library of attack plans (defined as polytree), belief updating algorithm is used to calculate the new belief of each node when a new evidence entered, and then the node with the highest score is considered as the most possibly occurring attack. But the causal network used in [4] is a polytree, as shown in the sequel its expressive power is too limited in illustrating attack plan.

A Bayesian Network (BN) G=(V, P) is represented as a directed acyclic graph G where V is a set of nodes, and each one of V stands for a variable. P is a set of edges

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 809–816, 2007.
© Springer-Verlag Berlin Heidelberg 2007

and each one of P denotes a causal relationship between a couple of variables. A polytree topology is defined as BN where for every pair of nodes (x_1, x_2), there is at most one path from x_1 to x_2 in the *underlying undirected graph*[6]. When defining the attack tree, security analysts decompose the final goals into subgoals iteratively until those of the lowest level are exercisable penetration points. In the above process, a causal network is expanded and its branches are built to identify the different subgoals. So it is common that x_1 and x_2 have two same children. Unfortunately in this case the path between them will be two (as shown in Fig.1). In order to escape the limitation of polytree in expression, there are naturally two ways to try. One is to remove nodes to make it polytree-structured by conditioning algorithms[5,6] or cluster schemes[7,8]. However, all these reductions are usually exponential in some aspect of the problem instance and not efficient. Another way is to increase the number of paths between pairs of nodes and broad the known classes of tractable Bayesian networks. [9] proved that Belief updating in max-k-connected networks is hard for all $k \geq 2$, even with no evidence. This paper puts forward max-1-connected Baysian Network(M1CBN) as the tradeoff in balancing its expressive power and inference efficiency. Unlike Polytree, M1CBN is defined as networks where for each couple of nodes (x_1, x_2) in DAG, there is at most one *directed path* form x_1 to x_2. The difference between them is illustrated in Fig.1. Clearly, all polytrees are M1CBN, but not vice versa. Based on M1CBN, an intrusion plan recognition algorithm (IPR) is introduced which not only benefits more powerful expressive power, but also retains polynomial performance. In practice many attack graphs are expressed by multiple connected BN, and it is more efficient to transform them to M1CBN and apply IPR directly than to transform them to polytree and apply Pearl's algorithm as [4].

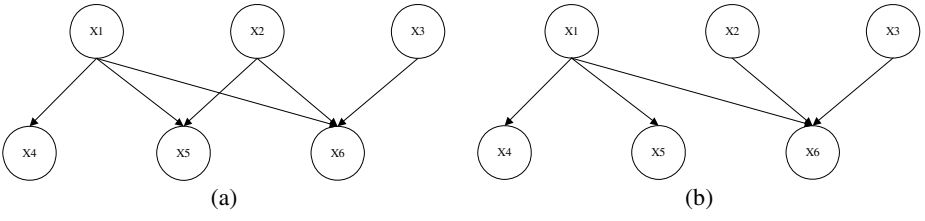


Fig. 1. (a) Max_1_connected BN; (b) Polytree

Intuitionally the task of plan recognition is to find an explanation for the observed evidence e. The explanation is usually composed of a set of hypotheses, and what these hypotheses are and the values of them are particularly concerned. In a BN G=(V, P) E denotes nodes that have been trigged by evidence e, and W=V-E denotes all variables without evidence. Any assignment of W that is consistent to e is called an explanation of e. The task of recognition is to find an assignment that makes

$$P(w \mid e) = \underset{W}{Max}\, P(W \mid e) \tag{1}$$

and it is called the most possible explanation (MPE). Usually people tend to ignore attack details and focus only on key steps and the final goal(that is, to focus on some specific variables of W). When the assignment w is a partial one the task of (1) becomes a MAP problem (to find a most probable instantiation of a set of variables

given evidence). Generally the MPE and MAP problem over Bayesian Network are NP hard but in some known tractable subclasses such as polytree, polynomial algorithms have been found for MPE problem, whereas MAP remains hard.

The rest of the paper is organized as follows. Section 2 shows some important results in inference in M1CBN. Based on section 2, respectively a new approximate belief updating algorithm BeliefUpdating and intrusion plan recognition algorithm IPR are addressed in section 3 and 4. Section 5 applies IPR in LLDOS1.0, a data set of DARPA 2000, and reports our experiment results. Finally, we conclude the paper and discuss future research directions on this topic in section 6.

2 Intrusion Plan Recognition on M1CBN

For the sake of convenience, we introduce some notations. Given a M1CBN B and any node X in B, we denote by $\prod(X)$ the parents of node X by $D(\prod(X))$ the set of all possible assignment on $\prod(X)$, and by $\prod^*(X)$ the set of all the ancestors of X, including X (* is reflexive and transitive closure to $\prod$). The ancestor graph $G^*(X)$ of a node X, induced by X, is composed of all the nodes in $\prod^*(X)$ and the arcs connecting them in B.

Proposition 1. MPE/MAP when restricted to M1CBN is NP-hard, even with no evidence.

Sketch of Proof. Clearly a 2_level BN is a M1CBN, as is shown in Fig.1(a). [10] proved MPE/MAP restricted to 2_level BN with evidence is NP hard, and MPE/MAP on 2_level BN without evidence remains NP hard [11]. And then MPE/MAP on 2-level M1CBN is NP hard. Therefore, MPE/MAP restricted to m-level (m≥2) M1CBN is NP hard as well.

Proposition 2. Belief updating in M1CBN is NP-hard.

This conclusion directly follows from the above proof.

Theorem 1. Top-down belief updating in M1CBN without evidence can be performed in time linear in the size of the network.

Proof. Given a BN without evidence, the marginal probability in any node $X_i = x_{i,j}$ is:

$$P(X_i = x_{i,j}) = \sum_{A \in D(\prod(X_i))} P(X_i = x_{i,j} \mid A) P(A) \tag{2}$$

In a max-1-conneted BN, any predecessor of X is d-separated, and thus equation (2) can be evaluated as Equation (3):

Equation (3) just formalizes the case of passing just π messages in polytree without evidence. So top-down belief updating in M1CBN without evidence can be performed

$$P(X_i = x_{i,j}) = \sum_{A \in D(\prod(X_i))} P(X_i = x_{i,j} \mid A) \prod_{X_m \in \prod(X_i)} P(X_m = A(X_m)) \tag{3}$$

in time linear in the size of the network.

Theorem 2. $G^*(X)$ forms an *X*-oriented polytree for every $X \in V$.

Proof. The proof of the proposition is constructive. According to the definition of ancestor graph, given node X, there is one and only one path from X to its parent Y. Likewise, there is an exclusive path from Y to its parent Z. So the exclusive path between X and Z is fixed. Iteratively following this way, we can get an exclusive path between X and any of its ancestors. Hence, $G^*(X)$ forms an X-oriented polytree.

Proposition 1 shows MAP is hard in M1CBN. Belief updating is a practically useful inference algorithm for approximating MAP for a number of reasons and has proven to be very effective and efficient in a variety of domains. Unfortunately, even belief updating is hard in M1CBN (shown in proposition 2). Based on theorem 1 and 2, a new belief updating algorithm restricted to M1CBN is described in the next section.

3 Belief Updating Algorithm in M1CBN

Based on the above discussion, local belief updating in M1CBN can be accomplished in two steps. *Step 1 Bottom-up propagation:* all nodes that have direct causal relations with X compose the ancestor tree of X. And as is proved in theorem 2, the ancestor tree is a polytree. So when a new evidence e triggers node X, belief propagation in $G^*(X)$ is the same as that in polytree [6]. *Setp 2 Top-down propagation:* the belief changes of nodes Y in step 1 will affect the likelihood of their children as effect. This step updates the believes of Y' s offspring using equation (3) while breadth first searching M1CBN. One note worth to mention is that belief updating in M1CBN propagates only twice and does not adopt iterative updating mechanism used by polytree, since most of causal influence can be evaluated in two propagations. Moreover, the simplified algorithm directly leads to polynomial performance. In algorithm BeliefUpdating N is a node in M1CBN, and each N has an evaluator J_N, which is responsible for evaluating the condition matching. To measure uncertain information, $J_N \in [50\%,100\%]$. All messages are initialized to 1.

Algorithm. BeliefUpdating(B, event$_i$)

```
Input: M1CBN B(V, E), hyper alert event_i
Output: updated believes in B, denoted as a vector Up-
datedBelief(aim:                                likelihood,
key_attack_step_1:likelihood,            …              ,
key_attack_step_i:likelihood,
key_attack_step_n:likelihood)
{
  If (node X is triggered by event_i) then {
    sign X as an observed node, and J_X = current belief
    of X,
    // Bottom-up propagation
    Breadth-first search G*(X) starting from X, for all
    node Y∈G*(X) do {
    //updating with Pearl's formulation[6]
      receive π_Y(U_i) from every Y's parent node U_i;
      receive λ_{C_i}(Y) from every Y's child node C_i;
       compute Belief(Y) and sign Y;
```

```
        compute λ_Y (U_i) for every Y's parent node U_i;
        compute π_Ci(Y) for every Y's child C_i;
      }
      // Top-down propagation
      Breadth-first search B, for any node N do
      If (the parent of N has been signed) then {
         update the believes of N with equation (3);
         sign N;
        }
      For i=1 to the number of key attack nodes of B do{
        UpdatedBelief [i]. key_attack_step= B[i].node;
        UpdatedBelief [i].likelihood = B[i]. belief;
      }
      Output UpdatedBelief;
    }
  }
```

4 Intrusion Plan Recognition Algorithm IPR

For raw alerts generated by NIDS, we aggregate and cluster them based on different srcIP or dstIP, and then prioritize them as [14]. The redundancy of resulting alerts is reduced, while important alert attributes retains. We denote each attack flow in the same cluster by a time series-based event vector Event($event_1$, $event_2$, … $event_i$, … , $event_n$), and each $event_i$ is called a hyper alert.

The complexity of IPR heavily relies on BeliefUpdating. BeliefUpdating propagates the diagnostic influence of ongoing evidence in its ancestor tree and thus reduces the problem's difficulty. In other wards, it tries to propagate causal influence as wide as possible. We denote by X the evidence node, by n the number of nodes in BN, and by m the average number of nodes in $G^*(X)$. As shown in [13], generally $m \prec\prec n$. If we measure the complexity of BeliefUpdating by the nodes it visits, then the complexity of Bottom-up propagation is O(m), and Top-down propagation is O(n). So the complexity of BeliefUpdating is O(m+n). Suppose the average number of matching Bayesian Network is k, the complexity of IPR is O(k(m+n)), which is polynomial in the size of attack graph.

Algorithm. IPR(Event)

```
Input: Event(event_1, event_2, … , event_i …)
Output: the aim of attack with the probability and
corresponding parameters, such as srcIP and dstIP.
{
  i=1;
  While (event_i is not NULL){
    Search attack plan library, and trigger BNs that
    include event_i;
    j = the number of triggered BN;
    for k=0 to j do{
    newBelief[k] ← call(BeliefUpdating(BN_k, event_i));
    If (newBelief[k].aim > threshold) then {
```

```
            output newBelief[k];
            output corresponding parameters, such as srcIP,
            dstIP, port and time;
            predict newBelief[k].aim as the aim of attack;
          }
        }
        i++;
      }
    }
```

5 Experiment

To evaluate the effectiveness of approximation made by these two algorithms, we applies them to LLDOS1.0, the first DDOS attack scenario created for DARPA to evaluate IDS. IPR is implemented on Monster3.0, a GNIDS developed by Southeast University. LLDOS1.0 includes 5 attack phases over the course of which the adversary probes, breaks in, installs trojan mstream DDoS software, and launches a DDoS attack against an off site server. Fig.2 illustrates the DDOS attack graph stored in Monster3.0, where the key parameters, such as CPT and prior probabilities, are listed on arcs and nodes respectively. Table 1 shows the hyper alerts of the attack flow in 172.16.115.0/24 after aggregating, clustering and eliminating redundancy, while the same alerts of other three subnets (172.16.112.0/24~ 172.16.114.0/24) are omitted for its clarity. In fact the alerts are far more than that listed in the table 1, for example, the hyper alert ICMP_PING_SWEEP represents 256 raw alerts. A denotes the aim DDOS, and B, C denote the other two key attack steps, Controlling a group of hosts and Launching attack respectively. In this case, Fig.2 and the alerts in column 2 are

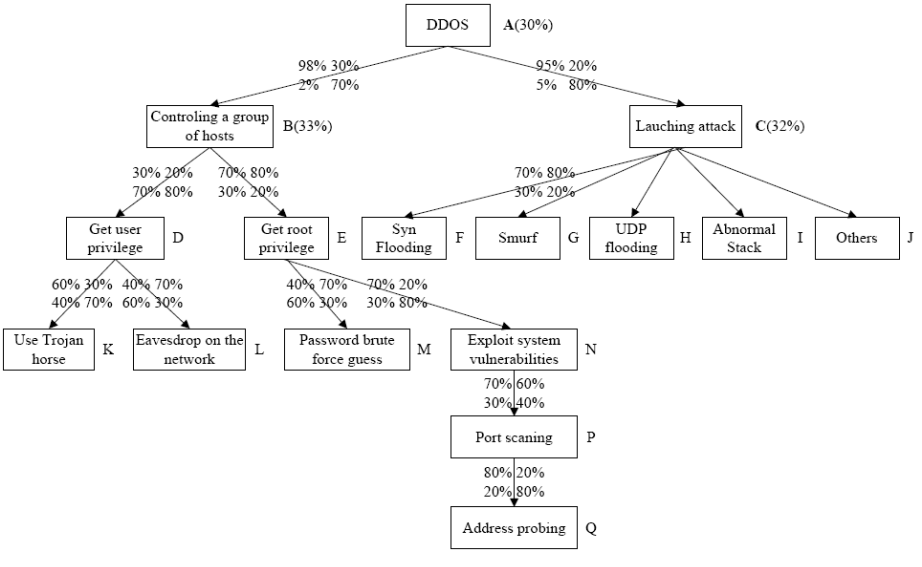


Fig. 2. Attack graph of DDOS

the input of algorithm IPR and the likelihood value of A, B and C are listed in each row as the output. The belief of the trigged node which is computed by its evaluator J, is listed in column 3. And the source IPs are 202.77.162.213, except for those that are spoofed and randomly generated by Syn flooding.

A clear attack track is shown in table 1 and believes of A, B and C increase stably in the course of attack. In the beginning, ICMP_PING_SWEEP comes, and the evaluators of node P and Q activate them with a probability of 100% (the adversary probes the subnet). The values of A, B and C suggests that ICMP_PING_SWEEP doesn't contributes much to DDOS, though the belief of A rises from 30% to 32.5%. Secondly, RPC_portmap_sadmind_request_UDP and RPC_sadmind_UDP_PING are probe steps to determine which hosts are running the remote administration tool, "sadmind". At this time, 172.16.115.20 is founded to be vulnerable. During the previous process, J_N gradually increase its belief from 80% to 100%, and A increases from 42.8% to 44.8% accordingly. In the third step, the adversary uses sadmind, buffer-overflow attack, to remotely break in 172.16.115.20, and three different stack pointer values are attempted, generating alerts RPC_portmap_sadmind_request_UDP, RPC_sadmind_query_with_root_credentials_attempt_UDP, and RPC_sadmind_UDP_NETMGT_PROC_SERVICE_CLIENT_DOMAIN_overflow_assttempt. When 172.16.115.20 responses the adversary by listing the directory (172.16.115.20 is sure to be conquered), J_E = 100% and A and B climb rapidly and reach 62.8%, 73.2% respectively. As soon as the real Syn flooding is launched, A soars to 80.4%, which is greater than threshold (70%), and then IPR outputs the attack vector as ((A, 80.4%), (B, 100%), (C, 80.9%)). From the circumstance variables, such as srcIp(172.16.115.20) and DstIp(131.84.1.31), one can conclude that the source IP of DDOS are not the forged ones appeared in the packets, but 172.16.115.20, which is remotely controlled by the adversary. In LLDOS1.0, 172.16.112.20 and 172.16.112.10 was conquered in the same way. So we can cut down the communication of these three hosts and prevent DDOS from happening.

Table 1. The experiment results

Dst IP	Hyper alert	Value of Evaluator	A(%)	B(%)	C(%)
172.16.115.0/24	ICMP_PING_SWEEP	$J_Q = 1$	32.5	50	32
		$J_P = 1$	41.8	40.4	42.6
172.16.115.0/24	RPC sadmind UDP PING	$J_N = 0.8$	42.8	59.5	43
172.16.115.20	RPC portmap sadmind request UDP	$J_N = 1$	44.8	65.3	46
172.16.115.20	RPC sadmind query with root credentials attempt UDP	$J_N = 1$	44.8	65.3	46
172.16.115.20	RPC_sadmind_UDP_NETMGT_ PROC_SERVICE_CLIENT_DOMAIN overflow attempt	$J_N = 1$	44.8	65.3	46
172.16.115.20	ATTACK-RESPONSES directory listing	$J_E = 1$	62.8	73.2	59.2
172.16.115.20	RSERVICES rsh root	$J_B = 1$	68.9	100	63.4
131.84.1.31	Syn flooding	$J_F = 1$	80.4	100	80.9

6 Conclusion

This paper proposes an algorithm IPR with polynomial complexity O(k(m+n)) to predict and recognize attack plan, which exceeds [4] in expressive power and performance.

The main improve relies on the following: 1) IPR broads the structure of attack plan depicted by Bayesian Network from polytree to max-1-connected Baysian Network, and thus expressive power becomes rich. 2) IPR gives up the iterative updating mechanism used by polytree, and adopts approximation to propagate causal influence as wide as possible. The approximation leads to polynomial performance and is effective as the experience shown. Moreover, IPR bears several advantages: firstly, it is able to detect multiple concurrent goals and partially ordered plan; Secondly, it has default reasoning ability and can deal with uncertain information. However, as a method based on predefined attack graph, it can not recognize unknown attacks. So how to recognize the new attack by correlation is a challenge in our future work.

Acknowledgments. This research is partially support by the National Basic Research Program (973 Program) No.2003CB314803, Jiangsu Province Key Laboratory of Network and Information Security BM2003201 and the Key Project of Chinese Ministry of Education under Grant No.105084.

References

1. Ning Zhuo, Gong Jian: A Survey on Network Intrusion Plan Recognition. Computer Science. 33(9) (2006) 4-6 (in Chinese)
2. Xinzhou Qin and Wenke Lee: Discovering Novel Attack Strategies from INFOSEC Alert. In Proceedings of The 9th European Symposium on Research in Computer Security (ESORICS 2004), Sophia Antipolis, France, September 2004.
3. Valdes A, Skinner K: Adaptive Model-based Monitoring for Cyber Attack Detection [EB/OL]. http://www.sdl.sri.com/projects.emerald/adaptbn-paper/adaptbn.html
4. Wenke Lee, Xinzhou Qin: Attack Plan Recognition and Prediction Using Causal Networks, In Proceedings of the 20th Annual Computer Security Applications Conference (ACSAC 2004), Tucson, Arizona, December 2004
5. F.J.Diez: Local conditioning in Bayesian Networks. Artificial Intelligence 87(1-2) (1996)1-20
6. J. Perl: Probabilistic Reasoning in Intelligent Systems. Networks of Plausible Inference, Morgan Kaufmann Publishers, Inc.San Mateo, CA, 1988
7. F.V. Jensen, K.G.Olsen, S.K.Andersen: An Algebra of Bayesian Belief Universes for Knowledge-Based Systems. Networks. 20(1990)637-660
8. S.L. Lauritzen, D.J. Speigelhalter: Local Computations with Probabilities on Graphical Structures and Their applications to Expert Systems, J. Royal Statist. Soc. 50 (1998)157-224
9. G.F. Cooper: The Computational Complexity of Probabilistic Inference Using Bayesian Belief Networks. Artificial Intelligence 42(2-3) (1990)393-405
10. Y. Peng, J.A. Reggia: A Probabilistic Causal Model for Diagnostic Problem Solving (parts 1 and 2). IEEE Trans. Systems Man Cybernet. SMC-17(1987)146-162,395-406
11. S.E. Shimony: Finding MPE&MAPs for Belief Networks is NP-hard. Artificial Intelligence 68(2) (1994)399-410
12. James D. Park, Adnan Darwiche: Complexity Results and Approximation Strategies for MAP Explanations. J. Artif. Intell. Res. 21(2004)101-133
13. A. Drwiche: Recursive Conditioning. Artificial Intelligence (Special Issue on Resource Bounded Reasoning). 125 (1-2) (2000)5-41
14. X. Qin and W. Lee: Statistical Causality Analysis of Infosec Alert Data. In Proceedings of the 16th International Symposium on Recent Advances in Intrusion Detection(RAID 2003)

Impact of Buffer Map Cheating on the Streaming Quality in DONet*

Yong Cui, Dan Li, and Jianping Wu

Department of Computer Science, Tsinghua University, Beijing, 100084, China
cy@csnet1.cs.tsinghua.edu.cn, lidan@csnet1.cs.tsinghua.edu.cn,
jianping@cernet.edu.cn

Abstract. Data-driven Overlay Network (DONet) is especially suitable for live stream in P2P environment. However, it demands the cooperation of individual nodes to exchange their buffer maps. Since these nodes are selfish and have their own interests, they might cheat about their buffer maps to reduce the forwarding burden. We analyze the impact of this kind of cheating behavior on the streaming quality of DONet by experiment in this paper. The experimental results show that buffer map cheating has considerable negative impact on the streaming quality in DONet.

Keywords: DONet, Buffer Map Cheating, Selfish.

1 Introduction

Multimedia applications, especially live stream, have been used more and more in Internet. In live stream, a large number of users are interested in the real- time data from a common source. Compared to other applications, live stream demands higher network bandwidth as well as node forwarding capacity. Given the multi-receiver nature of live stream, multicast is the ideal supporting technology. Currently, there are two kinds of multicast technologies. One is realized in the network layer, named as IP multicast [1]; and the other is realized in the application layer, named as overlay multicast [2~12].

IP multicast builds the data structure on routers, which is a tree, and thus achieves high scalability and high efficiency. However, IP multicast changes the "unicast" principle of traditional Internet, and a lot of problems in it, such as member management, congestion control, and pricing model, have not been solved well yet. All these lead to the difficulty of deploying IP multicast in Internet scale.

Overlay multicast is subsequently proposed, which constructs the data structure in the application layer. Compared with IP multicast, overlay multicast is lower in efficiency, but much more deployable and flexible. Currently, there are two kinds of overlay multicast, namely, tree-based overlay multicast [2~7] and mesh-based overlay multicast [8~12]. Like in IP multicast, the data structure in tree-based overlay multicast is also a tree, and data is propagated along the tree after its establishment. However, overlay nodes are unstable. There is a high frequency of node join, node leave, and

* Supported by the National Natural Science Foundation of China (No. 60473082, No. 90104002).

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 817–824, 2007.
© Springer-Verlag Berlin Heidelberg 2007

node crash in overlay network. Thus, the application-layer multicast tree might change from time to time, which will bring negative impact on live stream applications that have stringent demands on the streaming continuity.

Mesh-based overlay multicast is considered as a better choice to support live stream. In mesh-based overlay multicast, the data structure is no longer a tree, but a mesh. Data-driven Overlay Network (DONet) is a representative of this kind of protocols [8]. In DONet, the stream propagated in the overlay network is divided into multiple segments. Each node maintains a number of segments, and exchanges the buffer map of available segments with partner nodes. After learning the buffer maps of partner nodes, each node requests a certain segment from a suitable partner node that holds the segment. If a node receives segment requests from partner nodes, it replies to the requests by forwarding the corresponding segments within its outgoing bandwidth. Therefore, the leave or crash of a single node will not bring too much impact on other nodes.

However, overlay nodes are not only unstable, but also selfish and strategic. Selfish nodes in DONet might cheat about their buffer maps to reduce the forwarding burden to other nodes. In such non-cooperative scenarios, some nodes cannot have the actual information to request segments, and thus the streaming quality in DONet might be impact. We try to establish the model of buffer map cheating in DONet and analyze its impact on the streaming quality by experiment in this paper, which is not studied by previous work.

2 Related Work

Because of the difficulty of deploying IP multicast in Internet scale, researchers turn to overlay multicast to support live stream applications. The overlay multicast protocols proposed currently can be roughly classified into two categories, namely, tree-based overlay multicast and mesh- based overlay multicast.

In tree-based overlay multicast, each node selects a long- time parent from other participating nodes to receive stream data. The parent/children relationships among all nodes compose the data structure, i.e., the multicast tree. Once the multicast tree is established, data is propagated along the tree and there is no additional control overhead. Protocols belonging to this category of overlay multicast include NARADA [2], NICE [3], ZIGZAG [4], Scattercast [5], Yoid [6], Host Multicast [7] and etc.

In mesh-based overlay multicast, there is no explicit parent/ children relationship. The data structure to propagate the stream is a mesh. Therefore, it can tolerant node dynamics well and is especially suitable to support live stream applications. Representatives of this kind of protocols include gossip-based protocols and DONet. In gossip-based protocols [9~12], each node forwards available data to a set of randomly selected nodes. But in DONet [8], each node maintains several partner nodes, and data is transmitted among partner nodes, eventually to the whole overlay network. Compared to gossip-based protocols, the advantage of DONet is that data is flowing in a request-reply way, thus there is no redundant data consuming the precious network bandwidth.

An important characteristic of overlay network is that the overlay nodes are all selfish and strategic. The selfish overlay nodes might cheat about their private information to obtain higher interests, which could affect the performance of the overall

system. Selfish nodes in DONet might cheat about their buffer map to reduce the forwarding burden. The impact of this kind of cheating on the streaming quality in DONet is not systematically studied before, and it is discussed in this paper.

3 Model of Buffer Map Cheating in DONet

The stream propagated in DONet is divided into multiple segments with uniform length. A buffer map can represent the information of available segments on a node. Each node periodically exchanges its buffer map with partner nodes, and decides from which partner node to fetch a certain segment. If there are multiple partner nodes holding the same expected segment, various ways can be chosen to select the segment-providing node, for instance, the one with the shortest distance or the one with the highest outgoing bandwidth. The requests arrive at the requesting queue of the segment- providing node. When replying the requests, the segment-providing node selects some requests in the requesting queue and sends the corresponding segments within its outgoing bandwidth.

At first, we make some definitions and list the notations throughout this paper as Table 1.

Table 1. List of notations

B	Buffer window size
W	Playing waiting segment number
N	Number of nodes in the overlay network
M	Number of partner nodes each node maintains
Q	Number of segments of the stream
P	Cheating node percentage
H	Cheating degree
$Pc(i)$	Playing continuity on node i
$Pd(i)$	Playing delay on node i

Each node in DONet maintains a buffer window, which is used to store a number of continuous segments of the stream. The *buffer window size* of each node is denoted as B. To make the streaming more continuous, each node does not begin to play the stream immediately after receiving the first segment. Instead, it waits for the arrival of the first several segments and then begins to play the stream. The number of segments each node waits for before playing the stream is defined as the *playing waiting segment number*, denoted by W. Obviously, there is $W \leq B$.

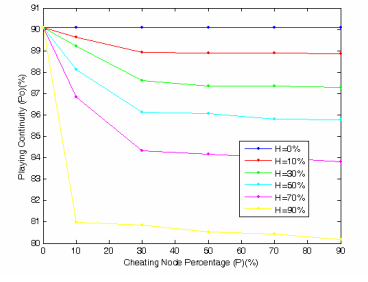
The streaming quality of each node can be measured by *playing continuity* and *playing delay*. The playing continuity of node i is defined as the number of segments arriving at node i no later than its playback time over the total number of segments of the stream, denoted as $Pc(i)$. And the playing delay of node i is defined as the average source-to-end delay of all segments playing on node i. Higher playing continuity and less playing delay indicate better streaming quality. The playing continuity of the DONet is defined as the average playing continuities of all nodes, and the playing delay of the DONet is defined as the average playing delays of all nodes.

We can establish the model of buffer map cheating in DONet as follows. Suppose the number of nodes in the overlay network is N, the number of partner nodes each node maintains is M, the number of segments of the stream propagated in the overlay network is Q, the buffer window size is B, the playing waiting segment number is W, and the percentage of the number of cheating nodes over the total number of overlay nodes is P. When exchanging the buffer map with partner nodes, each cheating node hides some of its available segments, and the proportion of the number of hidden segments over the total number of available segments is H. This model of buffer map cheating in DONet is denoted as $C(N, M, Q, B, W, P, H)$, where $N>0$, $M>0$, $0<W\leq B\leq Q$, $0\leq P\leq 1$, and $0\leq H\leq 1$.

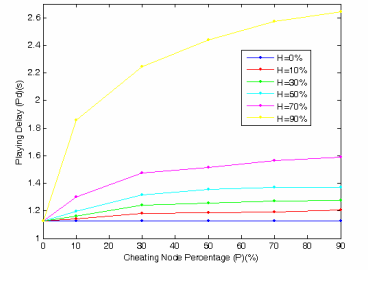
4 Experiment Study

We conduct experiments to study the impact of buffer map cheating on the streaming quality of DONet, using the cheating model $C(N, M, Q, B, W, P, H)$ established in the last section.

Suppose the number of overlay nodes in DONet is 500 (N=500), and the stream propagated in the network is composed of 5000 segments (Q=5000). If no node cheats about their available segments, the cheating node percentage is 0%. We compare the average playing continuity (Pc) and the average playing delay (Pd) of all nodes when



(a)



(b)

Fig. 1. Streaming quality of DONet when M=4, B=70, W=30. (a) playing continuity; (b) playing delay.

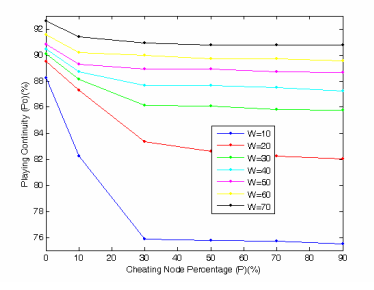
the cheating node percentage (P) is 0% with that of 10%, 30%, 50%, 70%, and 90%, each under different cheating degrees (H), different playing waiting segment numbers (W), different buffer window sizes (B), and different partner numbers (M).

1) Different Cheating Degree

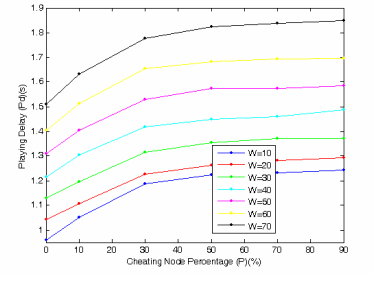
Let M=4, B=70, W=30, and the cheating degree H vary as 0%, 10%, 30%, 50%, 70%, and 90%. The playing continuity and the playing delay in these cases with different cheating node percentages are shown in Fig. 1(a) and Fig. 1(b), respectively.

From Fig. 1(a), we see that under all cheating degrees higher than 0%, the maximal playing continuity is achieved when there is no node cheating. When the cheating node percentage increases, the playing continuity decreases, but the decreasing gets smoother. We also see that the playing continuity is higher with lower cheating degree. When the cheating degree is 0%, which means that there is no node cheating, the playing continuity gets the optimal value. The change of playing continuity is especially steep when the cheating node percentage is low and the cheating degree is high.

Fig. 1(b) shows that under all cheating degrees more than 0%, the playing delay is least when there is no node cheating. As the node cheating percentage grows, the playing delay becomes more. In addition, the lower is the cheating degree, the less is the playing delay. It is similar to Fig. 1(a) that the change of playing delay is especially obvious when the cheating node percentage is low and the cheating degree is high.



(a)



(b)

Fig. 2. Streaming quality of DONet when M=4, B=70, H=50%. (a) playing continuity; (b) playing delay.

2) Different Playing Waiting Segment Number

Let M=4, B=70, H=50%, and the playing waiting segment number vary as 10, 20, 30, 40, 50, 60, and 70. The playing continuity in these cases under different cheating node percentages is shown in Fig. 2(a), and the playing delay is shown in Fig. 2(b).

Fig. 2(a) illustrates that no matter what the playing waiting segment number is, the playing continuity is maximal when there is no node cheating. The playing continuity decreases with the growth of cheating node percentage, but the curve becomes smoother. Additionally, more playing waiting segment number brings higher playing continuity.

From Fig. 2(b), we conclude that the playing delay increases when the cheating node percentage grows, and is shortest when there is no node cheating, no matter the playing waiting segment number. Also, when the playing waiting segment number is more, the playing delay is longer.

3) Different Buffer Window Size

Let M=4, W=30, H=50%, and the buffer window size vary as 30, 40, 50, 60, 70, 80, and 90. With different cheating node percentages, the playing continuity and the playing delay are shown in Fig. 3(a) and Fig. 3(b), respectively.

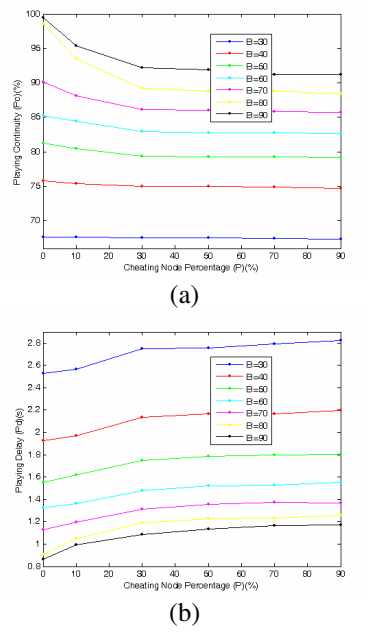


Fig. 3. Streaming quality of DONet when M=4, W=30, H=50%. (a) playing continuity; (b) playing delay.

Fig. 3(a) suggests that with all buffer window sizes, the playing continuity is highest when there is no node cheating, and gets lower with more cheating nodes. In addition, given the same cheating node percentage, the bigger buffer window size brings higher

playing continuity. The change of playing continuity is most obvious when the buffer window size is big and the cheating node percentage is small.

Fig. 3(b) tells that the playing delay increases with the growth of cheating node percentage, and decreases with the growth of buffer window size.

4) Different Partner Number

Let *B*=70, *W*=30, *H*=50%, and the partner number of each node vary as 4, 5, and 6. The playing continuity and the playing delay under different cheating node percentages are shown in Fig. 4(a) and Fig. 4(b), respectively.

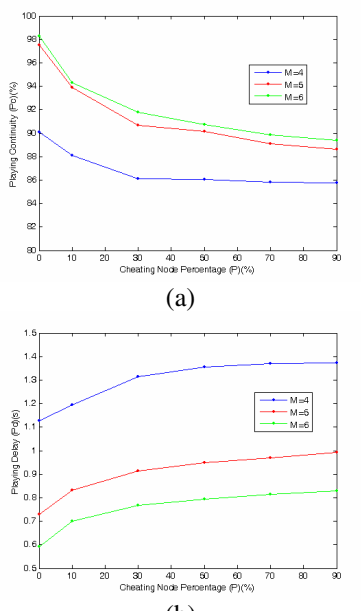


Fig. 4. Streaming quality of DONet when B=70, W=30, H=50%. (a) playing continuity; (b) playing delay.

With any number of partner nodes, the playing continuity decreases with the increase of cheating node percentage, and is maximized when there is no node cheating, as shown in Fig. 4(a). This is consistent with Fig. 1(a), Fig. 2(a), and Fig. 3(a). In addition, Fig. 4(a) shows that the playing continuity is higher when the number of partner nodes each node maintains is bigger.

On the other hand, Fig. 4(b) tells the similar conclusion as Fig. 1(b), Fig. 2(b), and Fig. 3(b), that the playing delay increases with the growth of cheating node percentage, and is shortest when there is no node cheating. Also, more partner nodes will bring shorter playing delay.

According to all the experimental results above, the streaming quality is always optimized when there is no node cheating, no matter what the network parameter is. Therefore, the buffer map cheating of selfish nodes indeed has considerable negative impact on the streaming quality of DONet.

5 Conclusion and Future Work

DONet is especially suitable for live stream applications. However, selfish overlay nodes might cheat about their buffer maps to reduce the forwarding burden. We establish the model of buffer map cheating in this paper, and analyze its impact on the streaming quality of DONet by experiment in this paper. The experimental results show that buffer map cheating of selfish nodes will indeed cause the streaming quality in DONet to decrease considerably.

As for future work, we should design incentive algorithms to defend buffer map cheating in DONet, and it can be reviewed as another way to improve the streaming quality in DONet considering the selfishness of individual nodes.

References

1. S. E. Deering: Multicast Routing in Internetworks and Extended LANs, In *Proceedings of ACM SIGCOMM'88*, Stanford, CA, USA, (Aug 1988)
2. Y. H. Chu, S. G. Rao, and H. Zhang: A Case for End System Multicast, In *Proceedings of ACM Sigmetrics'00*, Santa Clara, California, USA, (Jun 2000)
3. S. Banerjee, B. Bhattacharjee, C. Kommareddy: Scalable Application Layer Multicast, In *Proceedings of ACM SIGCOMM'02*, Pittsburgh, PA, USA, (Aug 2002)
4. D. A. Tran, K. A. Hua, T. Do: Zigzag: An efficient peer-to-peer scheme for media streaming, In *Proceedings of IEEE INFOCOM'03*, San Franciso, CA, USA, (Mar/Apr 2003)
5. Y. Chawathe: Scattercast: An Architecture for Internet Broadcast Distribution as an Infrastructure Service, *Ph.D. Thesis*, University of California, Berkekey, (Dec 2000)
6. P. Francis: Yoid: Extending the Internet Multicast Architecture, *White Paper*, http://www.icir.org/yoid
7. B. Zhang, S. Jamin, L. Zhang: Host Multicast: A Framework for Delivering Multicast to End Users, In *Proceeding of IEEE INFOCOM'02*, New York, NY, USA, (Jun 2002)
8. X. Zhang, J. Liu, B. Li, T. P. Yum: Data-Driven Overlay Streaming: Design, Implementation, and Experience, In *Proceedings of IEEE INFOCOM'05*, Miami, Florida, USA, (Mar 2005)
9. S. Banerjee, S. Lee, B. Bhattacharjee, A. Srinivasan: Resilient multicast using overlays, In *Proceedings of ACM SIGMETRICS'03*, San Diego, CA, USA, (Jun 2003)
10. P. Eugster, R. Guerraoui, A.M. Kermarrec, L. Massoulie: From epidemics to distributed computing, *IEEE Computer*, (2004) 37(5):60-67
11. S. Jin and A. Bestavros: Cache-and-relay streaming media delivery for asynchronous clients, In *Proceedings of NGC'02*, Boston, MA, USA, (Oct 2002)
12. Y. Cui, B. Li, K. Nahrstedt: oStream: asynchronous streaming multicast, *IEEE Journal on Selected Areas in Communications*, (2004) 22(1): 91-106

Architecture of STL Model of New Communication Network

Aibao Wang[1,2] and Guangzhao Zhang[1]

[1] Zhongshan University
[2] China Telecom
wangab@chinatelecom.com.cn, isszgz@mail.sysu.edu.cn

Abstract. New communication network taking Internet as representative is being used extensively, but traditional communication theory cannot give good explanations and instructions on these new communication networks. Taking anthropotomy as a reference model, we analyze the systematic integration and functional requirements of new communication network with Internet as its representative in this paper. Further, we put forward the STL model of new communication network. Through the illustration for STL model, we have new knowledge on the Internet-representative communication network, which has certain use for the development of related theory and technologies.

Keywords: Communication theory, network architecture, next generation network.

1 Introduction

Tracing back the development of communication technology, it has been more than 20 years since the digital technology acquired its application in communication field, especially the fast development of satellite communications, fiber optical communications, and mobile communications made communication much more easily in people's daily work and life. However, comparing with the extensive application and involvement in social life, the development of communication theory appears to be standing still for relevant communication theory cannot keep up with the development of communication technology.

As for the definition of communication, generally speaking, the traditional one is "the process of realizing the effective message transferring (or exchanging) from one site to another (or multi-site), using technologies like electromagnetism, photoelectricity, etc." But in recent years, the development of communication technology, especially the development of Internet and mobile communication technologies, has made the definition of communication in reality greatly surpass the definition of communication in communication theory.

Firstly, the purpose of communication is not only to transmit certain message, but also to take on more extensive bound, for example, telephone chat, joke transmitted by short messages, Internet games, and current popular BLOG. Strictly speaking, they are all kinds of communication, but the purpose of communication is entertainment or for the need of personal self-express.

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 825–832, 2007.
© Springer-Verlag Berlin Heidelberg 2007

Secondly, seeing from the obstacle communication can surmount, traditional communication only pays attention to the area crossing communication skills, but does not take two obstacles of spanning time and language into consideration; but seeing from the application of existing Internet, surmounting two obstacles of time and language has been possible. For example, EMAIL, which surmounts the time limit and realizes the intercommunication between two nodes at EMAIL sending and receiving time; and Text-To-Speech (TTS) technology based communication between text media and speech media has realized the communication between different languages. What's more, with the development of Internet technology and application, the expressing method has taken on diverse trends. Speech communication is no more the only way for people's daily communication. Text communication (message communicating and chat on-line), and video communication began to involve in people's daily life.

Therefore, the traditional definition of communication can not cover the ability possessed by real communication anymore. In order to instruct the development of communication technology and application effectively, communication science is in dire need of reknowing and redefining the definition of communication.

Since the appearance of Internet, especially the popular applications such as: WEB/Email/IM/BLOG, etc., people have realized that the Internet is in fact a communication network which can provide more abilities. The open character of Internet makes everyone have the opportunity to participate in the creation for all kind of contents and applications; and the abundant computing abilities together with the great capacity memory power promotes the Internet to explode vast intelligence application, which can be independent of people's intervene to Auto-Run. Therefore, Internet-representative new communication network is no longer a simple communication system, but a smart and elaborative intelligent processing system, (for convenience, new communication network refers to Internet-representative new generation network in this paper), its service object does not limit to the speech among people, but in a greater degree involves to all layers of social life. The intelligence of new communication network mainly presents the following aspects.

Taking anthropotomy as reference model, the purpose of this paper is to analyze the systematic integration and functional requirements of new communication network with Internet as its representative, and put forward the STL model and its framework consideration. Through the illustration for STL model, we have a whole new knowledge on Internet-representative new generation communication network, which has certain use for the development of relative theory and technology.

2 Related Work

Research on network communication principle and architecture has been the hotspot of academe for a long time. The famous expert in computer network D. Clark gave the object for development of Internet and main design principle based on "Edge Theory" in [1], which established basic architecture and developing direction for the Internet. In 1980s, standardization has been the masterstroke for the development of network theory and technology.

Internet is a successful example of communication network, but currently, people still cannot establish a theory model for complex and giant system as Internet. Because

of the continuous development of computer's traffic type, traditional mathematic model of network behavior such as Markov modulated Poisson Process and so on cannot reflect the real network behavior anymore. In recent years, relative researches domestically or abroad are all trying out new mathematic model reflecting certain real features of network, such as self-similarity and long-range dependence model, etc., through the analysis and research on a great deal traffic data. But due to the high burst and randomcity of modern network information transmission, up to now, satisfactory theory and model which can reflect network statement and behavior from all aspects are not found yet.

With people's increasing need for the Internet, many new thought and technology are introduced to Internet-relative theories and practices. Existing IPv4 protocol has the problem of inadequate address space. In order to provide more address space, IPv6 protocol became the IETF warmly recommended network layer protocol of new generation Internet [3]. IPv6 has integrated the IPSec security protocol, which can solve the problem of network security to some extend. Now, America, Europe Union, and Japan are all vigorously supporting the construction of IPv6 backbone network. IPv6 network already or being constructed include CERNET2 of our country, Abilene [4] of America, GEANT [5] of Europe Union, and APAN [6] of Asia-Pacific, etc.

Researchers have put forward a lot of new Internet architecture, such as NGI [7], NewArch [8-10], FIND [11], GENI [12], NSFCNET [13], etc. They are trying to design a new Internet which can better satisfy people's need in network communication. There are also some proposals to improve the current Internet architecture, such as IPNL [14], ROFL [15], NIRA [16], I3 [17], TRIAD [18] and etc. However, up to now, no researchers give better explanations and instructions on Internet-representative new communication network from the perspective of communication theory, which is also the main research purpose of this paper.

3 New Communication Network System

New communication network is a smart intelligence system and what kind of components and functions the intelligent system should possess? Now, we take anthropotomy as model theory of new communication network system, and take new communication as a "Human Body System" to anatomize, analyze, and research.

Before analyzing the new communication network system, let us first have a look at the components of human body. From the perspective of anthropotomy, human body consists of skeleton, muscle, nerve, cerebra, heart, blood vessel, and all other kinds of internal organs. Human as a special animal, its purpose of existence is fundamentally the same with other animals, which is "act", but people endow "act" with more meanings (e.g. labor, entertainment, sport, travel, etc). The mainly basic apparatus in human body system which realize "act" are muscle, skeleton and nerves. Therefore, view from the 'act' function, muscle, skeleton and nerves are the fundamental components of the human body system. And all other organs work together as the human body support-system. As a normal man, in order to maintain the 'act' function work well, his support-system such as heart, lungs etc. must be in powerful status.

When we take new communication as an intelligent "human body system" to analyze, the new communication system consist of following parts: network

architecture(signal transmission subsystem), information storage module, information expression module, resource management subsystem, power supply subsystem, application servers, operation supporting subsystem, security protection subsystem, etc. As new communication network taking "communication" as its basic function, its three basic parts are: network infrastructure, storage module, and information expressing module. Among them, network infrastructure refers to circuit transmission and switching system such as communication termination, access network, backbone network, and core network. Storage module includes not only independent information storage subsystem, but also attached storage equipment in network and termination equipments. And information expressing module mainly refers to all kinds of communication terminations, it also include internal software parts of equipment in network nodes for information recognizing and switching. possess the various information expressions and switching abilities from different languages.

As for other parts in the new communication network system, which are similar to the various supporting organs in human body system, belong to OSS(Operation Support System) which serve for the fundamental function of main body in communication system to help and make sure the network can provide 'communication services' with abundant contents continuously and smoothly.

So, from the viewpoint of 'human body system', the new communication network system should have:

- An architecture which realize the fundamental function of the network system. But the fundamental function 'Communication' now have more meaning than the traditional definition. More detailed about the architecture meaning will show in chapter4.
- An OSS which should run effectively and healthy in line with the architecture of the main body of the new communication network in order to guarantee the total network system run well.

4 STL Model Design Adapted to New Communication Network System

We may see from above, the new communication network realized communication content has surpassed the defined traditional communication in existing communication theory; at the same time, from the anatomy point of view, current Internet still cannot completely possess the function and corresponding components which new communication expected, for example, security and antivirus subsystem, resource operation and management subsystem , etc.

This has brought forward such a proposition for us: how should new communication network infrastructure be like? How should existing communication system theory evolve to guide the need for the continuing development of new communication network?

Trying to answer the above questions, we put forward the STL model of new communication network infrastructure on the basis of analyzing three direct correlative abilities in accomplishing new communication function, namely, network infrastructure, information storage unit, language expressing and switching unit.

4.1 STL Model for New Communication

We know from the function analysis needed in new communication network of the previous chapter that refresh understanding new communication network need surmount not only the geographic limitation (existing communication only consider the problem of space domain, and realizing the communication among users in different geographic places), but also the limitation of time domain and language domain, realizing the communication among communication bodies of different places, different time and different language expressions(Fig. 1). On mathematic model, we can take it as a three-domain (three variables) communication model, taking the three initials of words SPACE, TIME, LANGUAGE; we call the new communication network infrastructure model as STL model.

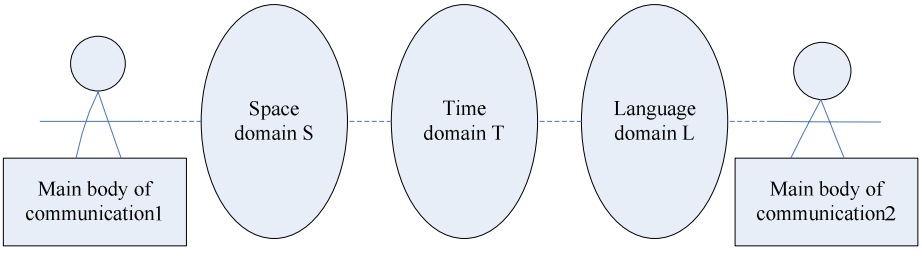


Fig. 1. Illustration of STL model of new communication network

The relationship of three–domain variables needed by communication network in fig.1 is serial interconnection, which is consistent with the space domain, time domain, and language domain where any communication process will come through in reality. In the fig.1, space-domain as "body (network infrastructure)" mainly undertakes the function of multicast delivery and multicast switching; time-domain as information storage section is mainly responsible for the information transmission on time latitude; and language-domain as the information expressing and computing section, is mainly responsible for the semantics recognition, expression and translation.

Because two latitudes are added on the basis of space domain, comparing with the traditional communication network, there will be a substantial jump change on the function of the new communication network. For example, after adding the time–domain, new communication network can realize the communication between present and future or past; after adding the language-domain, new communication network can not only realize the communication and conversation among different language group, but also realize the communication between people and machine, and even people and animals or plants.

Then we begin the analysis on the output result of several combine scenes of STL model in mathematic perspective.

Space-domain can be denoted as vector S, S=[S(1), S(2), …, S(k), …, S(n)], among which, k(1,2,…n) denotes the different geographic positions in the space and S(k) we called it position factor corresponding the output intensity of signal, for convenience, under default statement, S(k)=1 (k=1,…n).

Time-domain can be denoted as variable T, T=T (k), denoting time delay when in the position of k (time factor for short);

Language domain is denoted as variable L, L=L(k), denoting the language expression when in the position k (semantic factor);

Now we suppose a message x was input into the communication system, considering a communication process must pass through the three domains of communication network, namely, space domain, time domain, and language domain, thus the output message Y is:

$$Y = S.T.L.x$$
$$= [S(1).T(1).L(1), S(2).T(2).L(2), \ldots, S(k).T(k).L(k) \ldots, S(n).T(n).L(n)].x$$

After a message x enters the STL model network, there will be many output messages, the finite time, geographic position, and semantic expression used can all be different. Following are eight possible message output scenes.

1) Point-to-point real-time communication mode, adopting the same language expression.

At this time, n=1, S (1) =1, T (1) =1(no time delay) , L (1) =1(the same language). Thus Y=S (1).T (1).L (1) =x, i.e. the message output end acquired is consistent with the input end. Real scene corresponds to the point-to-point real-time communication mode like PSTN, instant messenge, etc. In these scenes, both sides have the on-line real-time communication with each other, adopting the same language.

2) Point-to-point real-time communication mode with output and input adopting different language expressions

At this time, output Y=L(1).x. i.e. output message is the product of original language message and semantic factor L(1). Real scene corresponds to phone-calling or text chating existing Synchronized Translation or Text-To-Speech function.

3) Point-to-point non-real-time communication mode, adopting the same language.

At this time, output Y=T (1).x. i.e. output message is the product of original language message and time factor T (1). Real scene corresponds to communication mode adopting EMAIL sending process. The message from sender will be received by the accepter after time T (1).

4) Point-to-point non-real-time communication mode with output and input adopting different language expressions.

At this time, output Y=T (1).L (1).x, i.e. output message is the product of original language message, time factor T (1), and semantic factor L (1). Real scene is similar to WEB visiting, which is when the visitor acquired the information of website; auto translation software will switch its webpage content to languages or characters he or she can understand.

5) Point-to-multipoint real-time communication mode, adopting the same language expression.

Thus, the output of multi-position point is: Y1=X, Y2=X,…, Yn=X. Real scene corresponds to the conference meeting scene where people communication by the same language.

6) Point-to-multipoint real-time communication mode with input and outputs adopting different language expressions.

At this time, Y1=L (1).x; Y2=L (2).x; …; YN=L (n).x. Real scene corresponds to the conference meeting scene where listeners need Synchronized Translation to understand the lecturer.

7) Point-to-point non-real-time communication mode, adopting the same language expression.

Thus, the output of multi-position point is: Y1=T (1).x; Y2=T (2).x; …; YN=T (n).x. Real scene corresponds to EMAIL send to group-accepter or WEB visited by multi-people, the reading time is delayed differently for each accepter or visitor.

8) Point-to-multipoint non-real-time communication mode with input and outputs adopting different language expressions.

Thus, the output of multi-position point is: Y1=T (1).L (1).x; Y2=T (2).L (2).x; …; YN=T (n).L (n).x. Real scene corresponds to webpage visiting of WEB mode (the webpage support the translation function among multi-language), the website visiting finite-time of everyone is different, and the semantic may be different, too.

4.2 Consideration of the STL Mode Based Network Architecture

Limited by the paper volume, the solution of the STL mode will not be revealed in this paper. Here we give some consideration of the solution when we do the STL network architecture design.

1) the objective of the STL mode is to provide a communication theory or mode which can explain the Internet representative communication network, and from the STL mode, we hope to find more value(good business mode) and more application from the new network.
2) STL mode based network not only should realize the 3-domain communication functions, but also should provide a perfect OSS. By the OSS, the STL network will not only solve the QOS and Security issues which are the hot topic in the current days, but also, will give the STL network intelligence, reliable and continuous developmental functions etc.
3) STL network should run on the top of IP4 base Internet while evolve a intelligent terminal driver. Among them, the solution for the space-domain architecture will refer to the principle of the Content Switch Network architecture[19]. And the architecture for time-domain and Language-domain will require new design.

5 Conclusion

The Internet has contributed a lot to human-being since its first appearance, but the traditional communication theory cannot give a satisfactory explanation to the Internet-representative new communication network. This paper analyzes the systematic components and function requirement of Internet-representative communication network, and on the basis of which brings forward a STL model and its interface of new communication network. Through the description of STL model, we have a whole new knowledge of Internet-representative new communication network, which is helpful to the development of communication science and relative technologies, and at the same

time provides a road of long-range development for network and service providers. Under STL model, network providers will not only go on providing traditional voice communication service, but also provide more fundamental service, like multi-point switching, storage, speech translation, and so on. While the service providers will have more abilities from the network to create more new creative application for the society, then make the maximum value of the new communication network.

Reference

[1] D. Clark, "The Design Philosophy of the DARPA Internet Protocols", In Proceedings of ACM SIGCOMM 1988
[2] "Internet Performance Measurement and Analysis", http://www.merit.edu/ networkresearch/projecthistory/ipma/index.php? Printvs=1
[3] "IP Version 6 Working Group (ipv6)", http://www.ietf.org/html.charters/ipv6-charter.html
[4] "Advanced Networking for Leading-edge Research and Education", http://abilene. Internet2.edu/
[5] "The GÉANT project", http://www.geant.net/
[6] "Asia-Pacific Advanced Network", http://www.apan.net/
[7] "Next Generation Internet", http://www.ngi-net.de/
[8] "NewArch project: future-generation Internet architecture", http://www.isi.edu/newarch/
[9] D. Clark et al, "Tussle in Cyberspace: Defining Tomorrow's Internet", In Proceedings of ACM SIGCOMM 2002
[10] D. Clark et al, "Addressing Reality: An Architectural Response to Real-World Demands on the Evolving Internet", In Proceedings of ACM SIGCOMM 2003 Workshops
[11] "FIND: future Internet network design," http://find.isi.edu
[12] "GENI: global environment for network innovations," http://www.geni.net
[13] "NSFCNET", http://www.nsfcnet.net/
[14] P. Francis, R. Gummadi, "IPNL: a NAT-extended Internet architecture," ACM SIGCOMM,'02 Aug 2002.
[15] M.Caesar, T. Condie, J. Kannan et al "ROFL: Routing on Flat Labels", ACM SIGCOMM'06, Aug 2006
[16] X. Yang, "NIRA: a new Internet routing architecture," SIGCOMM Workshop on Future Directions in Network Architecture (FDNA), Aug 2003.
[17] I. Stoica, D. Adkins, S. Zhuang, S. Shenker, S. Surana, "Internet indirection infrastructure," ACM SIGCOMM, Aug 2002.
[18] D. Cheriton, M. Gritter, "TRIAD: a scalable deployable NAT-based Internet architecture," Technical report, Jan 2000.
[19] Aibao Wang□"The framework of content switch network oriented to the new generation Internet', 2004.9□Telecom Science□□

Experience with SPM in IPv6

Mingjiang Ye, Jianping Wu, and Miao Zhang

Department of Computer Science, Tsinghua University, Beijing, 100084, P.R. China
yemingjiang@csnet1.cs.tsinghua.edu.cn
{zm,jianping}@cernet.edu.cn

Abstract. The lack of source IP address checking makes it easy for the attackers to spoof the source address. Spoofing Prevention Method (SPM), as a prospective candidate to be deployed in the Internet, is a newly proposed scheme to solve this problem. However, there is no work on SPM prototype system. In this paper, we present our experience in achieving a prototype system for SPM. In addition to realizing the basic idea of SPM described in the original paper, we make the following three contributions: First, the detail design is made for the whole SPM system architecture and detail mechanisms. Second, several important issues for SPM system are addressed, e.g., how to carry the key required by SPM, the MTU problem, etc. Third, a prototype system is made and some experiments are done with this prototype system.

Keywords: Source Address Spoofing, Spoofing Prevention, Security.

1 Introduction

The Internet is a decentralized system which basically provides best effort, packet-based data forwarding service. In most cases, the source IP address in the IP packet is not checked in the forwarding process. In the Spoofer Project [1], the authors found that approximately one-quarter of the observed addresses, network address blocks and Autonomous Systems (AS) permit full or partial spoofing. The attacks employing source address spoofing remain a serious concern. Source address spoofing is utilized by some DDOS attacks, such as TCP SYN flooding attacks [2], and smurf attacks. The lack of source address validation provides hackers with anonymity, as it is much harder to trace back the source of the attack facilitated with source address spoofing.

Existing schemes to handle IP source address spoofing include [9]: (1) Tracing back the source of the forged packets with the cooperation of routers [3,4]. (2) Filtering forged packets online [5,6,7,8,9]. (3) Using cryptographic authentication, such as IPSec [10].

Though a lot of methods have been proposed, few of which are really adopted in the Internet. There are two important reasons: one is the lack of incentive to deploy. Most of the proposed methods do not bring direct benefit to the ISPs which deploy them. For example, Ingress filtering can only prevent the hosts in an ISP from sending spoofed traffic to other ISPs, but it cannot prevent receiving spoofed traffic from other ISPs.

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 833–840, 2007.
© Springer-Verlag Berlin Heidelberg 2007

SPM [6] is a newly proposed method to defense against the spoofed attacks, and it is good at providing incentive for both deployment and incremental deployment. The users in the networks who deploy SPM have direct relative benefits from it, so network operators have strong incentive to implement it. Beside that, even if deployed by only a fraction of the Internet networks, SPM still has significant benefit. These two properties make SPM an attractive solution to the problem of source address spoofing.

SPM filters spoofed packets by checking a unique temporal key, which is associated with each ordered pair of source destination AS (Autonomous System) networks. Each packet leaving a source network S is tagged with the key K(S, D), where D is the destination network. Upon arrival at the destination network, the key is verified and removed.

In the original paper for SPM, the basic idea of SPM is described, and the efficiency of SPM is analyzed. However there is no detail specification for the detail mechanisms and prototype system for SPM. There are also some important issues that are critical to realize SPM in the real world, which have not been discussed in the original paper.

The contribution of this paper is making investigation on these un-addressed issues in implementing and deploying SPM. Firstly, the whole SPM system architecture and detail mechanisms are presented. The functions of SPM are split into three components: register server, AS control server, and AS edge router. Secondly, several important issues for SPM system are addressed. Some issues are related to carrying the key required by SPM in the IP packets. Some issues are related to deploying SPM in a transit AS. The mechanism for managing the SPM key is also discussed.

We made a prototype system, and have done some experiments with this prototype system. The result demonstrated that SPM is capable of preventing the spoofed traffic and working seamlessly with the existing network mechanisms.

2 System Overview

In the SPM paper , the authors made some discussion on how to construct the SPM system. Here a concrete design of the SPM system is proposed as is shown in figure 1, which consists of three components.

- Register Server (RS). RS acts as a rendezvous of the SPM system.
- AS Control Server (ACS). ACS in one AS negotiates SPM key and exchange the AS-Prefix with ACS in other AS. ACS is also responsible for configuring the filter in the AS edge routers.
- AS Edge Router (AER). AER is in charge of adding AS key to each outgoing packet, and verifying the AS key of the incoming packets.

All ASes deployed SPM constitutes a SPM Alliance for protecting against the packet with spoofed source address. Member ASes in the SPM Alliance build up a trust relationship with each other via RS. For each pair of ASes in the SPM Alliance, the negotiation of the key and the exchange of AS-Prefix information are handled by ACS in each AS.

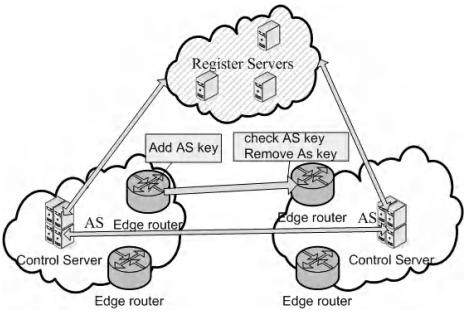


Fig. 1. System Architecture

3 Issues in Designing SPM System

In designing the prototype system for SPM, we find some issues that have not been fully addressed before. They are critical to realize SPM in the real world. Due to the limitation of space, only several important issues are presented in this section.

3.1 How to Carry the Key

It is required by SPM mechanism that a key should be carried in every IP packet. Here how to carry the key in the packet is considered more concretely.

Since there is little space left in the basic IPv6 header, it is necessary to make use of extension header. There are two ways to carry the key with extension header: designing a new type of extension header, or designing a new option in the hop-by-hop extension header. A new extension header is easier for routers to deal with, but it will not be compatible with the current on-shell IPv6 routers. It is expected that SPM will be deployed in an incremental and graceful way, so the focus is put on the second choice.

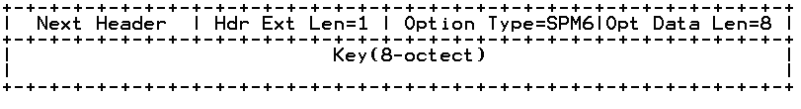


Fig. 2. IPv6 Hop-by-Hop Options

A new type of Hop-by-Hop Options header [11] is designed for SPM. The option type of this new option is 00001100 in binary format. The highest-order two bits 00 specify that the routers will skip over this option and continue processing the header if this option can not be recognized. The third-highest-order bit 0 specifies that the Option Data can be changed en-route to the packet's

final destination. In current implementation, the option type number is set to 01100, which is not used by other option headers.

The new option header of SPM is compatible with end to end authentication mechanisms because the edge routers of the receiving AS will remove the option header from the packets.

3.2 The MTU Problem

The MTU problem arises from the insertion of SPM key into the IP packet. With the additional SPM key, the whole packet length may exceed the path MTU.

The scenario of MTU problem is depicted in figure 3. The path MTU of the link between router1 and router2 is 1400 bytes, and the path MTU of the link between router2 and router3 is 1320 bytes. When a host is in SPM, AS sends out a packet with 1320 bytes, the edge router of the AS1 tags a 12 bytes long SPM option header with key to the packet. The length of the packet becomes 1332 bytes. Then the packet arrives at router2. Router 2 finds that the length of the packet exceeds the link MTU, and it sends back an ICMP Packet-Too-Big packet with the MTU 1320 bytes. When the host receives the ICMP packet, it may be confused since it sent a packet not exceeding 1320 bytes.

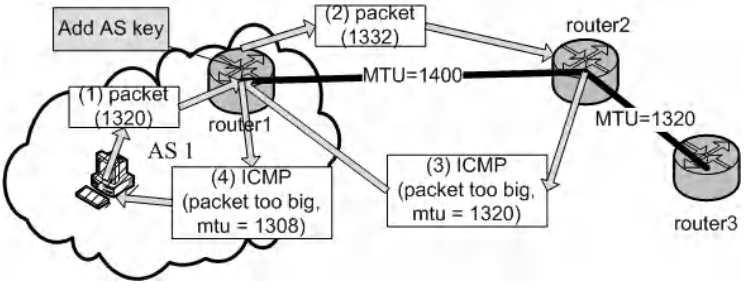


Fig. 3. MTU Problem

As to this problem, a black hole would be set up. The host keeps sending 1320 bytes long packets and it keeps receiving the Packet-Too-Big notification which indicates that the packets are dropped because it is larger than 1320 bytes.

To solve the problem mentioned above, a mechanism is designed to be installed at the AS edge routers to capture and modify the incoming ICMP Packet-Too-Big packets. All the incoming Packet-Too-Big notification packets destined to the local AS will be processed. If the original packets are sent from the local AS to another AS in the SPM Alliance, the MTU value in Packet-Too-Big packets will be modified to a smaller value to reserve the room for the SPM key. For example, in figure 3, the MTU value in ICMP6 packet is modified from 1320 bytes to 1308 bytes.

3.3 Expression for AS-Prefix Ownership

In the SPM, it is necessary to express the ownership of prefix for each member AS of the SPM Alliance. Some discussion should be made on this topic, since it is not as simple as it seems to be.

The expression is very simple for a stub AS. The AS owns all the prefixes assigned to it.However, it is more complex for a transit AS. A fraction of the assigned address of one transit AS may belong to another AS. For example, in figure 4, one multi-homing stub AS3 may have global AS number. And it is allocate a small fraction of address from the provider AS1.

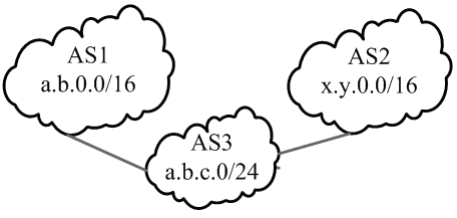


Fig. 4. Multihoming

So the expression for AS-Prefix ownership should be considered more carefully. Two situations are discussed: (1) AS1 and AS3 are both in SPM alliance, (2) AS1 is in the alliance while AS3 is not.

For AS1, it should be explicitly announced that the address space a.b.c.0/24 is not owned by it. If AS3 does not belong to the alliance, ASes in alliance will only receive the announcement for a.b.c.0/24 from AS1. The address space will be marked as non-protected address. If AS3 belongs to alliance, it will announce the ownership for a.b.c.0/24. ASes in alliance receive both announcements for a.b.c.0/24 from AS1 and AS3, so they can conclude that the address space a.b.c.0/24 is owned by AS3 while other parts of a.b.0.0/16 are owned by AS1.

According to the prefix table of AS2 in situation, only AS1 belongs to SPM alliance is shown in table 1. An entry in table will explicitly mark the a.b.c.0/24 as non-protected space. Also, the source of an entry is recorded for update since AS2 may receive new announcement for a.b.c.0/24 in the future if AS3 joins the SPM alliance.

Table 1. AS-Prefix Table

IP address prefix	AS number	Protected address	Announcement Source
x.y.0.0/16	2	Yes	AS2
a.b.0.0/16	1	Yes	AS1
a.b.c.0/24	N/A	No	AS1
x.y.0.0/16	2	Yes	AS2

The longest prefix matching should be used in searching the AS-Prefix table to distinguish the entries such as a.b.c.0/16 and a.b.c.0/24.

3.4 Key Management

Key management is a very important issue for SPM. Our work focused on two points: key negotiation and key switching.

Key Negotiation. The key negotiation happens between the ACS of a pair of ASes in the SPM Alliance. In the original paper, key negotiation is proposed to be sender-driven. The sender AS initiates the key negotiation and the key is generated at the sender AS. Considering the issue of synchronization and the incentive of enabling SPM, receiver-driven key negotiation is suggested. It gives more decision power to receiver AS instead of sender AS.

With receiver-driven scheme, the receiver AS can decide the policies of key management, including:
(1) Whether to enable the SPM anti-spoofing function with some AS in SPM alliance or not. It is not necessary to enable SPM anti-spoofing function between all peer ASes in the SPM Alliance. An AS can make decision based on the situation.
(2) Life cycle of the keys for different ASes. For one AS, it can choose different time interval of changing the key for different peer AS.

Key Switching. Key switching is another important issue. When one AS change the key tagged in the packets to another AS, the packets tagged with old key and the packets tagged with new key may coexist in the network at the same time. To avoid dropping those packets tagged with old key, the AS-In Key table keeps both the old key and new key to check the key of the incoming packets, as is shown in Table 2.

Table 2. AS-in Key Table

AS number	Old key		new key	
	Value	Status	Value	Status
M	FE:12:34:CA:89:76:32:45	Valid	32:54:81:29:FF:00:60:21	Valid
N	89:12:34:89:BC:76:FE:3E	Invalid	22:33:65:78:24:70:AB:C0	Valid

The packets tagged with old key should not be allowed infinitely. After having negotiated the new key, the old key should be set to be invalid after a period of time. The length of this period is a parameter which will control how long the old key will be invalid after the new key has been used. Two minutes is a suggested value since it is enough for the packets with old key disappear in the network.

4 Experiment

SPM prototype, including registration server function, controller function and data plane function (separate device schema), is implemented in Linux 2.6.

The prototype system is tested in an IPv6 environment to valid the design issue illustrated above. The small test environment showed in figure 5 has five ASes. Three shadow ASes are ASes in SPM alliance, while the other two ASes are normal ASes. AS1 is a transit AS, connected with a multi-homing stub AS3. A fraction address of AS1 is allocated to AS3.

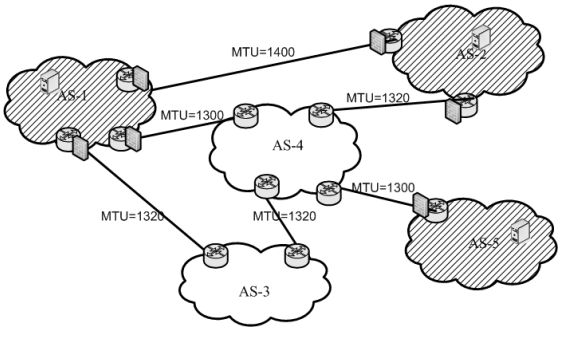


Fig. 5. Experiment Enviroment

The issues mentioned in previous section are tested in the experiment. The experiment results are showed in the following sections. Three types of traffic were generated in the experiment.

- The traffic from AS5 to AS2. It used the spoofed source address belonging to AS4.
- The traffic from AS3 to AS2. It used the spoofed source address belonging to AS1.
- The traffic from AS3 to AS2. It used the spoofed source address belonging to AS4.

The first type of the traffic was filtered inside of the AS5 by Ingress filtering. None of this type of traffic could be observed in the out link of the AS5. The second type of the traffic successful arrived at the board routers of the AS2, but was filtered by SPM. By inspecting the log of SPM, it was confirmed that the traffic was filtered since carrying the wrong key. The third type of the traffic successfully arrived at the victim. Since the AS3 and AS4 did not belong to SPM alliance, the AS2 could not distinguish the traffic from them. But the rate limitation or other measurement can be taken to protect the victim while not infecting the traffic from the ASes in SPM alliance.

The experiment tested the basic function of SPM. SPM can efficiently protect the ASes in SPM alliance from spoofing attack. The spoofed traffic using the source address belonged to ASes in SPM alliance could be identified and filtered.

5 Conclusion

In this paper, we have discussed the basic principle, the advantage and disadvantage of SPM. The original paper have proposed the SPM mechanism and analyzed the benefit of it. But there are a lot of un-addressed issues left in implementing and deploying SPM. These issues are discussed and solved in the paper.

To validate the architecture and the important issues discussed in the paper, a prototype system is implemented. To our best knowledge, it is the first implementation of the SPM. The experiment results demonstrated that the prototype system is capable of preventing the spoofed traffic and works seamlessly with the existing network mechanisms.

In future work, we will focus on the performance issues in data plane and the deployment SPM system in the real world.

References

1. Beverly, R., S. Bauer: The spoofer project: inferring the extent of source address filtering on the Internet. Proceedings of USENIX Steps to Reducing Unwanted Traffic on the Internet Workshop (SRUTI 2005), Cambridge, MA (2005) 53–59
2. C. Schuba, I. Krsul, M. Kuhn, E. Spafford, A. Sundaram, D. Zamboni: Analysis of a denial of service attack on TCP. Proceedings of the 1997 IEEE Symposium on Security and Privacy. IEEE Computer Society, Washington, DC, USA (1997) 208–223
3. A. Yaar, A. Perrig, D. Song: Pi: A Path Identification mechanism to defend against DDoS attacks. Proceedings of the 2003 IEEE Symposium on Security and Privacy. IEEE Computer Society, Washington, DC, USA (2003) 93–107
4. J. Li, M. Sung, J. Xu,L. Li: Large-scale IP traceback in high-speed Internet:Practical techniques and theoretical foundation. Proceedings of the 2004 IEEE Symposium on Security and Privacy.IEEE Computer Society, Washington, DC, USA (2004) 115–129
5. P. Ferguson, D. Senie: Network ingress filtering: Defeating denial of service attacks which employ ip source address spoofing. RFC 2267 (2000)
6. Bremler-Barr, A., Levy, H: Spoofing Prevention Method. IEEE INFOCOM Vol. 1, (2005) 536–547
7. C. Jin, H. Wang, K. G. Shin: Hop-count filtering: An effective defense against spoofed DDoS traffic. Proceeding of the 10th ACM International Conference on Computer and Communications Security (CCS 03).ACM Press, New York, USA (2003) 30–41
8. T. Peng, C. Leckie, R. Kotagiri.: Protection from distributed denial of service attacks using history-based IP filtering. Proceedings of the IEEE International Conference on Communications Vol. 1. Anchorage, Alaska, USA (2003) 482–486
9. J. Li, J. Mirkovic, M. Wang, P. Reiher, and L. Zhang: SAVE: Source Address Validity Enforcement Protocol. IEEE INFOCOM 2002 Vol. 3, (2002) 1557–1566
10. S. Kent, R. Atkinson: Security architecture for the Internet protocol. RFC 2401 (1998)
11. S. Deering, R. Hinden: Internet Protocol, Version 6 (IPv6) Specification. RFC 2460 (1998)

Dongting Lake Floodwater Diversion and Storage Modeling and Control Architecture Based on the Next Generation Network

Lianqing Xue, Zhenchun Hao, Dan Li, and Xiaoqun Liu

State Key Laboratory of Hydrology-Water Resources and Hydraulic Engineering, Hohai University, Nanjing, 210098,P.R. China
lqxue@hhu.edu.cn

Abstract. Dongting Lake is one of the most important floodwater diversion and storage regions in Yangzi River basin. The development of the next generation network and information technology makes it possible to implement the remote monitoring and floodwater control effectively. By taking advantage of the next generation network and lots of 16-byte addresses of IPv6, different kinds of remote sensing devices can be accessed and monitored data can be collected and transferred to the two-dimension hydraulic model for Dongting Lake. And on the basis of real-time network data transmission and parallel algorithm the model can be operated with more unstructured meshes and the hydrological regime and gates are virtually controlled by remote sensors. When analyzing the modeling results, the input data and computing results can be revised in time according to the feedback from the high-speed internet-based system and the project operation can be effectively managed as well.

Keywords: the Next-Generation Network, Network Application, Parallel Computation Algorithm, Hydraulic Model.

1 Introduction

At present, Yangzi River flood control and its hydrological information management are the most important issues in hydrology study field. Because of its size and complicated geographic features, Dongting Lake region has become the most important flood diversion route and water storage area after water diversion from Sikou, on the south bank of Jingjiang River, to Dongting Lake. The efficiency of flood control and regulation command dispatching, however, has been greatly lowered by the previous single motor running and manual data input, and can not satisfy the requirement of hydraulic management using information technology.. Therefore, it is important to utilize the new generation network with overall regulation for real-time monitoring and operation to model the floodwater diversion and storage in Dongting Lake region using high-speed parallel computing method; to set up a platform for water information monitoring and fast transmission; to enhance flood forecast and control; and to use high-tech for flood control regulation, decision-making, evaluation and management after flood disaster. The transition from flood control to flood management will

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 841–848, 2007.
© Springer-Verlag Berlin Heidelberg 2007

play a critical role in flood control of the middle and lower reaches of the Yangzi River. It is also important for the economy development of the Yangzi River basin.

The current problems in floodwater diversion and storage of Dongting Lake region are as follows. (1) Dongting Lake region covers a large area and has a large amount gates. It is difficult to carry out real-time monitoring and information transmission. Therefore, a high-speed information transmission platform should be set to transfer the collected data timely to the model for decision-making; (2) the involved regions are very complicated. The related theory model for floodwater diversion and storage can not be simulated exactly. It needs continuous feedback and revision for the regulation results. The internet-based floodwater diversion and storage management system for Dongting Lake region has the functions of remote data acquisition, monitoring alarm, simulation, real-time operation and revision by the feedback. Next generation network and the standard TCP/IP protocol are essential because of the requirement of video images transmission and varieties of sensors and data.

The system architecture design mainly involves two aspects: (1) real-time data acquisition; and (2) simulation. Firstly, it is simulated according to the real-time data. Secondly, the remote gates are operated based on the simulation results; and at the same time, the real-time data is obtained for subsequent monitoring and modification. The internet-based floodwater diversion and storage control system will be discussed below, including the simulation algorithm and the results.

2 Monitoring and Control System Based on Internet

By using a lot of 16-byte addresses of IPv6, different kinds of sensing devices are accessed and the monitoring data for each diversion area such as environment or earthquake monitoring data can be collected in time, which shows the “more powerful and more convenient ”characteristic of next generation network.

Fig.1. shows the floodwater diversion and storage control system for Dongting Lake region. It mainly consists of two parts: (1) a local area network composed of sensors and controllers. The collected data is transmitted by a gateway through internet to the center. There are different kinds of sensors in the system used for monitoring and control. They are water level sensor, flow velocity sensor, water flow direction sensor, etc. Another kind of data is the video image. The real-time information can be transmitted through high-speed network to the center by the pickup cameras that are laid on some important places, on the gates for instance. In order to operate the gates successfully, some remote controllers are needed. Then each component needs an IP address for overall control. For enough address space, Ipv6 can be set in each component; (2) the center, composed of high-speed computers with parallel computing capacity. The control center makes simulation according to the real time data, controls the controllers based on the simulation results, and thereafter achieves the modification and recycling monitoring.

Each aspect should be considered during the system architecture design, such as communication, sensor, simulation, and control. The network and data control, monitoring, regulation, simulation, feedback and web browser should also be included in this system. The TCP/IP and web technology make it possible for the system be

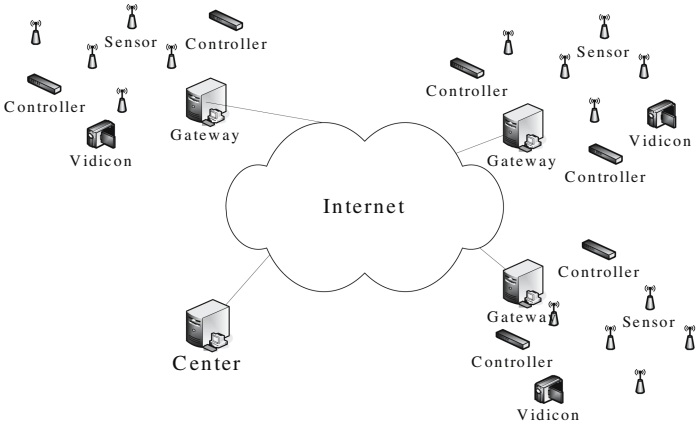


Fig. 1. General Architecture of the internet-based system

effectively controlled at any time and at any place. The internet-based system should include the functions of combined control, simulation, data collection and analysis, management.

There are two essential differences between this system and those traditional internet-based systems: (1) more simulation computing; (2) real-time information transmission.

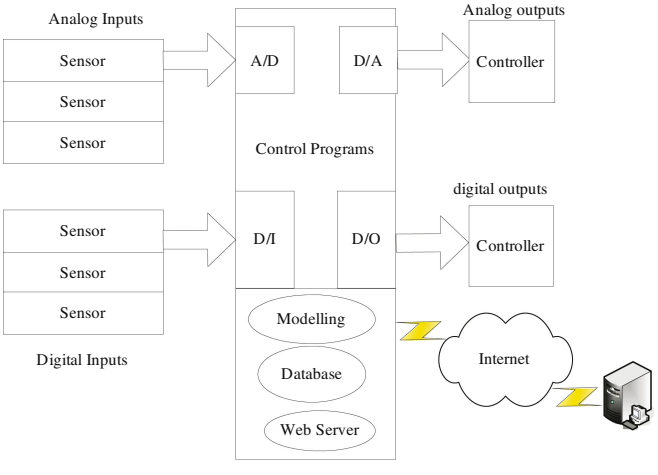


Fig. 2. the Internet-based system for floodwater diversion and storage modeling

Fig.2. shows the architecture of the internet-based remote system for floodwater control and management. Most of them are same as that of traditional ones, such as transmission protocol, safety control, sensor and database. The major difference is that there is a special modeling layer in the internet-based architecture for data simulation, which will be discussed in next section.

3 Two-Dimension Hydraulic Model for Flood Routing

Dongting Lake region covers more than 5000km^2 with 24 diversion areas. Therefore, it is very complicated to simulate this region. The internet-based remote control system introduced above is constructed to solve this problem. This section will mainly discuss the simulation of the model. The model is verified by the historical records. The results show that next generation network and high-speed parallel computers make it possible to do remote real-time control for the Dongting Lake region. The following will discuss the theoretic model and give some examples.

3.1 Hydraulic Model Theory

With the development and application of computer technology, study on adjusting the capacity of Dongting Lake is based on mature hydraulic theory. A two-dimension model is set up to simulate the flood routing of Dongting Lake region and its diversion areas, and to quantify the flood storage and detention capacity of the wetland. As FVM can be directly used in irregular region, it's employed with unstructured mesh for area partition here. In the paper, the two-dimension shallow water equation is transformed from conservative form to discrete form. The two-dimension shallow water equation is set up by selecting water level and flow as controlling variables.

(1) two-dimension hydraulic equation

$$\frac{\partial h}{\partial t}+\frac{\partial (hu)}{\partial x}+\frac{\partial (hv)}{\partial y}=0 \tag{1}$$

$$\frac{\partial (hu)}{\partial t}+\frac{\partial (hu^2+gh^2/2)}{\partial x}+\frac{\partial (huv)}{\partial y}=gh(s_{0x}-s_{fx}) \tag{2}$$

$$\frac{\partial (hv)}{\partial t}+\frac{\partial (huv)}{\partial x}+\frac{\partial (hv^2+gh^2/2)}{\partial y}=gh(S_{0y}-S_{fy}) \tag{3}$$

Where, h is water depth; u and v are vertical mean velocity along x and y direction separately; S_{0x} and S_{fx}, S_{0y} and S_{fy} are lake bed slope and friction slope along x and y direction separately; g is gravity acceleration.

Set $\vec{q}=\{h,hu,hv\}^T$, a state column vector composed of variables that represent flow conservation. set $\vec{f}(q)=\{hu,hu^2+gh^2/2,huv\}^T$ as the flux vector along x direction, and $\vec{g}(q)=\{hu,huv,hv^2+gh^2/2\}^T$ as the flux vector along y direction. When only the bed slope and friction are considered, the source and sink term $\vec{b}(q)$ can be written as follows:

$$\frac{\partial \vec{q}}{\partial t}+\frac{\partial \vec{f}(q)}{\partial x}+\frac{\partial \vec{g}(q)}{\partial y}=\vec{b}(q) \tag{4}$$

(2) Discretization principle of FVM

The explicit form is adopted for space discretization of the continuous equation, and the leapfrog method for time discretization. Crossing grid is used for physical variables calculation. The input or output flow and momentum flux along normal in each discrete unit boundary are worked out. By water and momentum balance equation, the averaged water depth and velocity at the end time are gained at last (under the assumption of averagely distributed hydraulic variables). The details of discretization will not be discussed in this paper.

According to the above theory, the procedures of FVM model are: First, derive water depth h_L, flow velocity u_L, v_L and wave velocity C_L of the calculated unit grid from those in earlier period. Second, derive those of the adjacent grid as well as normal flux $f_{LR}(\cdot)$ by the known water depth and flow velocity. Repeat step 2 for other units next to the already calculated ones. Finally, calculate the water depth h, flow velocity u and v in current period.

3.2 Parallel Computing of Flood Routing

In order to improve the efficiency of flood routing modeling, it is necessary to selectively accept and renew real-time monitoring information, and to take advantage of the parallel computing on internet. Recently, the domain decomposition method has gained more and more attentions from people and it is becoming an important branch of the parallel computing science. If the common grids of adjacent areas can be reduced during region partition, the time of calculation will be much shorter with higher efficiency by using the high-speed network transmission. Since the key problem of parallel procedure design is task allocation and communication, in this paper, Message Passing Interface (MPI) is adopted as the information transmission database, which performed well in communication.

3.3 Grid Generation

On the basis of the above theory, the corresponding procedures are programmed. The mesh is generated using GIS to digitize the 1:10000 groundwater measured data in 1995. Then the flood regulation in Dongting Lake region is simulated, which are the theoretical basis for flood control in Yangzi River basin.

The measured data of the east Dongting Lake, south Dongting Lake and Muping Lake in 1995 are collected. Program interface is developed based on GIS. The quadrilateral mesh for hydraulic calculation is generated by the grid topology program. The principle of grid arrangement is to adopt as few grids to represent the trend of topography as possible, and to make sure that they are reasonably organized and gradually varied. In order to reduce the modeling time, it is essential to adopt different grid scale. Considering the complicated condition in Dongting Lake region, different grid density is used, which not only guarantee the simulation of the flood diversion capacity of Caowei River in the north, but also reflect flood routing characteristics of the south lake as well as the general distribution of topography. The smallest grid is 60m*62m, and there are grids up to 100,664 in the south lake area. Fig.3. shows the grid arrangement.

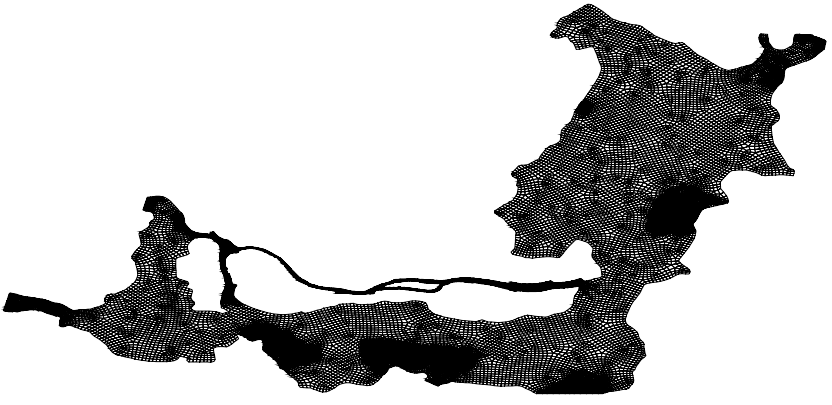


Fig. 3. Calculation Grids of Dongting Lake

3.4 Simulation Analysis

In this paper, the flood routing from 0:00 am, July 21th to 23:00 pm, August 2th in typical year 1998, is simulated and compared with the parallel computing results. The results are in good agreement with measured flood routing trend. The model gives a reliable simulation with the average water level discrepancy of 0.11m. Therefore, it can be used for quantificational prediction of flood diversion process.

While there are 24 diversion areas in Dongting Lake region, only one mode of gate operation is listed in this paper. The results show that if the gates were opened, the total flood storage volume of five diversion areas would be 64.6*10^{8}m^3. The five diversion areas are Gongshuangcha, Qiannianghu, Datonghu Jianshe and Jianxin. For the flood in 1998, lake area met the floodwater diversion and storage request from July 21th till August 2th, and the volume of storage water was 54*10^{8}m^3. Water level of outside lake dropped 0.3m. However, it became uniform with natural water level

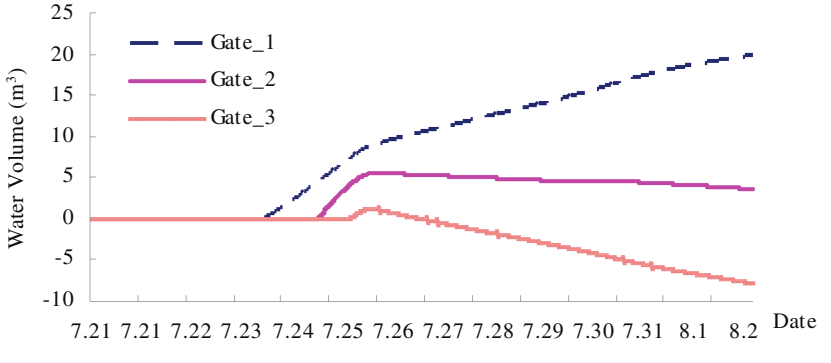


Fig. 4. Water inflow and outflow hydrograph for gate1, gate2 and gate3 (10^{8}m^3)

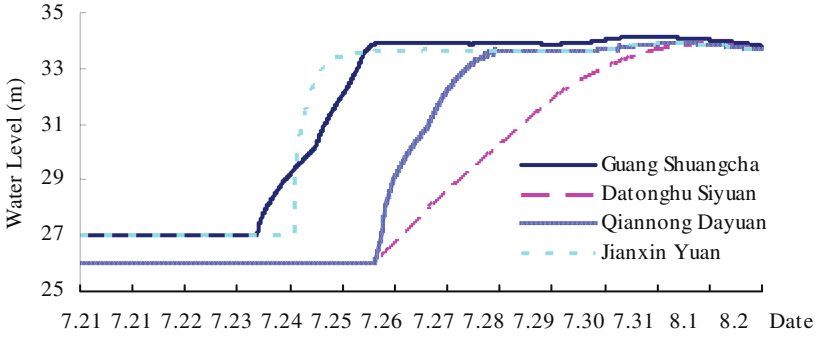


Fig. 5. Water level hydrograph in 4 different places

since August 2nd. The diversion areas lost their floodwater diversion and storage function. Gate No. 1 in Gongshuangcha lies in upstream, where the relative water level is higher. So it was opened earlier than Gate No. 2 and Gate No. 3. It can be observed from Fig.4. Gate No. 1 was opened all the time, while Gate No. 2 was opened after Gate No. 1. The volume of passed flow in Gate No. 1 and Gate No. 2 are positive. Gate No. 3 was closed in the beginning and opened later. The volume of passed flow changed from positive to negative. The above Results show that it is inadequate to use diversion areas in Dongting Lake region. Therefore, it is very important to adopt high-speed parallel computing method. Fig.5. shows the water level hydrograph.

4 Conclusions

By taking advantage of new generation network, floodwater diversion and storage control system for Dongting Lake region is constructed and discussed in this paper. Satisfying application results shows that the system is proficient in real-time data acquisition, high-speed information transmission, and remote control, which can effectively improve the efficiency of flood control and regulation command dispatching. Meanwhile, comparing the internet-based system for Dongting Lake with the traditional systems, we find that the main difference is the use of high-speed parallel computing for simulation in the internet-based system. Furthermore, the two-dimension hydraulic model for Dongting Lake employs Finite Volume Method (FVM) and unstructured mesh for area partition. The flood routing in a typical year is modeled and the regulation effects of 24 diversion areas are predicted, during which the hydrological regime and gates are virtually controlled by remote sensors and controllers. It is also proved that the application of new generation network can be fully broadened in future because it's more powerful and more convenient. Moreover, it is of great practical value in future flood control.

Acknowledgments. This paper has been funded by the National Natural Science Foundation of China under Grant No. 50609006, the Key Project of Chinese Ministry of Education under Grant No. 105084, the Natural Science Foundation of Jiangsu Province under Grant No. BK2006092, and the Excellent Youth Teacher of Southeast University Program under Grant No. 4009001018.

References

1. Paul I-Hai Lin, Harold L. Broberg, Internet-Based Monitoring and Controls for HVAC Applications, IEEE Industry Applications Magazine, (Jan-Feb 2002), 49-54
2. Paul I-Hai Lin, Web Programming and Applications. Fort Wayne: Indiana University-Purdue University, (2001).
3. Chen Jin, Huang Wei, Inquiry into system risk analytic method of flood control engineering. Journal of Yangtze River Scientific Research Institute, (2001) 18(5): 37-40.
4. Piotr Małoszewski, Przemysław Wachniew and Piotr Czupryński, Study of hydraulic parameters in heterogeneous gravel beds: Constructed wetland in Nowa Słupia (Poland). Journal of Hydrology, Volume 331, Issues 3-4, 15 December (2006) 630-642
5. Zhang Yanzhao, Dong Jie, Han Min. Research of flood routing simulation system based on GIS. Chinese Journal of Scientific Instrument, (2006)27(6): 936-937.
6. Yu Yunli, Wang Deguan, Wang Zhigang, Lai Xijun, Numerical Simulation of Thermal Discharge Based on FVM Method, Journal of Ocean University of China. (2006) 5(1): 7-11
7. Minhyung Kim, Sangkyun Kim and Arcy J. Kong, High performance AAA architecture for massive IPv4 networks. Future Generation Computer Systems, Volume 23, Issue 2, (February 2007) 275-279.
8. A.Bejan and S.Lorente, Constructal tree-shaped flow architecture. Applied Thermal Engineering, Volume 27, Issue 4, (March 2007) 755-761

Query Processing to Efficient Search in Ubiquitous Computing

Byung-Ryong Kim[1] and Ki-Chang Kim[2]

[1] Department of Computer and Science Engineering, Inha University, Incheon, Korea
[2] School of Information and Communication Engineering, Inha University, Incheon, Korea
{doolyn,kchang}@inha.ac.kr

Abstract. Both ubiquitous computing and mobile ad-hoc networks (MANET) have recently attracted a lot of attention in the research community as well as the industry. Both the domains share certain similarities, primarily the fact that both are instances of self-organizing decentralized systems. Distributed hash table (DHT) provides a very effective and reliable search scheme in Both Networks. However, when the search involves query consist of a set of common words, it suffers heavy network traffic due to the passing around of a large inverted list among nodes. In this paper, we suggest a technique based on the concept of distance between keywords that can remove irrelevant indices from the list. It utilizes the concept of distance between keywords in the query and removes those entries in the inverted list that are going to be dropped sooner or later. We prove this prediction is accurate and effective such that it reduces the size of the inverted list.

Keywords: ubiquitous computing, query processing, distributed hash table.

1 Introduction

Distributed Hash Table(DHT) is used to map a keyword to the peer responsible for holding the corresponding reverse index. Partition-by-Keyword uses DHT to store the indices of files, and uses these tables to find the location of a target document. It hashes the keyword of a document and stores the index of the document (that is its actual location such as URL address) in a node which controls the hash table corresponding to the keyword's hash value. Since a document can contain several keywords, the index of this document can be stored in a number of nodes. To find the location of a document that contains a keyword x, the client peer simply hashes it and send a query to a node which controls the corresponding hash table. However it still poses some performance problems, especially when the query contains many keywords. For query with many keywords, the query should be sent to all nodes that control these keywords. Since the inverted lists (lists mapping the keyword to a list of documents that contain it) from the relevant nodes need to be combined through intersection, the query is passed around with the inverted list being updated at each node. If the first keyword in the query was a common keyword, the starting inverted list

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 849–852, 2007.
© Springer-Verlag Berlin Heidelberg 2007

would be huge, and transmitting this huge list to the next node would cause a heavy traffic. Usually the keywords in a particular query consist of a set of common words and another set of specific words that occur only in a limited number of documents. Most of the indices of documents in the starting inverted list, therefore, will most likely be dropped through intersection sooner or later.

It would be beneficial if we can cut down these irrelevant indices from the inverted list before transmitting it to the next node. But how can we know which indices will eventually be dropped? In this paper, we suggest a technique based on the concept of distance between keywords that can remove irrelevant indices from the list. Our technique is explained in detail later in the paper, and the preliminary result shows the technique is very promising. The rest of the paper is organized as follows: Section 2 explains the basic operations of DHT-based P2P system and surveys searching techniques for query with many keywords; Section 3 explains the proposed technique in detail and Section 4 gives a conclusion.

2 Related Works

Numerous researches have been performed to reduce the network traffic in DHT overlay network for keyword search. Multi-level partitioning (MLP) transmits the query only to some limited number of peers to decrease the traffic. In order to select appropriate peers, the technique divides the node space into k logical groups[1]. [2,3] uses Bloom Filter and previous search results to compress the intermediate URI list. Bhattacharjee proposed another technique to improve efficiency: result caching[4]. Keyword Fusion[5] uses Fusion Dictionary, which contains a set of common keywords, to achieve an efficient searching. SCALLOP[6] utilizes a balanced lookup tree to avoid the hot spot problem. In order to reduce query overhead, KSS(Keyword Set Search) by Gnawali partitions the index by a set of keywords[7]. Hybrid-indexing[8] extracts a set of important keywords from the document and utilizes it together with the inverted URI list. mSearch[9] also employs Hybrid-indexing, but it defines a multicast tree for each keyword in order to multicast the query only to some relevant nodes. pSearch[10] reduces the number of visiting nodes for a given query and still achieves high accuracy.

The above techniques have been successful in reducing the network traffic in some aspect, but either the reduction rate is not enough, or they require another type of system resource such as memory space. MLP introduces additional cost for communication between nodes to maintain the grouping information. Using Bloom Filter can cause the problem of false positive in which the hit rate varies considerably depending on the number of hash functions and the size of the bit vector. The dictionary in Keyword Fusion takes time to build and needs to be updated frequently, which costs additional traffic. SCALLOP requires additional storage for lookup table. KSS also causes increasing storage overhead as the number of keyword combination increases. The multicast tree used in mSearch demands additional space overhead, and Hybrid-indexing requires additional space to store major keywords for each document.

3 The Concept of Distance

In this study distance is applied to find out unmatched entry in inverted lists administered by nodes responsible for keyword. Distance means how far the keywords administered in hash range in network are located each other. Node participating in network complies with the following process when distributing document. All the keywords in the document are drawn out and the keywords are hashed using hash function such as SHA-1. Then after arranging the hashed values in descending order, the each difference between two hash values is computed in order.

These computed values are distance. When the number of keywords included in the document is n, the distance of keyword k_i, $d(k_i) = hash(k_{i+1}) - hash(k_i)$ and when i is n (keyword with the largest hash value among keywords in the document), $d(k_n) = 0$. Inverted list of keywords included in the document is shown as <document's url, k_i, $d(k_i)$> added the value of $d(k_i)$. This $d(k_i)$ value makes it possible to tell which keyword with larger hash value than k_i is included in the document including k_i , or if there is any definite chance of inclusion or not. For example to know whether w is included in document x including k_i, if $d(k_i)$ out of every entry of k_i 's inverted list is larger than (hash(w)- hash(k_i)) it clearly means that w is not included in x, if smaller it means that w may or may not be included, and if it equals to the value, it means w is included for sure.

To apply this notion of distance, the search process shall be different from general one. Let's assume the process of search (x AND y) for document including keyword x and y. First of all keyword x, and y included in query are hashed. If hash(x) > hash(y), node y administering inverted list is firstly visited and if not, distance technique can be applied only after visiting node x administering x's inverted list first. When hash(x) > hash(y), the entries, other than with $d(y) >$ (hash(x)- hash(y)) among y's inverted list, are forwarded to node responsible for keyword x because it is sure that the excluded entries contain the document URL never including x according to the definition of distance.

Every inverted list of y has been sent to the node responsible for x in the existing techniques but using distance effectively diminishes traffic since only entries surely including x or with high possibility of the inclusion are sent. Tests were conducted to evaluate the performance of the proposed technique, the number of nodes is 1,000 and the number of keywords included in query is randomly set at maximum 5. Test results showed that the traffic was reduced by approx. 40% with 2 keywords in query and the overall traffic was reduced by about 30% on average. The distance forwards the inverted list with any chance of inclusion but the amount of inverted list forwarded this way is not small yet. Hence further study on the distance extension will be needed to solve this problem.

4 Conclusion

DHT is very effective for fast keyword-based-search of a file in ubiquitous computing. However query with many keywords has been known to cause a heavy traffic due to a large inverted list that has to be passed around. We have proposed a technique based on the concept of distance which can cut down the size of the inverted list

considerably and therefore reduce network traffic. According to the test results when the number of keywords included in query was 2 with the application of distance, traffic was reduced by 40%. With 2 to 5 keywords the traffic was reduced by about 30% on average. Consequently with the decrease in the size of inverted list to be sent at intersection at each node, it can say a great contribution to ubiquitous application service.

Acknowledgments. This work was supported by INHA UNIVERSITY Research Grant.

References

1. Shuming Shi, Guangwen Yang, Dingxing Wang, Jin Yu, Shaogang Qu, Ming Chen. Making Peer-to-Peer Keyword Searching Feasible Using Multi-Level Partitioning. In IPTPS, (2004).
2. Patrick Reynolds Amin Vahdat. Efficient Peer-to-Peer Keyword Searching. In Middleware, (2003).
3. B. Bloom. Space/time trade-offs in hash coding with allowable errors. Communications of the ACM, 13(7) (1970) 422-426,.
4. B. Bhattacharjee and S. Chawathe and V. Gopalakrishnan and P. Keleher and B. Silaghi,Efficient peer-to-peer searches using result-caching, in The 2nd International Workshop on Peer-to-Peer Systems(IPTPS'03), (2003).
5. Lintao Liu, Kyung Dong Ryu, Kang-Won Lee. Keyword fusion to support efficient keyword-based search in peer-to-peer file sharing, CCGRID 2004, (2004), 269-276..
6. Jerry C. Y. Chou, Tai-Yi Huang, Kuang-Li Huang, Tsung-Yen Chen. SCALLOP: A Scalable and Load-Balanced Peer-to-Peer Lookup Protocol. IEEE Trans. Parallel Distrib. Syst, 17(5) (2006) 419-433.
7. O. Gnawali. A Keyword Set Search System for Peer-to-Peer Networks, Master's thesis, Massachusetts Institute of Technology, (2002).
8. C. Tang and S. Dwarkadas. Hybrid Global-Local Indexing for Efficient Peer-to-Peer Information Retrieval, In Proceedings of the Symposium on Networked Systems Design and Implementation (NSDI), (June 2004).
9. Ajay Gulati and Supranamaya Ranjan. Efficient Keyword Search using Multicast Trees in Structured p2p Networks submitted to Middleware (2005).
10. C. Tang, Z. Xu, and S. Dwarkadas. Peer-to-Peer Information Retrieval Using Self-organizing Semantic Overlay Networks. In Proceedings of SIGCOMM (2003).

Service and Management for Multicast Based Audio/Video Collaboration System on CERNET

Xuan Zhang[1,2], Xing Li[1,2], and Qingguo Zhao[1]

[1] Network Research Center
[2.]Department of Electronic Engineering,
Tsinghua University ,Beijing China, 100084
{zhangx,xing}@cernet.edu.cn ,zhao-qg@mails.tsinghua.edu.cn

Abstract. Multicast based audio/video collaboration is one of representative applications in next generation internet. IP multicast has the advantage of saving bandwidth for group communication on large scale collaboration. The lack of ubiquitous multicast supporting and lack of effective service for multicast based A/V collaboration limit its application. This paper introduces the service and management of multicast based audio/video collaboration on China education and research network (CERNET). The multicast service, user and session management, performance monitoring and flow control are presented. The collaboration system has been applied on CERNET successfully.

Keywords: service, audio/video collaboration, multicast.

1 Introduction

Multi-party audio/video (A/V) collaboration systems are important applications in Next Generation Internet [1].In multi-party A/V collaboration, users could receive all group users' A/V stream with strong interactive characters. Among the solutions multicast based system may be ideal way for group collaboration. IP multicast could save bandwidth for many-to-many media communication, which makes it scalable. Some multicast based A/V systems as AccessGrid[2] have been employed recently. But for some reality reasons, the native multicast is not available ubiquitously, which confines the multicast applications extensively. On the other hand, the current A/V collaboration system lacks effective service and management, for example, lack of guarantee on quality of service. In this paper, we introduce the service and management for multicast based A/V collaboration on CERNET. The IP multicast service on CERNET, user and session management, performance monitor and flow control are presented. We have implemented and applied the multicast A/V collaboration successfully.

2 The Multicast Service on CERNET

The native multicast is deployed at network level routers with multicast protocols. Now we have configured native multicast on CERNET backbone covering all the provincial capitals. The multicast routing protocol deployed on CERNET backbone is PIM-SM (Protocol Independent Multicast-Sparse Mode) [3]. The PIM-DM [4] is

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 853–856, 2007.
© Springer-Verlag Berlin Heidelberg 2007

deployed in regional networks, and MSDP [5] are used for routing between different AS domains. In practice on deploying PIM-SM, we adopted 10 anycast RP(Rendezvous Point)s with same IP address via MSDP[5].

To users not in multicast domain we supply reflectors for them to join multicast group. Reflectors are application layer gateway which could convert unicast streaming data into multicast data and convert the multicast data into unicast data conversely.

Figure 1 shows the communication mechanism of reflector. The unicast users send their A/V streaming to reflector, reflector receive the unicast data and retransmit them to multicast group. On the other hand, the multicast A/V data from multicast users are received by reflector and resent to unicast users. By this way, users could receive all group users' A/V streaming whether the users are in unicast or multicast domain.

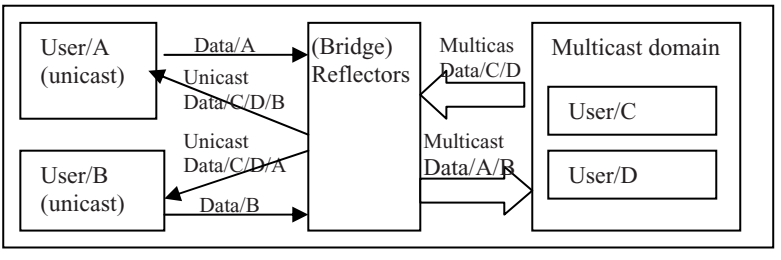


Fig . 1. The mechanism of reflector for unicast access

3 User Management and Session Service

We establish one website http://mvc.edu.cn[6] at CERNET national network center as collaboration service center for multicast A/V collaboration (CSC-MVC).

3.1 Users Manage and Venue Generation

Users in A/V collaboration could be classified into two types: chairman and participates. Chairman is the sponsor of collaboration venue. Commonly chairman applies for new venue and manages the venue. When chairman applies for create new venue, the Administrator of CSC-MVC would generate one new venue and add it to venue-list. The administrator assigned the session information for the venue.

Some authorities on venue manage may be endowed to the chairman by CSC-MVC administrator .The chairman manages the participant users of this venue. If one user want to attend one venue, he must apply to the chairman or be invited by the chairman. In our service, the sponsor adds the invited users' name to the participants list, only the users in the invited participants list are permitted to enter the sponsor's venue room, otherwise he would be denied.

3.2 Session Service

Session information is created by CSC-MVC administrator. The session information mainly consists of IP/port pairs for audio/video. To users in native multicast domain,

the multicast IP addresses are supplied, to users in unicast domain, the unicast IP addresses of reflectors are supplied.

The user first accesses the CSC-MVC website and login, authorized user could access venue room list (as figure 4 shows) and enter the prefer venue room if the user has been invited by chairman of this venue room. The session information stored in the venue is sent to end users when users enter venue. By parsing session information, the end user would get the multicast (and unicast) IP addresses/ports to launch their audio/video tools. Secret key and other session information are also used for launching audio/video. When at least two users enter the venue and communicate each other via audio/video tools, the multicast session is established. The session is terminated when all the users leave from the venues room. To temporary venue room, the CSC-MVC would delete the venue and take back the multicast IP /port pairs.

4 Performance Monitor and Flow Control

To guarantee the quality of multicast based A/V collaboration, we design and implement the performance monitor and flow control scheme. We build a server for performance monitor and flow control charged by chairman. As Figure 2 shows, during collaboration the endpoints send their performance messages to server. The messages include the user's sending traffic, end-to-end delay and packet loss rate.

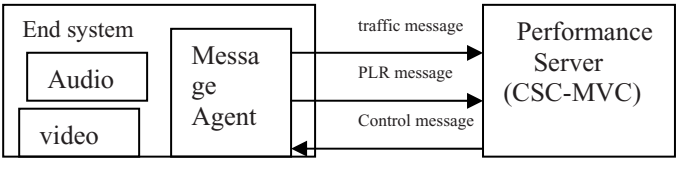


Fig .2. The performance monitor and flow control service

The end-to-end delay and PLR are calculated through RTP/RTCP protocols in audio/video tools. For example, PLR are calculated by received RTP packets. In a group with N users, each user can get N-1 PLR from other senders, figure 3 shows one user's end-to-end PLR report message. When all N users send their PLR message to the performance server, the server could get N*(N-1) end-to-end PLRs. Chairman monitors the application performance from server. When congestion occurs, chairman would send rate control message to endpoints to adjust their sending bit rate. By this simple manual control, serious congestion could be avoided.

Message Type	Seq_no	Timestamp
Local source	Local Cname(IP addr)	
Source1	Cname1 (IP addr)	PLR value
Source2	Cname2 (IP addr)	PLR value
...Source(i)	Cname(i)(IP addr)	PLR value

Fig .3. End user's end-to-end PLR message

5 Service and Application on CERNET

The multicast based audio/video collaboration service has been provided on CERNET since 2002.In 2003 SARS period, the system advanced the A/V collaborations among domestic universities. Now the service has expanded to all provinces of China and become the A/V communication platform for CERNET members. Figure 4 shows one venue room list on collaboration service center for multicast A/V collaboration.

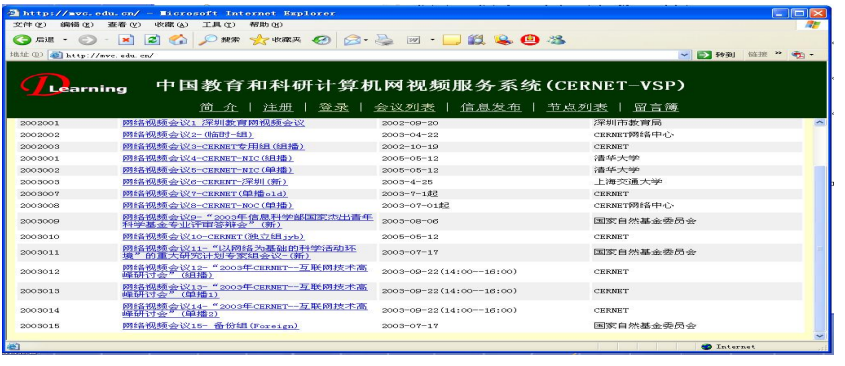


Fig .4. One venue room list on CERNET CSC-MVC from http://mvc.edu.cn

6 Conclusions

This paper introduces the service of multicast based audio/video collaboration systems on CERNET. The multicast service, users manage and session service, performance monitor and control are presented. The multicast based audio/video collaboration has been applied on CERNET successfully in nationwide.

References

[1] Internet2 consortium, http://dv.internet2.edu/ .

[2] Access Grid Project, http://www.accessgrid.org

[3] L. Wei, Protocol Independent Multicast-Sparse Mode (PIM-SM): Protocol Specification, Internet experimental RFC 2117, June (1997).

[4] Adams, J. Nicholas, W. Siadak, Protocol Independent Multicast - Dense Mode (PIM-DM): Protocol Specification (Revised), Internet experimental RFC 3973, January (2005).

[5] D. Kim, D. Meyer, H. Kilmer, D. Farinacci, Anycast Rendevous Point (RP) mechanism using Protocol Independent Multicast (PIM) and Multicast Source Discovery Protocol (MSDP), Internet informational RFC 3446, January (2003).

[6] Multicast based video service on CERNET, http://mvc.edu.cn .

A Double-Sampling and Hold Based Approach for Accurate and Efficient Network Flow Monitoring

Guang Cheng[1], Yongning Tang[2], and Wei Ding[1]

[1] College of Computer Science & Engineering,
Southeast University, Nanjing, P.R. China, 210096
[2] School of Computer Science, Telecommunications and Information Systems,
DePaul University Chicago IL USA 60604
gcheng@njnet.edu.cn

Abstract. One crucial challenge in network flow monitoring is how to accurately and efficiently monitor the large volume of network flows. Several approaches proposed to address this challenge either lack flexibility adapting to greatly varying network traffic (e.g. sNetFlow), or require intensive computing resources (e.g. ANF). In this paper, we propose a novel double-sampling and hold approach for network flow monitoring to tackle this challenge. We take a coarse-grained packet sampling to initially reduce the total number of monitored packets; then, an enhanced fine-grained sample and hold algorithm (ESHA) is adopted to selectively add packets into flow cache. By optimally adjusting the ESHA sampling rate and taking Early Removal flow cache management scheme, the flow information can be maximized with given limited system resources. Extensive simulation and experiment studies show that our approach can significantly improve both the accuracy and efficiency in network flow monitoring than other methods.

Keywords: Sampling Methods, Internet Traffic, Measurement.

1 Introduction

Network traffic monitoring and analysis is crucial for many network applications, such as network planning, network management, and network security applications. Network packets passing through the monitoring system can be classified into flows based on their header 7-tuple information [1], which can be further analyzed to present more significant implications. NetFlow [2], implemented in Cisco routers, can generate and output flow records, and keep a flow cache into memory as flow records to describe the passing traffic. Current flow monitoring approaches, which require recording flow records into memory to keep the flow status, usually run into problems if the number of flows too huge to hold in the memory. Several approaches have been proposed to address this challenge. However they are either lacking of flexibility adapting to greatly varying network traffic (e.g. sNetFlow), or requiring intensive computing resources (e.g. ANF).

In this paper, we propose a novel approach to tackle this issue by using a double-sampling and hold scheme to make the monitoring system self-adjustable to the

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 857–864, 2007.
© Springer-Verlag Berlin Heidelberg 2007

varying monitored traffic. In this approach, firstly, we take a coarse-grained packet sampling to initially reduce the total number of monitored packets to satisfying CPU processing capability in the monitoring system; then, an enhanced fine-grained sampling and hold algorithm (ESHA) is adopted to selectively add packets into flow cache. By optimally adjusting the ESHA sampling rate and taking Early Removal flow cache management scheme, the flow information can be maximized with given limited system resources. The paper is organized as follows: in Section 2, we discussed the related work. In Section 3, we elaborate our approach with detailed discussion on the corresponding algorithms. Comprehensive experiments are conducted and the results are discussed in Section 4. Section 5 concludes this work.

2 Related Work

With the increasing demands from various areas such as network security, network flow monitoring has gained more and more attentions by research community. The IETF Packet Sampling working group (PSAMP) [6] is chartered to define a standard set of capabilities for network elements to sample subsets of packets statistically. Furthermore, sampling techniques have been proposed [3,4,5] and employed in network products [2]. Kumar [10] proposed a novel SCBF that performs per-flow counting without maintaining per-flow state in and an algorithm for estimation of flow size distribution [11]. This is followed by [13] in which the flow distribution is inferred from the sampled statistics. Duffield [7, 12] studied the statistical properties of packet-level sampling using real-world Internet traffic traces, and developed a simple model to predict both the export rate of flow packet-sampled flow statistics and the number of active flows. The measured numbers of flows and the distribution of their lengths can been used to evaluate gains in deployment of web proxies [8].

However Sampled NetFlow suffers from two problems: (1) the number of flow cached in system memory could be significantly increasing if the monitored network hit by certain burst traffic such as DDoS; (2) pre-selected sampling rate cannot be adapted to the varying network traffic. Estan [1, 9] has proposed a family of bitmap algorithms for counting active flows and an adaptive NetFlow (ANF) was proposed. The adaptive NetFlow algorithm keeps resource within fixed limits, and uses renormalization to reduce the number of NetFlow records after a decrease in sampling rate. This algorithm divides traffic stream into fixed time interval bins. However, the algorithm has the following limitations: (1) it consumes a lot of CPU resource during renormalization because it needs repeatedly to analyze previously collected flow records that have already been processed and stored in system memory; (2) it inevitably loses monitoring accuracy due to static sampling rate; (3) it has to frequently adjust its sampling rate to adapt to varying traffic flow statistics.

Obviously, there is a trade-off between monitoring accuracy and limited system resources (e.g. memory size, CPU speed). How to select an optimal sampling rate to achieve satisfactory accuracy with given system resource is a significant challenge. In this paper, we will try to tackle this issue.

3 Double-Sampling and Hold Based Flow Monitoring

3.1 System Overview

The double-sampling and hold based flow monitoring system consists of two components: Flow Information Updating (FIU) and Flow Cache Management (FCM). FIU updates flow information by considering new received packets, and FCM adopts an Early Removal policy to remove the number of least significant flows (e.g. small flows) from the flow cache to accommodate memory space for more informative flows. FIU adopts a double-sampling and hold mechanism, which consists of a coarse-grained sampling module, an enhanced fine-grained sampling and hold module. The coarse-grained sampling module is to protect the monitoring system from overwhelmed by effectively controlling the number of measured packets to decrease the load of CPU, processing the large volume of incoming traffic. Currently it simply samples one of every N packets. N is a system parameter configured based on CPU capacity and average traffic statistics. Fine-grained Sampling and hold module [4] records all packets information which belong to the existed flows, and then updates the corresponding flow entries. Otherwise, the packets are sampled with a probability and new entries will be created accordingly. In this manner, the heavy-tailed flows will be given higher priority, which also contain more information. Thus, this approach can maximize flow information and estimation accuracy with given flow cache size. FCM takes Early Removal policy remove small flows to control the total number of flows in the flow cache. Figure 1 show the architecture of the system.

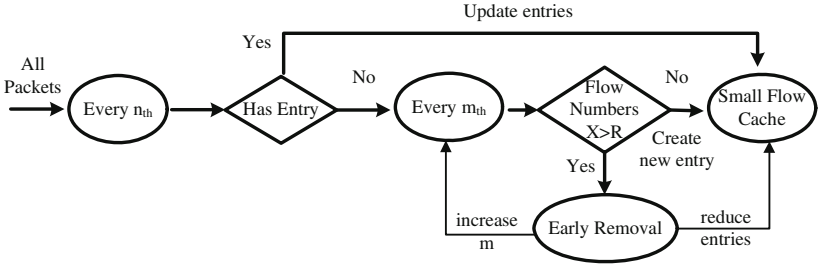


Fig. 1. This algorithm architecture. In this figure, all packets are sample by one out of n^{th}, and all sampled packets are processed by hold function. If there is a flow entry in the flow memory, then the flow entry is updated, otherwise the packet is sampled again using one out of m^{th} to decide whether a new flow is created. If the number of flows in the flow memory is larger than a threshold R, then a removal algorithm will be started up to remove some flow entries.

3.2 Algorithm Analysis

In the following, we first define some notations: N1 and N2 are the sampling rates used in coarse-grained and fine-grained samplings respectively; S is a collection of flows in the flow cache; T1 and T2 are the thresholds to trigger and stop Early Removal processing for the flow cache; $F = |S|$; m_1 and m_2 are the number of packets sampled by coarse-grained and fine-grained samplings respectively; m is the total number of incoming packets.

Network flows (called original flows) based on 7-tuple header information, which are, can be further aggregated to form new flows (called generated flows) based on a subset of 7-tuple header information. Assuming an aggregated flow x contains t original flows, and the number of flows in the flow cache S is T, the number of packets that are selected by coarse-grained sampling but dropped by fine-grained sampling is m_1-m_2. Apparently, based on our sampling and removing mechanism, these dropped m_1-m_2 packets must belong to small flows. Because of the applied hold algorithm, the packet number in flow cache is equal to the packet number of the sampled flows. Since all dropped flows and packets belong to small flows, which have the same random distribution, we have that the ratio of the flow number in the aggregated flow x to the number of flows in the flow cache is equal to the ratio between dropped packets of the aggregated flow x and the original packets, that is t/T=x/(m_1-m_2), x is the unsampled packets. According to the ratio, we can estimate x=(m_1-m_2)T/t. Thus we can add up the two parts to estimate the sample result $\hat{x} = m_x + (m_1 - m_2) \cdot T / t$. If we consider the first sample information, the equation (1) is the estimated result of the aggregation flow x.

$$\hat{x} = (m_x + (m_1 - m_2) \cdot T / t) / p_1 \tag{1}$$

Let the ratio between the number of packets of aggregation flows and the total number of packets be f, so the packet number of aggregation flows is f*m. The distribution of sampled aggregation has a binomial distribution with mean p*f*m, and variance p(1-p)f*m. Since the estimate for the number of packets in the aggregation stream can be computed by multiplying the number of sampled packets by 1/p, and the estimated variance is $(1/p^2)p(1-p)f \cdot m = (1-p)f \cdot m / p$. The standard deviation of this estimate is $SD(x) = \sqrt{(1-p)f \cdot m / p}$, p=m_x/(m_2N_1), and its relative error RE(x) is equation (2).

$$RE(x) = \frac{\sqrt{(1-p)f \cdot m / p}}{f \cdot m} = \sqrt{\frac{1-p}{f \cdot m \cdot p}} = \sqrt{\frac{1 - m_x / m_2 N_1}{f \cdot m \cdot m_x / m_2 N_1}} = \sqrt{\frac{m_2 N_1 - m_x}{f \cdot m \cdot m_x}} \tag{2}$$

3.3 Early Removal

The goal of Early Removal is to adapt fine-grained sampling rate and further control the total number of entries in the flow cache. The adaptive algorithm reduces the fine-grained sampling rate whenever the number of flow entries exceeds a predefined threshold. In this paper, we remove all flow entries whose size is less than a threshold d. If the number of the removed entries cannot satisfy the system requirement on the flow cache, then the threshold d increased by Δ and removing process will be conducted recursively until the requirement is satisfied.

The system may take long time to remove a predefined number of flows if we have to go through the entire flow cache multiple times. Here we introduce a quick search algorithm, which only searches the flow cache one time. Firstly, we set a k size dimension K. When the flow cache is updated, the dimension K is also updated. The index in the dimension K is the length of flow in the flow cache. If the length of a

flow s is x, and x < k, then the packet which belongs to s will be sampled and s will be updated accordingly. So we can set K[x]=K[x]-1, K[x+1]=K[x+1]+1. If x=k, then K[x]=K[x]-1. If a flow size is larger than k, then the flow won't be recorded in the dimension K. If a new flow is created in the flow cache, then K[1]=K[1]+1.

So if we want to set H as the removal threshold, according to the dimension K, the length of removal flow can be computed in equation (3). In the equation (3), if we set the removal threshold d=j, then the removal threshold H can be obtained by only going through the flow cache one time. If the k is too small, it may be $H > \sum_{i=1}^{k} K[i]$, then we can set d = k+1, and loop the flow cache to remove the threshold H flows. In this experiment, we set k be 100.

$$\sum_{i=1}^{j-1} K[i] < H \le \sum_{i=1}^{j} K[i] \text{ , j <= k} \tag{3}$$

4 Experiment Analysis

We use one team packet header traces gathered at NLANR [5]. This team traces use OC192MON hardware to collect data on August 19, 2004, from 13:40pm to 14:40pm. These traces information is displayed in table 1.

Table 1. Traces used in this experiment

No.	Time	File Size	#packets	#flows
1	2004/08/19 13:40	173MB	8434885	144813
2	2004/08/19 13:50	154MB	6922629	146665
3	2004/08/19 14:00	172MB	8251311	184213
4	2004/08/19 14:10	157MB	7111907	154750
5	2004/08/19 14:20	161MB	7388868	140297
6	2004/08/19 14:30	176MB	8560571	143341
7	2004/08/19 14:40	162MB	7527445	149196

In this experiment, we set the time bin 10 minutes, and the flow cache size is 2^{13}=8192, 2^{14} = 16384, 2^{15} = 32768, 2^{16} = 65536, 2^{17} = 131072, 2^{18} = 262144, respectively. Firstly, we give a definition of estimated error. Let the size of aggregation flows be X, and its estimated value be $\hat{X}$, then we can get a relative estimated error metrics, $error = (X - \hat{X}) / X \times 100\%$, and a metrics about absolute value sum of relative estimated error, $sum_error = \sum_{i=1}^{n} (| X - \hat{X} | / X)$. Figure 2 is the relationship between the SPORT aggregation flow and their relative estimated error. Figure 3 shows a comparison based on sum error metrics between double-sampling algorithm and ANF algorithm and figure 4 gives the different percent relative error in the different time bins.

Table 2. The comparision between Estan ANF algorithm and Double-Sampling algorithm (DS). The initial sampling rate are all 1/2 in the two algorithms.

No	Flow memory	Ratio ANF	#packet ANF	#flow ANF	Ratio DS	#packet DS	#flow DS
1	4096	1/1024	9214	2708	1/512	3560390	3277
2	8192	1/256	35765	6790	1/128	3685570	7689
3	16384	1/128	72847	11603	1/64	3792411	12868
4	32768	1/32	260982	30183	1/16	3915760	28561
5	65536	1/16	486643	50297	1/8	4046263	49747
6	131072	1/4	1858527	112875	1/2	4181913	97189
7	262144	1/2	4217442	129862	1	4217442	129862

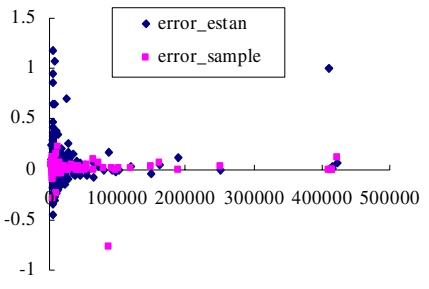


Figure 2(a)

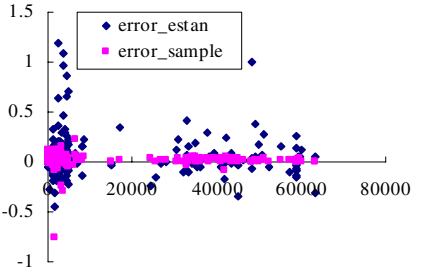


Figure 2(b)

Fig. 2. Relative estimated error of SPORT aggregation flows. In this experiment, we let flow cache be 8192, and experiment data is the first team data in the table 1, and only analyze the SPORT aggregation flows whose size is larger than 0.1% of the total packets. The figure 2(a) is the relationship between the aggregation size and relative estimated error, and figure 2(b) shows the relationship between the SPORT number and relative estimated error. X axis in the figure 2(a) means the size of sport aggregation flows, and X axis in the figure 2(b) means SPORT number. Y axis in the two figures also means relative estimated error.

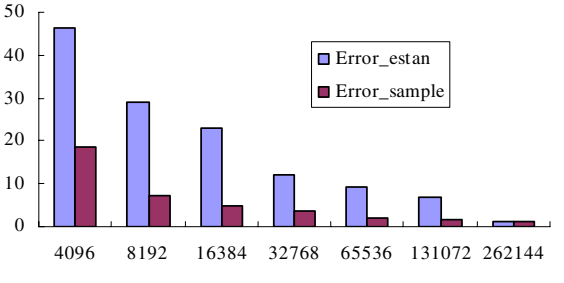


Fig. 3. A comparison based on metrics in equation (5) between double-sampling algorithm and Estan algorithm. X axis in this figure is set 4096, 8192, 16384, 32768, 65536, 131073, 262144 respectively. We notice that the error of Estan algorithm is larger than that of double-sampling algorithm, but when the flow cache size is equal to 262144, the two errors are equal. The reason is that the size of flow cache is larger than the flow number of the measured traffic, so the two algorithms will keep all flow information.

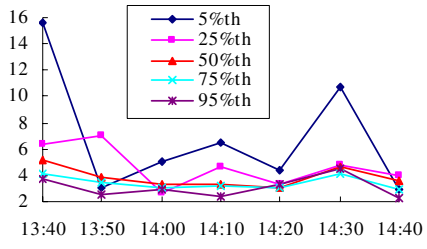


Figure 4(a)

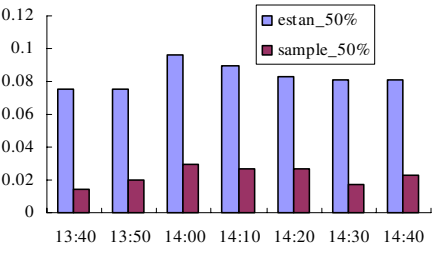


Figure 4(b)

Fig. 4. The different percent relative error in the different time bins. Figure 4(a) is the relative error ratio between Estan's algorithm and the double-sampling algorithm in 5%, 25%, 50%, 75%, 95% percentile respectively. Figure 4(b) is the relationship between Estan's algorithm and the double-sampling algorithm in 50% percentile. The two figures show that the relative error of Estan's algorithm is at least 3 times larger than that of the double-sampling algorithm.

The experiments show that the accuracy of the double-sampling algorithm is higher than that of Estan's algorithm with the same system resources. There are two reasons for this result. Firstly, Estan's algorithm loses partial packet information of each flow in the flow cache during renormalization. The Early Removal algorithm in this paper only removes least significant small flow entries, but keeps all information of heavy-tailed flows. Secondly, only sampled packets can be added into the flow cache, but the double-sampling and hold algorithm in this paper can collected all packets after their flows are recorded into the flow cache. Therefore, the double-sampling based approach can keep more packets than Estan's algorithm with the same flow cache size, and thus it can obtain more accurate traffic flow information.

5 Conclusion

This paper proposes an adaptive NetFlow algorithm on double sample and hold, which includes two sample process, one hold process, and one early removal process. We can control the flow number in the flow memory by the hold and early removal process, and hold process can record more packets into the flow memory, so the algorithm has two advantages: firstly, it can control and adapt the flow cache size; secondly, it can improve the estimated precision; thirdly, it can save the load of CPU during early removal. We also use NLANR data to compare the performance between the double sample ANF algorithm in this paper and the Estan ANF algorithm. We analyze the SPORT aggregation flows over 0.1% of the total traffic, and the experiments shows that the double sample algorithm has better precision than the Estan ANF algorithm under the same flow memory size and configure parameters.

Acknowledgments. This work has been supported by the Natural Science Foundation of Jiangsu Province under Grant No. BK2006092, the Key Project of Chinese Ministry of Education under Grant No. 105084, the National Grand Fundamental Research 973 program of China under Grant No. 2003CB314804, and the Excellent Youth Teacher of Southeast University Program under Grant No. 4009001018, the National Natural Science Foundation of China under Grant No. 50609006.

References

1. C. Estan, Ken Keys, David Moore, George Varghese, Building a Better Netflow, SIGCOMM, (August 2004)
2. Cisco IOS NetFlow Introduction, http://www.cisco.com/en/US/products/ps6601/products_ios_protocol_group_home.html
3. Sampled NetFlow Documentation, http://www.cisco.com/univercd/cc/td/doc/product/software/ios120/ 120newft/120limit/120s/120s11/12s_sanf.htm
4. C. Estan, G. Varghese.: New directions in traffic measurement and accounting. In SIGCOMM, (Aug. 2002).
5. Abilene-V Trace Data, http://pma.nlanr.net/Special/ipls5.html
6. Packet Sampling (psamp), http://www.ietf.org/html.charters/psamp-charter.html, (2002).
7. Duffield, N.G., Lund, C., Thorup, M.: Properties and Prediction of Flow Statistics from Sampled Packet Streams, ACM SIGCOMM IMW 2002, (November, 2002).
8. Feldmann, A. , Caceres, R. , Douglis, F. , Glass, G., Rabinovich, M.: Performance of Web Proxy Caching in Heterogeneous Bandwidth Environments, IEEE INFOCOM 99.
9. Estan, C. andVarghese, G., Bitmap algorithms for counting active flows on high speed links, in Proc. ACM SIGCOMM Internet Measurement Conference, (Oct. 2003).
10. Abhishek Kumar, Jun Xu, Li Li, and Jia Wang, Space Code Bloom Filter for Efficient Traffic Flow Measurement, ACM/USENIX IMC, Miami, FL, (October 2003).
11. Abhishek Kumar, Minho Sung, Jun (Jim) Xu and Jia Wang, Data streaming algorithms for efficient and accurate estimation of flow size distribution, ACM SIGMETRICS 2004.
12. Duffield, N.G., Lund, C. , Thorup, M.: Properties and Prediction of Flow Statistics from Sampled Packet Streams. ACM SIGCOMM IMW 2002, (November, 2002).
13. Duffield, N.G., Lund, C. , Thorup, M.: Estimating Flow Distributions from Sampled Flow Statistics. ACM SIGCOMM . 2003, Karlsruhe,Germany. (August 325-336).

A Power Saving Scheme for Heterogeneous Wireless Access Networks*

SuKyoung Lee, LaeYoung Kim, and Hojin Kim

Dept. of Computer Science, Yonsei University, Seoul, Korea
sklee@cs.yonsei.ac.kr

Abstract. In integrated WLAN and cellular networks, we propose a power saving scheme that completely turns off WLAN interface in the idle state and wakes it up when there is incoming data from long-lived traffic. We also develop a Mobile IPv6 (MIPv6)-based signaling to turn on the WLAN interface only for long-lived traffic. It is shown via simulations that the proposed power saving scheme improves power efficiency over typical WLAN system.

Keywords: heterogeneous networks, power saving, MIPv6.

1 Introduction

In the integrated Wireless LAN (WLAN) and cellular networks, power efficiency is still one of the most important aspects as in existing wireless networks. To keep the WLAN card operating, mobile nodes (MNs) with dual interfaces have to remain on, but in idle state. In [1], although the WLAN interface can be turned off in the idle state, MNs should still listen to the lower-power radio for the idle state. In [2], a power saving scheme is proposed to completely turn off the WLAN interface, nevertheless, this scheme cannot easily support third-party WLAN since it is based on tightly-coupled interworking architecture. Therefore, we propose a power saving scheme based on loosely-coupled interworking architecture to completely turn off WLAN interface of dual-mode MN and wake it up when there is a need to receive data. Further, a MIPv6-based signaling is developed to turn on the WLAN interface only for long-lived traffic so that the WLAN interface should not be turned on and off repeatedly for a momentary traffic, dissipating power. We show via simulations that the proposed power saving scheme improves power efficiency over existing WLAN.

2 Idle Power Saving Scheme

A typical WLAN interface has active and idle states. In idle state with power-save mode, power consumption level for the WLAN interface is several times as high as that of cellular interface [1]. In our scheme, to save the power in idle

* This work was supported by grant No.R01-2006-000-10614-0 from the Basic Research Program of the Korea Science & Engineering Foundation.

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 865–868, 2007.
© Springer-Verlag Berlin Heidelberg 2007

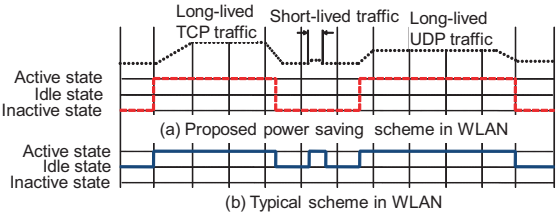


Fig. 1. Idle power saving for WLAN interface

state (so called "idle power"), WLAN interface is turned off without any periodic wake-up during the idle state, that is called "inactive" state in this paper and is woke up when there is incoming data from long-lived traffic depending on the duration and amount of data (e.g., file transfer, streaming service). Once the WLAN interface is turned on, it enters active state to receive the data and then returns to inactive state after the data session is completed as shown in Fig. 1. In this study, we focus on downlink traffic since it is envisioned that 4^{th} generation wireless system's traffic pattern will be highly asymmetrical. Our proposed scheme targets on the loosely-coupled interworking (based on MIP) the cost of which is lower compared to the tightly-coupled interworking because WLAN is deployed as an independent operator connected to the Internet.

In our system model, the cellular interface is assumed to be able to detect Access Points (APs) by listening to paging channel. Basing our system on 3GPP system, Gateway GPRS Support Node (GGSN) is responsible for controlling IP connections between users and external data networks (i.e. Internet) with its own buffer. To reduce the time taken to wake up the WLAN interface, each AP is configured with a special Service Set Identifier (SSID) string which provides prefix information of the attached 802.11 gateway as in [3], resulting in fast configuration of MN's new Care-of-Address (CoA). The SSID field in beacon message consists of prefix of WLAN gateway's IP address, slash, prefix-length, one blank and identity of service set (e.g., 2001:1302:A83B:1104/64 AP1). As shown in Fig. 2, the developed signaling procedures proceed as follows :

1. When a Correspondent Node (CN) has a long-lived traffic, it sets the flow label field in the IP packets to one (i.e., 0x00001) and two (i.e., 0x00002) for UDP and TCP traffics, respectively.
2. When the GGSN receives the packet whose flow label field is not zero, it sends a WAKEUP_WLAN message to the MN. This message is introduced as a new MIPv6 message whose Mobility Header (MH) type is 10. The GGSN keeps all the TCP packets of loss-sensitive application, destined to the MN in its buffer to prevent the packet loss while switching to the WLAN interface. On the other hand, the GGSN continuously sends the UDP packets of delay-sensitive application to the MN.
3. On receiving the WAKEUP_WLAN message, the MN turns on its WLAN interface and is associated with an appropriate AP by performing beacon

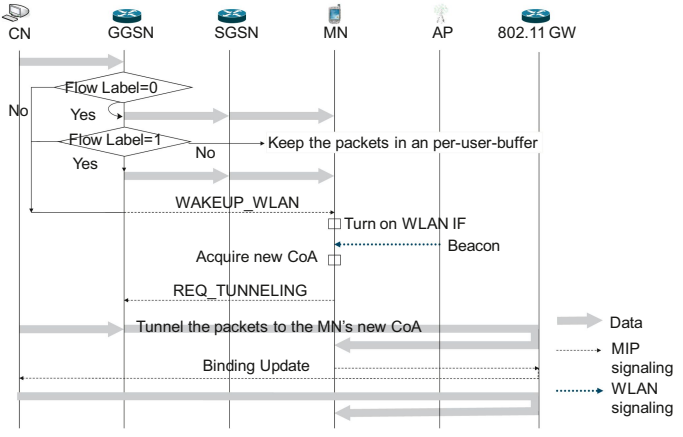


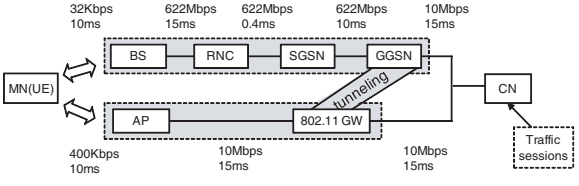
Fig. 2. MIPv6-based signaling procedures for incoming data traffic

scanning. Then, the MN configures a new CoA based on the prefix of the gateway's IP address obtained from the SSID field in the beacon message.

4. Once the new CoA is configured, the MN sends an REQ_TUNNELING message containing the new CoA, to the GGSN. This message is carried in a new MIPv6 message whose MH type is 11. The MN also sends a Binding Update (BU) message to inform the CN and its home agent of the new CoA.
5. On receiving the REQ_TUNNELING message, the GGSN tunnels on-the-fly packets as well as the buffered packets (if there is any) to the new CoA.
6. When the CN receives the BU, it sends the packets to the MN's new CoA indicated in the BU.

3 Performance Evaluation

Simulation experiments are carried to evaluate the performance of the proposed scheme by using ns-2 with wireless extensions. The data session holding time is exponentially distributed with means of $5min$ and $1min$ for long-lived and short-lived traffics, respectively. Fig. 3 shows the network topology and the powers (W)



(a) Network Topology

Interface	State	Power(W)
WLAN	active	1.65
	idle	1.15
Cellular	active	1.254
	idle	0.125

(b) Power Consumption

Fig. 3. Network topology and power consumption parameters

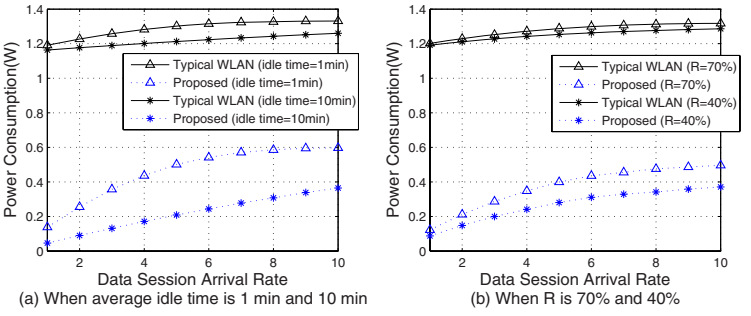


Fig. 4. Total power consumption versus data session arrival rate

consumed by WLAN and cellular network interfaces in active and idle modes, which are obtained from [1] and [4].

Fig. 4 (a) shows the total power consumption (W) versus data session arrival rate when the average idle time is $1min$ and $10min$. The power consumption for our scheme is less than the typical WLAN over all the ranges of data session arrival rate. The performance improvement by our scheme is higher when the average idle time is $10min$ than when $1min$. Fig. 4 (b) shows the total power consumption versus data session arrival rate for the portion of long-lived traffic among total data sessions, $R = 70\%$ and 40%. The proposed scheme achieves better power efficiency than the typical WLAN for the two values of R. However, the average improvement by the proposed scheme over the typical WLAN for $R = 70\%$ is smaller than that for $R = 40\%$ because the larger the portion of long-lived traffic is, the longer our scheme has to keep the WLAN interface on.

4 Conclusion

The simulation results showed that the proposed power saving scheme outperforms typical WLAN with respect to power saving by completely turning off the WLAN interface of dual-mode MN for the idle state and waking it up for long-lived traffic.

References

1. Shih, E., Bahl, P., Sinclair, M.: Wake on Wireless: An Event Driven Energy Saving Strategy for Battery Operated Devices. Mobicom, (Sep. 2002) 160-171.
2. Lee, S., Seo, S., Golmie, N.: An Efficient Power-Saving Mechanism for Integration of WLAN and Cellular Networks. IEEE Commun.Letters, Vol.9, No.12 (Dec. 2005) 1052-1054.
3. Jordan, N., Poropatich, A., Fleck, R.: Link-layer Support for Fast Mobile IPv6 Handover in Wireless LAN based Networks. IEEE LANMAN, (Apr. 2004) 139-143.
4. Baiamonte, V., Chiasserini, C.: Investigating MAC-layer Schemes to Promote Doze Mode in 802.11-based WLANs. IEEE VTC, Vol.3 (Oct. 2003) 1568-1572.

Efficient GTS Allocation Algorithm for IEEE 802.15.4

Youngmin Ji[1], Woojin Park[2], Sungjun Kim[2], and Sunshin An[2]

[1] Department of Telecommunication System Technology, Korea University, 5Ga 1, Anamdong, Sungbukku, Seoul 136-701, Republic of Korea
ymji@dsys.korea.ac.kr
[2] Department of Electronics and Computer Engineering, Korea University, 5Ga 1, Anamdong, Sungbukku, Seoul 136-701, Republic of Korea
{wjpark,sjkim,sunshin}@dsys.korea.ac.kr

Abstract. Recently issued standard IEEE 802.15.4 is a novel standard to achieve some critical features of wireless sensor networks. In this paper, we propose the algorithm to allocate GTSs according to the packet arrival rate from end devices and the number of devices in beacon enabled mode of IEEE 802.15.4 star topology networks. From simulation results, we show that our proposed algorithm enhances throughput comparing with the original IEEE 802.15.4 that uses only CAP duration for transmission[1].

Keywords: IEEE 802.15.4, Wireless Sensor Networks, GTS, Zigbee.

1 Introduction

In this paper, we introduce the algorithm to enhance network throughput by allocating GTSs efficiently in beacon enabled mode of IEEE 802.15.4 star topology networks. We use ns-2 simulations to prove our proposed algorithm. The simulation is performed by implementing our algorithm and GTS module in IEEE 802.15.4 of ns-2 [3][4]. Our simulation results indicate that proposed algorithm improves throughput comparing with the original IEEE 802.15.4.

2 Efficient GTS Allocation Algorithm

In this section, we present the proposed algorithm to allocate efficiently GTSs based on the information from end devices in IEEE 802.15.4 star topology networks. The information is the packet arrival rate and the number of devices. They are selected as the criteria of our algorithm due to follows: 1) Packet arrival rate: the devices with the higher packet rate can cause the more collisions and the much delay compared to devices with the lower packet rate during trying to transmit their data packets. Allocating GTSs to devices with the higher packet rate can minimize packet collisions and delay of the network. 2) Number

[1] This research was partly supported by SK Telecom (SKT) in Korea.

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 869–872, 2007.
© Springer-Verlag Berlin Heidelberg 2007

of devices: GTSs are designed to be used at ranges less than 8 slots in IEEE 802.15.4 [1]. Therefore, the coordinator can allocate the maximum 7 GTSs in the range of 0 through 6. Based on the criteria, our algorithm allocates one GTS slot to each node in order of high packet arrival rate and the maximum number of GTS allocation (n_G) depends on the size of the network as shown in (1).

$$n_G = \llcorner \frac{n}{2} \lrcorner \quad (\text{if } n \leq 15), \qquad n_G = 7 \quad (\text{if } n > 15) \tag{1}$$

In this algorithm, a coordinator manages each end device's T_{CA} (Current Arrival Time), T_{PA} (Previous Arrival Time), T_{CI} (Current Interarrival Time), T_{AI} (Average Interarrival Time) and δ (Parameter to control the relative weight of recent and past interarrival rate). If δ is 0, T_{AI} is ignored. If δ is 1, T_{CI} cannot affect T_{AI}. We have to choose δ at ranges of larger than 0 and less than 1. We set δ as 0.9, T_{AI} is more weighted than T_{CI} (Note: $T_{AIj} = T_{AIj}\delta + T_{CIj}(1-\delta)$ of Algorithm 1) since data packet is frequently received. The detailed sequence of the proposed algorithm is described in Algorithm 1, which runs in the coordinator.

Algorithm 1: *Record Interarrival Time*

```
01: while(TRUE)
02:     if i = m then
03:         Sort the T_AI in descending order;
04:         i = 0;
05:         changing Beacon using the top n_G of the T_AI
06:     else
07:         i = i + 1;
08:     end if
09:     receive message in CAP and CFP
10:     for j=1 to n (the number of end devices) then
11:         if end device j send message then
12:             T_CIj = T_CAj - T_PAj;
13:             T_PAj = T_CAj;
14:             T_AIj = T_AIj δ + T_CIj(1 − δ);
15:         end if
16:     end for
17: end while
```

3 Simulation

3.1 Simulation Environment

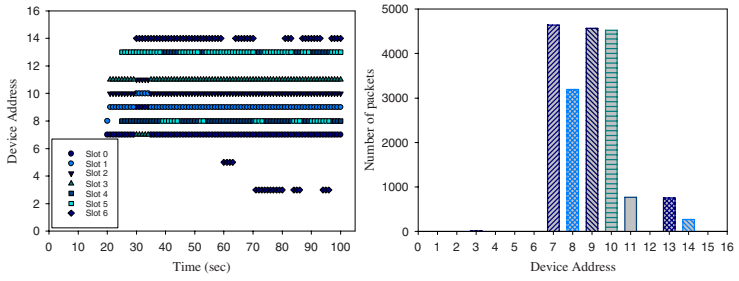
The network is a star topology that consists of 17 devices containing a PAN coordinator. We set *BO* (Beacon Order) and *SO* (Superframe Order) to 7 so that coordinator and end devices are always in active mode and never switch to sleep mode since we do not need to use the inactive periods. Simulations are conducted with three groups that consist of Group 1 (Device 1 to 6), Group 2

(Device 7 to 11) and Group 3 (Device 12 to 16). Each group has the different transmission rate (1, 0.01 and 0.1 in order of Group) to show that GTSs are dynamically allocated in order of transmission rates. The size of a transmitted packet is 80 bytes. The transmission rate of a device is 240 kbps.

3.2 Simulation Results

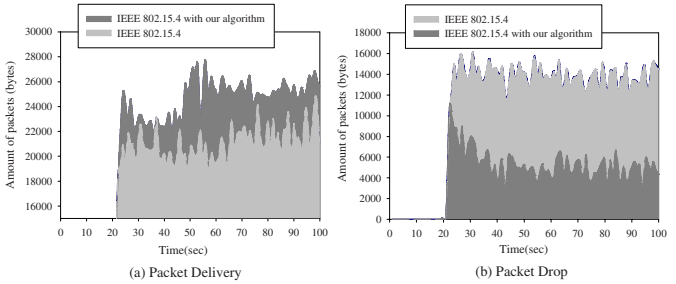
Fig.1 represents the history of allocating GTSs on each device. This result indicates that the occupation rate of time slots depends on the traffic rate. We can know the amount of successful transmissions on each device, which use the guaranteed time slots.

Fig.2 shows the comparison of packet delivery and packet drop. As expected, the average throughput of IEEE 802.15.4 is up to 168 kbps; however the average throughput of our proposed algorithm is up to 195 kbps. We can get 16% improvement in throughput in simulation. By using this result, we can understand the efficiency of proposed scheme.



(a) History of GTS allocation (b) Successful Packet Transmission

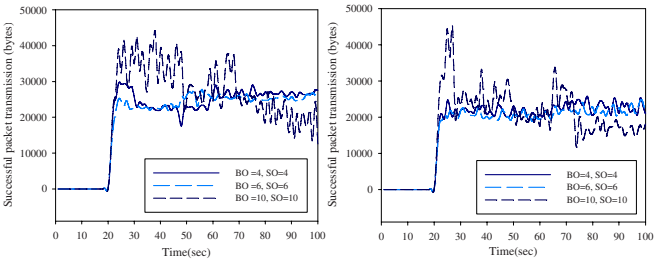
Fig. 1. GTS allocation and successful transmission on each device with our algorithm



(a) Packet Delivery (b) Packet Drop

Fig. 2. Comparison of packet delivery and drop number

In addition, when the device employs IEEE 802.15.4 with our algorithm, the amount of drop packets caused by a lot of collisions decreases as shown in Fig.2. The average packet drop rate of IEEE 802.15.4 is up to 110 kbps but average



(a) IEEE 802.15.4 with our algorithm (b) IEEE 802.15.4

Fig. 3. Comparison of successful transmission on various parameters

packet drop rate of our proposed algorithm is up to 43 kbps. This result also indicates that our proposed algorithm is more efficient than existing protocol. Fig.2 and Fig.3 represent the throughput of IEEE 802.15.4 with our algorithm and IEEE 802.15.4 with various parameters respectively. In a case of IEEE 802.15.4 with our algorithm, the average throughput is up to 206 kbps. In a case of IEEE 802.15.4, the average throughput is up to 174 kbps. While changing several parameters we have obtained better results through the IEEE 802.15.4 with our algorithm. Finally, we can get 18% improvement in average throughput in these scenarios.

4 Conclusion

In this paper, an efficient GTS allocation algorithm was proposed to guarantee the throughput for sensor devices and proved through the simulation. Our proposed algorithm showed the improved throughput comparing with the original IEEE 802.15.4 that uses only CAP duration for transmission. This algorithm is adaptable not only star topology but also multi-hop topology for beacon enabled mode.

References

1. IEEE Standard for Wireless Medium Access Control (MAC) and Physical Layer (PHY) Specification for Low-Rate Wireless Personal Area Networks (LR-WPAN), Oct. 2003
2. Z.Alliance, "Zigbee specification v1.0," June 2005
3. USC Information Sciences Institute, Marina del Rey, CA. Network Simulator - ns-2. (http://www.isi.edu/nsnam/ns)
4. Samsung/CUNY, "ns-2 simulator for IEEE 802.15.4," http://www-ee.ccny.cuny.edu/zheng/pub [referenced September 13, 2005]

Hybrid Search Algorithms for P2P Media Streaming Distribution in Ad Hoc Networks*

Dong-hong Zuo, Xu Du**, and Zong-kai Yang

Dept. of Electronics and Information Engineering, Huazhong Univerisity of Science and Technology, Wuhan, Hubei, 430074, China
{sixizuo,duxu,zkyang}@mail.hust.edu.cn

Abstract. Media streaming delivery in wireless ad hoc networks is challenging due to the stringent resource restrictions, potential high loss rate and the decentralized architecture. To support long and high-quality streams, one viable approach is that a media stream is partitioned into segments, and then the segments are replicated in a network and served in a peer-to-peer fashion, however, the searching strategy for segments is one key problem with the approach. This paper proposes a hybrid ant-inspired search algorithm (HASA) for P2P media streaming distribution in Ad Hoc networks. It takes the advantages of random walkers and ant-inspired algorithms for search in unstructured P2P networks, such as low transmitting latency and less redundant query messages. We quantify the performance of our scheme in terms of response time and network query messages for media streaming distribution. Simulation results show that it can effectively improve the Search Efficiency for P2P media streaming distribution in Ad Hoc networks.

Keywords: Ad Hoc networks, media streaming distribution, search algorithms, peer to peer.

1 Introduction

Media streaming distribution in wireless ad hoc networks is attractive and is also challenging due to stringent resource restrictions at the mobile hosts, dynamic network connectivity, and potentially high loss rate. To support media streaming applications with limited resources, one viable approach is: partitioning a media stream into segments and managing them in a peer-to-peer fashion [1, 2]. Unfortunately, this approach brings us challenges on searching for multiple segments and reassembling them in such unstructured P2P networks.

Flooding based search algorithms [1, 2] are adopted due to the reliability and low latency. However, they incur large volume of unnecessary traffic to networks. Statistics based searching algorithms [3, 4] are proposed to avoid the large volume of redundant messages. But they introduce the partial coverage problem. In order to effectively reduce the redundant traffic and alleviate the partial coverage problem,

* Supported by Province Natural Science Foundation of Hu Bei under Grant No.2005ABA264.
** Corresponding auther.

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 873–876, 2007.
© Springer-Verlag Berlin Heidelberg 2007

hybrid search algorithms [5] are proposed. But they are designed for the file sharing system and don't concern the correlation of continuous queries for media streaming segments.

We propose a hybrid ant-inspired search algorithm (HASA) for P2P media streaming distribution in Ad Hoc networks, coping with the P2P network churn and concerning the correlation of the searched media segments. It takes the advantages of K random walkers [4] and ant-inspired algorithms [6] for searching in unstructured media streaming distribution P2P networks.

2 Description of the Hybrid Ant-Inspired Search Algorithm

HASA uses the ant colony theorem [7] to collect and maintain the statistics information. At each peer P_i, pheromone trails are maintained in a pheromone table of size m×n, where m is the number of the media segments and n is the number of peer Pi's outgoing links to neighbor peers P_u where $\{1, ..., n\}$.

2.1 Overview of HASA

Every peer in the network behaves as a media segments query originator peer, media segments providing peer, and/or intermediate peer. When a peer wants to find some media segment, a forwarding ant is generated. Originator peers and intermediate peers forward the forwarding ants according the forwarding rule. When the query is hit at one peer, a backward ant is generated, and is sent back following the reverse path of its corresponding forward ant. When ants move in the P2P network, they collect the paths' information and update the pheromone tables of peers in their paths according to the pheromone updating rule.

2.2 Forwarding Rule

Both intermediate and originator peers forward the ants in the same way. When a forward ant is received or generated, this ant will be forwarded to K neighbor peers except the incoming one. The next peer selection is done using the pheromone tables. If the amount of pheromones of all neighbors for the segment that the forward ant is searching for is τ_{init}, then the forwarding ant will be forwarded to K randomly selected neighbor peers. Otherwise, if more than K links' pheromones for segment s are refreshed, the forwarding ant will be forwarded to these K neighbors (K is far less than the number of neighbor peers.) that the amount of pheromone is more than the rest. If only L (L<K) links' pheromones are updated, the forwarding ant will be forwarded to these L neighbors, and M($M \leqq K-L$) neighbors those have more pheromone trails for segment s-1. If M+L<K, then other K-M-L neighbors will be randomly selected from the rest to be forwarded to.

2.3 Pheromone Updating Rule

In HASA the pheromone table is updated when peers receive either forward ants or backward ants. When a peer joins the media distribution network, we calculate the elapsed time into equal intervals. At startup, all peers' pheromone table entries are

initialized with the same small value τ_{init}. When receiving ants, peers update pheromone table using function (1). At the beginning of the nth interval, peers update pheromone table using function (2).

$$\tau_{s,i,j}(n)=\begin{cases}\tau_{s,i,j}(n)+\nabla\tau_{s,i,j}(n,a_m) & \text{the incoming link}\\ \tau_{s,i,j}(n) & \text{other links}\end{cases} \quad (1)$$

$$\tau_{s,i,j}(n)=\tau_{s,i,j}(n-1)\rho_s \quad (2)$$

Where $\tau_{s,i,j}(n)$ is the pheromone value for segment s corresponding to neighbor j at node i after passing n intervals. ρ_s is the evaporation parameter of pheromone trails for segment s. $\nabla\tau_{s,i,j}(n,a_m)$ is the pheromone left by the mth ants for segment s corresponding to neighbor j at node i in the nth interval. The amount of $\nabla\tau_{s,i,j}(n,a_m)$ are different at different peers which is left by the same ant. Example of such function is: $\nabla\tau_{s,i,j}(n,a_m)=w_pI_{path,i}+w_h(C-h_{s,i})$, where C is a constant number, it should be larger than the longest hop for searching any segment s in the network. $h_{s,i}$ is the hop that the ant has traveled from its generating peer to peer i. $I_{path,I}$ is the path information from the originator or providing peer to the intermediate peer i. w_p is the path information weight, and w_h is the hop weight, they control the $w_p+ w_h=1$. The path information is collected by ants.

3 Performance Evaluation

HASA is evaluated by simulating a peer-to-peer system with 1600 peers and by comparing its performance against that of the well-known K random walkers search[5] approaches. In this experiment, a long media stream is divided into 20 segments. Initially, all segments are stored in one peer in the network. Then streaming accesses are triggered at random points in the network. To model the popularity of the segments, a Zipf-like distribution is used. Each node has cache spaces for 3 segments. In the experiment described here, the parameter values were chosen as follows: K=2, ρ_s=0.2, w_p=w_h=0.5, TTL = 15. To provide for comparison fairness, the TTL and K of K random walkers algorithm are set to 15 and 2 separately.

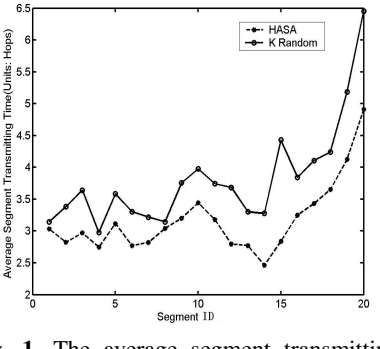


Fig. 1. The average segment transmitting delay

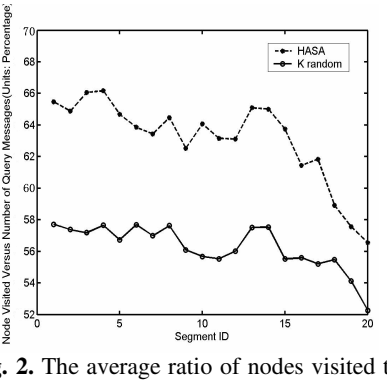


Fig. 2. The average ratio of nodes visited to number of query messages

Fig.1 and Fig.2 show the performance comparison of the media transmitting latency and messages efficiency for K random walkers and the hybrid ant-inspired search algorithms.The simulation results show that the hybrid search algorithm is better than the K random walkers algorithm. HASA algorithm forward queries following the trails for latter generated searches, but K random walks still forwards queries randomly, so HASA algorithm can find the queried segment more quickly, and it can effectively avoid forwarding the queries to the useless peers.

4 Conclusion

HASA uses the K random walkers based search strategy for searching a segment at the beginning, when receiving queries and responses, peers update their pheromone tables immediately, then use the pheromone trails to forward k ants for searching segments at each intermediate peers or originator peers. It collects the advantages of k random walkers and ant-inspired search algorithms, it exploits the correlation of streaming access. It can reduce most redundant messages, but still keeps the low transmitting latency for media streaming distribution in wireless Ad Hoc networks.

References

1. Jin, S.: Replication of Partitioned Media Streams in Wireless Ad Hoc Networks. Proc. of the 12th ACM Int. Conf. on Multimedia. ACM press, New York (2004) 396 - 399
2. Shahram, G., Bhaskar, K. and Song, S.: Placement of Continuous Media in Wireless Peer-to-Peer Networks. IEEE Trans Multimedia, Vol. 6, No. 2. IEEE Computer Society, New Jersey (2004) 335-342
3. Yang, B. and Garcia-Molina, H.: Improving search in peer-to-peer networks. Proc. of the 22nd Int. Conf. on Distributed Computing Systems. IEEE Computer Society, Washington (2002) 5-14
4. Qin, L., Cao, P., Edith, C., Li, K. and Scott, S.: Search and Replication in Unstructured Peer-to-Peer Networks. Proc. of the 16th Int. Conf. on Supercomputing. ACM Press, New York (2002) 84-95
5. Gkantsidis, C., Mihail, M. and Saberi, A.: Hybrid search schemes for unstructured peer-to-peer networks. Proc. of IEEE INFOCOM 2005. IEEE Computer Society, New York (2005) 1526-1537
6. Yang, K.H., Wu, C.J., Ho, J.M.: AntSearch: An ant search algorithm in unstructured peer-to-peer networks. IEICE Trans Commun, Vol. E89-B, No. 9. Oxford University Press, Bunkyo,Tokyo (2006) 2300-2308

Improving Search on Gnutella-Like P2P Systems

Qi Zhao, Jiaoyao Liu, and Jingdong Xu

Department of Computer Science, Nankai University,
Tianjin 300072, China
{qizhao6688,ljy,xjd}@ mail.nankai.edu.cn

Abstract. Gnutella has many weaknesses. Queries are handled identically no matter how popular the querying objects are. Blind flooding causes a great amount of redundant messages. In this paper, we explore how to retain the simplicity of Gnutella, while addressing its inherent weaknesses. We propose GPP, a content location solution in which multiple proxy peers are adopted to help a source peer locate desired objects. The underlying philosophy is to route queries to the areas where more results can be found. Towards this end, peers in GPP collaborate to achieve better search performance. Simulation results show that GPP outperforms Modified BFS and s-APS.

Keywords: Unstructued P2P, Search, Efficiency.

1 Introduction

Improvements to Gnutella's flooding mechanism have been studied along three dimensions: blind search, guided search, and group-based search. GPP (Gnutella with Proxy Peers), the search mechanism proposed in this paper, lies in the category of guided search mechanisms. Unlike most of the guided search mechanisms, GPP does not maintain state of the outgoing links or immediate neighborhood. Each peer maintains the performance of its proxy peers, which are distributed over the whole network. The search process of GPP is adaptive such that a targeted performance level can be achieved by the search.

In a guided search, peers locally store metadata that assist in the query process. In DBFS [1], queries are forwarded to a set number of neighbors that are more likely to return successful results. Intelligent BFS [2] uses peer ranking mechanism to forward queries to "good" neighbors. Local Indices [1] and Routing Indices [3] improve search performance via indexing objects stored at other peers. The forwarding process in APS [4] is probabilistic based on feedback from previous searches. In Popularity-Biased Random Walks [5], each step of the random walks is determined based on the content popularities of current neighbors. The authors of [6] introduce a number of local search strategies that utilize high degree peers. [7] proposes an adaptive search mechanism that uses estimate of popularity of a resource in order to choose the parameters of random walk.

2 System Design

Every peer in the system can ask any other peer to be its proxy peer. When a new peer joins the system, it does not know any peer except its neighbors. Therefore, it first

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 877–880, 2007.
© Springer-Verlag Berlin Heidelberg 2007

attempts to get proxy peers from its neighbors. The communication process consists of three steps. First, the new peer sends a *Ping_Proxy* message to its neighbors. Next, each neighbor selects a set number of proxy peers from its proxy list. It then sends a *Pong_Proxy* message back to the new peer, which contains the IP addresses of the selected proxy peers. The *Ping_Proxy* and *Pong_Proxy* messages are modified from the *Ping* and *Pong* messages in Gnutella 0.6 protocol [8].

A peer *P* 's proxy list consists of two parts: a *Primary Proxy List* (*PPList*) and a *Secondary Proxy List* (*SPList*). Each item in the proxy list is the IP address of a proxy peer. When *P* receives a *Pong_Proxy* message, it first extracts the IP addresses of the proxy peers, and then inserts them into its *PPList*. If its *PPList* is full, these IP addresses will be inserted into its *SPList*. Since *P* knows the IP addresses of its proxy peers, it can directly communicate with them.

PPList and *SPList* have different roles. When *P* initiates a query, it first forwards the query message to some items in its *PPList*. *P* will visit *SPList* only if all the items in its *PPList* have been used and the search is still unsuccessful. The items in *SPList* can enter *PPList* if they outperform the items in *PPList*. Through competition the best items are selected to stay in *PPList*. Obviously, *SPList* is used to memorize the IP addresses of many other peers. It gives *P* more choices when *PPList* is not enough for a search. *PPList* should be small and fresh. Contrarily, as the backup of *PPList*, *SPList* should be large. It is unnecessary and costly to maintain *SPList* fresh.

In GPP, *P* maintains its *PPList* fresh by periodically executing the following procedures. *P* first selects an item *Q* from its *PPList*, and then sends a *Ping_Proxy* message to *Q*. If *Q* does not respond, *P* will know that *Q* has left the system. In that case, *P* will delete *Q* from its PPList. If *Q* is alive, it will respond with a *Pong_Proxy* message, which contains some items selected from *Q* 's *PPList*. *P* will insert these items into its *PPList* with certain probability. The remaining items will be inserted into *P* 's *SPList*. Hence, peers can share proxy peers with each other. If *P* is not a proxy peer of *Q*, *Q* will insert *P* into its *PPList* or *SPList* with probability. These procedures will continue until *P* leaves the system.

2.1 Query Process

When a source peer *S* initiates a query, it forwards the query messages to some of its proxy peers. On receiving the query message, a proxy peer *X* initiates a local search on behalf of *S*. This local search starts a MBFS with dynamic fraction parameter f_h at different hops h when $h \leq n$. For $h > n$, it switches to the Random Walks. This combination of MBFS and Random Walks can greatly cut down redundant query messages [9].

Another important scheme of GPP is *Adaptive Termination*. To explain our idea, we first list the symbols used in this section. N_{mf} denotes the maximum number of proxy peers to query in a search process. N_{rd} denotes the number of results desired. S_{ppl} ($S_{ppl} \leq N_{mf} - 1$) denotes the size of *PPList*. If *S* decides to forward a query to *K* ($1 \leq K \leq N_{mf}$) proxy peers each time, then what is an appropriate value for *K*? If $K = N_{mf}$, the maximum number of queries will be initiated simultaneously. Therefore, we call this search scheme *Parallel Search*. It means that *S* wants the shortest response time. If $K = 1$, *S* will query its proxy peers one by one, until N_{rd} results have been received.

It shows that S aims to minimize redundant query messages. If $1 < K < N_{mf}$, the search process will be a trade-off between response time and query traffic. Given $1 \leq K < N_{mf}$, S will stop querying its proxy peers if N_{rd} results have been received. The search process is controlled by the source peer S, and it can be terminated adaptively. Therefore, we call this scheme *Adaptive Termination*.

In an *Adaptive Termination* search, the source peer S first selects the items in its *PPList*. If all the items in its *PPList* have been selected and S has not received N_{rd} results, it will select more proxy peers from its *SPList*. If S has sent a query to an offline proxy peer, it will select another proxy peer to forward the query. This process continues until S has sent the query to an online proxy peer. If S determines to initiate N_{mf} queries simultaneously, it will use all the items in its *PPList* and select $N_{mf} - S_{ppl}$ items from its *SPlist*.

2.2 Ranking Proxy Peers

Since $S_{ppl} < N_{mf}$, *SPList* is visited in every *Parallel Search* process. Hence, items in *PPList* can be compared with those in *SPList* after each process. The performance of proxy peers are evaluated in terms of *Query Hits*. The worst performer in *PPList* will be deleted, and the best performer in *SPList* will take its place.

The ranking process in *Adaptive Termination* search is a little complex. If the querying object is popular, the visited proxy peers are likely to be only within the *PPList*. In such cases, we are unable to compare *SPList* items with *PPList* items. Therefore, we need another *PPList* update strategy. We describe the strategy through the following example.

Peer X 's *PPList* is $\{P_0, P_1, P_2, P_3, P_4, P_5, P_6, P_7, P_8\}$. It is divided into two segments: $\{P_0, P_1, P_2, P_3, P_4\}$ and $\{P_5, P_6, P_7, P_8\}$. X initiates an *Adaptive Termination* Search. It receives N_{rd} *Query Hits* after querying l proxy peers.

If $1 \leq l \leq 5$, the queried proxy peers are within the first segment. Take $l = 3$ for example. The queried proxy peers are P_0, P_1, and P_2. Assume the *Query Hits* rank is $\{P_2, P_0, P_1\}$. Because P_1 has the lowest rank, it is transferred to the second segment, and P_5 is promoted to the first segment. The updated *PPList* is $\{P_2, P_0, P_3, P_4, P_5, P_1, P_6, P_7, P_8\}$.

If $6 \leq l \leq 9$, both segments will be visited. Take $l = 7$ for example. The queried proxy peers are $\{P_0, P_1, P_2, P_3, P_4\}$ from the first segment and $\{P_5, P_6\}$ from the second segment. Assume the *Query Hits* rank is $\{P_4, P_0, P_3, P_5, P_1, P_2, P_6\}$. P_6, an item in the second segment, has the lowest rank. It will be eliminated from *PPList*, and an item from *SPList* will be inserted into *PPList*. The updated *PPList* will be $\{P_4, P_0, P_3, P_5, P_1, P_2, P_7, P_8, P_{SPList}\}$. There is another case, in which the *Query Hits* rank is $\{P_4, P_0, P_3, P_5, P_1, P_6, P_2\}$. P_2, an item in the second segment, has the lowest rank. The *PPList* will be updated into $\{P_4, P_0, P_3, P_5, P_1, P_6, P_2, P_7, P_8\}$.

In the other cases ($10 \leq l \leq N_{mf}$), the worst performer in *PPList* will be deleted, and the best performer in *SPList* will be inserted into *PPList*. *PPList* will be updated according to the *Query Hits* rank.

3 Conclusion and Future Work

In this paper, we proposed GPP, an efficient search mechanism in unstructured P2P networks. The underlying philosophy is to route queries to the areas where more results can be found. Towards this end, peers in GPP collaborate to achieve better search performance. We discussed a variety of design issues and trade-offs in implementing GPP. As future work, we plan to give more formal analysis to GPP and other guided search methods through mathematical modeling.

References

1. B. Yang, H. Garcia-Molina: Improving Search in Peer-to-Peer Systems, *In Proceedings of ICDCS*, (2002).
2. V. Kalogeraki, D. Gunopulos, D. Zeinalipour-Yazti: A Local Search Mechanism for Peer-to-Peer Networks, *In Proceedings of CIKM*,(2002).
3. A. Crespo, H. Garcia-Molina: Routing Indices for Peer-to-Peer Systems, *In Proceedings of ICDCS*, (2002).
4. D. Tsoumakos, N. Roussopoulos: Adaptive Probabilistic Search (APS) for Peer-to-Peer Networks, Technical Report CS-TR-4451, Univ. of Maryland, (2003).
5. M. Zhong, K. Shen: Popularity-Biased Random Walks for Peer-to-Peer Search under the Square-Root Principle, *In Proceedings of IPTPS*, (2006).
6. L. A. Adamic, R. M. Lukose, A. R. Puniyani, B. A. Huberman: Search in Power-Law Networks, *Physical Review E*, 64, (2001).
7. N. Bisnik, A. Abouzeid: Modeling and Analysis of Random Walk Search Algorithms in P2P Networks, *In Proceedings of Second International Workshop on Hot Topics in Peer-to-Peer Systems*, (2005).
8. The Gnutella protocol specification 0.6, http://rfcgnutella. sourceforge.net.
9. H. Wang, T. Lin: On Efficiency in Searching Networks, *In Proceedings of IEEE INFOCOM*, (2005).

Non-preemptive Fixed Priority Scheduling of Hard Real-Time Periodic Tasks*

Moonju Park

Ubiquitous Computing Lab., IBM Korea, Seoul, Korea
mjupark@kr.ibm.com

Abstract. This paper addresses the problem of scheduling periodic tasks on a uniprocessor using static priority assignment without preemption. The problem of non-preemptive fixed priority scheduling has received little attention until recently, while real-life applications are often based on non-preemptive systems especially in case of embedded systems. In this paper, we show that Rate Monotonic priority assignment is optimal for non-preemptive scheduling when each task's relative deadline is equal to its period. We have derived a schedulability bound for non-preemptive Rate Monotonic scheduling by using period ratio of the tasks to provide a guarantee that tasks will meet their deadlines. Since the obtained bound is relatively small comparing with the preemptive scheduling, we also propose a method for designing high-utilization non-preemptive system to enhance the utilization bound.

Keywords: real-time scheduling, non-preemptive, periodic tasks, rate monotonic, utilization bound.

1 Introduction

In this paper, we address the problem of fixed priorities non-preemptive scheduling of hard real-time periodic tasks. In many real-time scheduling problems such as I/O scheduling, preemption is often impossible or prohibitively expensive due to properties of hardware device and software handling the device [5]. For example, [10] showed that non-preemptive fixed priority scheduling can be used for real-time signal processing applications because the amount of processor state including the stack to be stored can be reduced. Also non-preemptive scheduling on uni-processor systems eliminates the synchronization overhead of resource protecting mechanism. Since we can expect much lower overhead and smaller memory requirement at run-time with non-preemptive scheduling, non-preemptive scheduling is widely used in embedded systems, especially devices such as cell phones whose behaviour largely depends on network I/O.

Non-preemptive schedulers have received less attention than the preemptive ones, which have been extensively studied for both fixed and dynamic priority assignments since well-known results of Liu and Layland [8]. It is known that the

* This work was supported in part by MIC & IITA through IT Leading R&D Support Project.

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 881–888, 2007.
© Springer-Verlag Berlin Heidelberg 2007

general problem of finding a feasible schedule in an idling non-preemptive context is NP-complete. But some results are known for non-idling dynamic priority non-preemptive scheduling. [5] showed that Earliest Deadline First (EDF) is optimal among non-idling schedulers, non-preemptive scheduling of concrete periodic tasks is NP-Hard in the strong sense, and a necessary and sufficient condition for schedulability under non-idling EDF. It was shown in [4] that the worst case response time can be calculated using the results of [12].

Contrary to the preemptive context and dynamic priority scheduling, less results are known about fixed priority non-preemptive scheduling. It was shown that Deadline Monotonic is not optimal when relative deadline differs from period [4], but we do not know yet whether there is an optimal fixed priority scheduler when relative deadline is equal to period. Since there is little guidance for designing fixed priority non-preemptive systems, engineers in such an environment as small embedded devices are developing their systems by trial-and-error. They build systems with a certain priority assignment, and if the system crashes, they add more resources, change periods, or change priorities. The main purpose of this paper is to establish results on schedulability and optimal schedulers, and to give a guidance of designing fixed priority non-preemptive systems, for example, utilization bound.

This paper is organized as follows. Section 2 summarizes assumptions and terms used in this paper. Schedulability conditions and optimal schedulers will be discussed in section 3. Section 4 gives a utilization bound for non-preemptive Rate Monotonic scheduler, and some guides for achieving higher utilization. Finally, section 5 concludes our work.

2 Computational Model and Assumptions

A periodic task is denoted by τ_i. A periodic task set is represented by the collection of periodic tasks, $\tau = \{\tau_i\}$. Each τ_i is associated with (T_i, D_i, C_i); period, relative deadline, and the worst case computation time respectively. The total utilization of a periodic task set τ is given by $U_\tau = \sum_{\tau_i \in \tau} \frac{C_i}{T_i}$. For a given task τ_i, we define $hp(\tau_i)$ as the subset of τ consists of tasks with priority equal to or higher than τ_i. On the other hand, $lp(\tau_i)$ is the set of tasks with lower priority than τ_i.

A concrete task has a specified release time, or the time of the first invocation. The difficulty of scheduling tasks can be affected by the release time. A periodic task set is said to be schedulable if and only if all concrete task sets that can be generated from the periodic task set are schedulable. We will consider only periodic task sets in this paper.

The following is a summary of assumptions used in this paper.

- There is only one processor.
- Scheduling overhead can be ignored.
- $C_i \leq D_i \leq T_i$
- Tasks are sorted in non-decreasing order by period. That is, for any pair of tasks τ_i and τ_j, if $j < i$ then $T_j \leq T_i$.

– Tasks are all independent and cannot be suspended by themselves.
– Tasks become ready when arrive. (*i.e.*, there is no inserted idle time.)

3 Schedulability of Scheduling Non-preemptive Periodic Tasks

The following lemma from [4] shows the concept of level-i busy period [7] is also useful in non-preemptive scheduling.

Lemma 1. *A periodic task τ_i has the largest response time in a level-i busy period obtained by releasing all tasks τ_j with $\tau_j \in hp(\tau_i)$ simultaneously at time $t = 0$, and the task τ_k with $\tau_k \in lp(\tau_i)$ and $C_k = \max\{C_l | l > i\}$ just before time $t = 0$.*

Readers may notice that the level-i busy period given in Lemma 1 is basically simialr to the critical interval of dynamic priority driven scheduler in [5]. The interval given in Lemma 1 is an extension of the concept of the critical instant in [8].

Based on Lemma 1, the following theorem establishes a necessary and sufficient condition for schedulability for any periodic tasks with fixed priorities.

Theorem 1. *A periodic task set τ is schedulable using a fixed priority scheduler if and only if*

$$C^i_{\max} + \sum_{\tau_j \in hp(\tau_i)} \left\lfloor \frac{D_i}{T_j} \right\rfloor C_j \leq D_i \tag{1}$$

for $\forall \tau_i \in \tau$ where $C^i_{\max} = \max\{C_k | k > i\}$.

Proof. Theorem 1 is derived directly from Lemma 1 and non-preemptibility of tasks.

(if-part) Since a level-i busy period starts from time $t = 0$, the maximum processor demand of the tasks with priority higher than or equal to τ_i before the deadline of τ_i is given by:

$$\sum_{\tau_j \in hp(\tau_i)} \left\lfloor \frac{D_i}{T_j} \right\rfloor C_j$$

Note that we use floor function rather than ceiling function, because tasks are non-preemptive. Since tasks are non-preemptive, we do not need to examine the last instances released after the level-i busy period. Because a task with lower priority than τ_i is already executing just before time $t = 0$, the execution time of the task $C^i_{\max}$ should be accounted. Thus the total processor demand in the level-i busy period is less than or equal to:

$$C^i_{\max} + \sum_{\tau_j \in hp(\tau_i)} \left\lfloor \frac{D_i}{T_j} \right\rfloor C_j$$

By Lemma 1, if τ_i meets its deadline in the level-i busy period, it will never miss its deadline. Therefore if all tasks satisfy equation (1), the task set is schedulable by a fixed priority scheduler.

(only-if-part) Let us assume that τ is schedulable but the equation (1) is not satisfied for some τ_i. Using Lemma 1, we can make a concrete periodic task set such thas τ_i has the processor demand equal to lefthand side of the equation (1). Since the processor demand exceeds its period, τ_i should miss its deadline, which leads to a contradiction.

The condition in Theorem 1 is similar to the results of Priority Ceiling Protocol (PCP) [11] or Stack Resource Policy (SRP) [3] by considering the processor as a shared resource. However, using Theorem 1, we can determine the schedulability problem of non-preemptive fixed priority scheduling in polynomial time, while the results for preemptive scheduling has pseudo-polynomial time complexity.

Now we will investigate the problem of optimal priority assignment for non-preemptive scheduling. Suppose that each task τ_i has the same relative deadline as its period, *i.e.*, $D_i = T_i$. In this case, it is well known that Rate Monotonic (RM) scheduling algorithm [8] is optimal for preemptive scheduling. The following theorem shows RM scheduling is also optimal for non-preemptive scheduling.

Theorem 2. *Rate Monotonic is an optimal priority assignment for non- preemptively scheduling periodic task sets with* $D_i = T_i$.

Proof. Let $\tau = \{\tau_1, \tau_2, \ldots, \tau_i, \tau_{i+1}, \ldots, \tau_n\}$ be a set of n tasks sorted in non-decreasing order by period. Suppose that a scheduling algorithm assigns τ_{i+1} higher priority than τ_i and the task set is schedulable using the priority assignments. Let us consider Rate Monotonic priority assignment with τ_i being the higher priority one by interchanging the priorities of τ_i and τ_{i+1}. For task τ_i, the interference due to higher priority task is reduced by C_{i+1} and $C^i_{\max}$ is given by $\max\{C_j | j = i+1, i+2, \ldots, n\}$. And $C^i_{\max}$ for τ_i when it has a lower priority than τ_{i+1} was $\max\{C_j | j = i+2, i+3, \ldots, n\}$. The interference τ_i experience after priority exchange is

$$\max\{C_j | j = i+1, i+2, \ldots, n\} \leq \max\{C_j | j = i+2, i+3, \ldots, n\} + C_{i+1}. \quad (2)$$

Because the total interference for τ_i is reduced, τ_i will remain schedulable .

Now let us consider task τ_{i+1}. It is easy to see that lowering the priority does not increase $C^{i+1}_{\max}$. The only increased interference due to the priority exchange is τ_i's execution, $\left\lfloor \frac{T_{i+1}}{T_i} \right\rfloor C_i$. As pointed out in [8], if when τ_{i+1} has higher priority than τ_i execution time of $C_{i+1} + C_i \leq T_i$ can be serviced during T_i, the increased interference time due to priority exchange, $\left\lfloor \frac{T_{i+1}}{T_i} \right\rfloor C_i$, can be successfully handled during T_{i+1} because $\left\lfloor \frac{T_{i+1}}{T_i} \right\rfloor T_i \leq T_{i+1}$. Therefore τ_{i+1} will meet its deadlines also with Rate Monotonic priority assignment. As stated in [8], in non-preemptive scheduling also, since the Rate Monotonic priority assignment can be obtained from any priority ordering by a sequence of pair-wise priority reordering as above, we prove the theorem.

For tasks with $D_i \leq T_i$, it is known that Deadline Monotonic priority assignment [2] is optimal for preemptive scheduling. But for non-preemptive scheduling, it was shown in [4] that the Deadline Monotonic is no longer optimal. They also showed that the optimal priority assignment method for preemptive scheduling described in [1] is still optimal in non-preemptive systems. Therefore, since the priority assignment method runs in $O(n^2)$, we can obtain a valid priority assignment method for tasks with $D_i \leq T_i$ in polynomial time by using Theorem 1 and the assignment method of [1].

4 Utilization Bound

We can find a task set with arbitrarily small utilization which is unschedulable by non-preemptive fixed priority scheduling. However, by restricting the ratio of minimum and maximum period (*i.e.*, T_n/T_1), we can achieve relatively higher utilization bound. The following theorem gives us a utilization bound with which task sets can be successfully scheduled by Rate Monotonic non-preemptive scheduler.

Theorem 3. *A set of n periodic tasks with $D_i = T_i$ is schedulable by the non-preemptive Rate Monotonic scheduler if*

$$\sum_{i=1}^{n} \frac{C_i}{T_i} \leq \frac{1}{r} \tag{3}$$

where $T_n = rT_1$.

Proof. From equation (3), $T_1 \geq T_n \sum_{i=1}^{n} \frac{C_i}{T_i} = \sum_{i=1}^{n} \frac{T_n}{T_i} C_i$. For a given task τ_i,

$$C^i_{\max} + C_i + \sum_{j=1}^{i-1} \left\lfloor \frac{T_i}{T_j} \right\rfloor C_j \leq C^i_{\max} + C_i + \sum_{j=1}^{i-1} \frac{T_n}{T_j} C_j \leq T_n \sum_{j=1}^{n} \frac{C_j}{T_j} \leq T_1 \leq T_i$$

The task set is schedulable by Theorem 1 since above equation holds for $\forall \tau_i$.

As shown in Theorem 3, the utilization of a schedulable task set decreases as the ratio of the maximum and the minimum period increases. It was shown in [6] that a preemptive task set is schedulable by RM scheduler if the utilization of the task set is less than or equal to $\ln r + 2/r - 1$ where $r = T_n/T_1$ and $1 \leq r \leq 2$. This utilization bound is further enhanced in [9] using additional information of T_n/T_{n-1}, but we will use the result of [6] in this paper for comparison purpose. Fig. 1 compares the utilization bound for preemptive and non-preemptive scheduling using these results. As shown in Fig. 1, the utilization bound for non-preemptive fixed priority scheduling becomes smaller in comparison with preemptive scheduling as the period ratio increases.

To further enhance the utilization bound, we can make use of the following theorem.

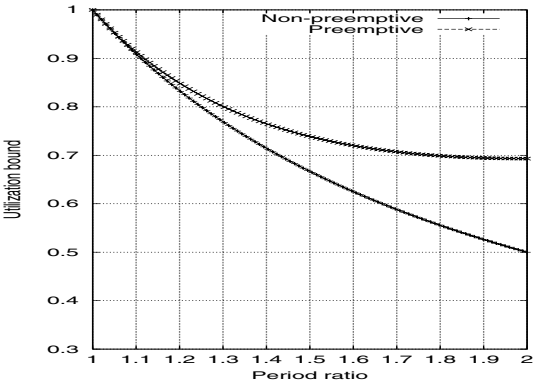


Fig. 1. Utilization bound for preemptive and non-preemptive scheduling for different values of r

Theorem 4. *A task set τ of n tasks with $D_i = T_i$ is schedulable by the non-preemptive Rate Monotonic scheduler if*

$$\max_{\tau_i \in \tau} \left\{ \frac{C_i}{T_i} \right\} \leq \frac{1}{r+n} \tag{4}$$

where $T_n = rT_1$ and $n \geq 2$.

Proof. For a given task τ_i, let us suppose $C^i_{\max} = C_k$. By dividing the lefthand side of equation (1) by T_i, we have

$$\frac{C^i_{\max}}{T_i} + \frac{C_i}{T_i} + \frac{1}{T_i} \sum_{j=1}^{i-1} \left\lfloor \frac{T_i}{T_j} \right\rfloor C_j \leq \frac{C_k}{T_k} \cdot \frac{T_k}{T_i} + \frac{C_i}{T_i} + \sum_{j=1}^{i-1} \frac{C_j}{T_j} \leq$$

$$\frac{r}{r+n} + \sum_{j=1}^{i} \frac{C_j}{T_j} \leq \frac{r}{r+n} + \sum_{j=1}^{i} \frac{1}{r+n} \leq \frac{r}{r+n} + \frac{n}{r+n} = 1$$

Therefore if $\max_{\tau_i \in \tau} \left\{ \frac{C_i}{T_i} \right\} \leq \frac{1}{r+n}$, since the condition in Theorem 1 is satisfied, the task set is schedulable.

When it is hard to limit each task's utilization as given in Theorem 2, we can use the utilization bound calculated using the maximum task utilization as in the following corollary derived from Theorem 4.

Corollary 1. *A set of n periodic tasks with $D_i = T_i$ is schedulable by the non-preemptive Rate Monotonic scheduler if*

$$\sum_{i=1}^{n} \frac{C_i}{T_i} \leq 1 - \alpha r \tag{5}$$

where $T_n = rT_1$ and $\alpha = \max_{\tau_i \in \tau} \left\{ \frac{C_i}{T_i} \right\}$.

Proof. For a given task τ_i, let us suppose $C^i_{\max} = C_k$. By dividing the lefthand side of equation (1) by T_i, we have

$$\frac{C^i_{\max}}{T_i} + \frac{C_i}{T_i} + \frac{1}{T_i}\sum_{j=1}^{i-1}\left\lfloor \frac{T_i}{T_j} \right\rfloor C_j \leq \frac{C_k}{T_k}\cdot\frac{T_k}{T_i} + \frac{C_i}{T_i} + \sum_{j=1}^{i-1}\frac{C_j}{T_j} \leq \alpha r + \sum_{j=1}^{n}\frac{C_j}{T_j} \leq 1$$

which proves the corollary.

If we limit each task's utilization by the condition given in Theorem 4, by combining Theorem 4 and Corollary 1 (*i.e.*, letting $\alpha = \frac{1}{r+n}$) we can achieve utilization bound as large as $\frac{n}{r+n}$.

As shown in Theorem 4, since tasks are non-preemptive, dividing tasks into several tasks can enhance the utilization bound. In this sense, Theorem 4 gives us a guide for designing real-time systems with non-preemptive fixed priority scheduling. To achieve higher system utilization, tasks should be divided so that we have large number of tasks, and the ratio of the maximum and the minimum period should be kept as small as possible. For example, when $r = 2$, the utilization bound given in Theorem 3 is only 50%. But by dividing some tasks to limit each task's utilization by 10% and to have 8 tasks, the possible utilization bound can be achieve is $\frac{8}{(2+8)} = 80\%$. Note that dividing a task does not help enhancing the utilization bound for preemptive scheduling.

Fig. 2 compares the utilization bound of non-preemptive and preemptive fixed priority scheduling when $r = 2$. The utilization bound for a given number of tasks is calculated using the result of [6], that is $(n-1)(2^{1/(n-1)} - 1) + 2/r - 1$. As the number of tasks increases, the utilization bound of preemptive scheduling approaches to the Liu and Layland's bound, $\ln 2$, while that of non-preemptive scheduling approaches 100%.

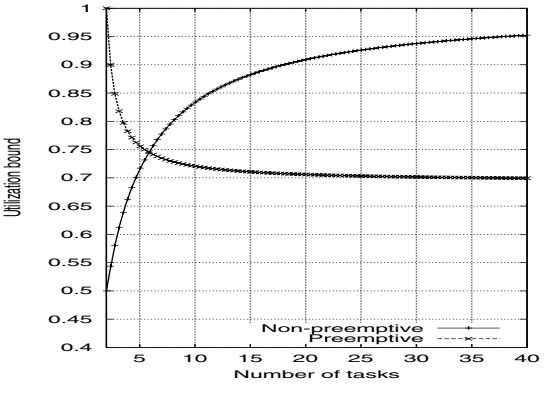


Fig. 2. Utilization bound for preemptive and non-preemptive scheduling for different number of tasks

5 Conclusions

In this paper, we have shown Rate Monotonic priority assignment is optimal for non-preemptive scheduling as in preemptive scheduling and schedulability for periodic tasks can be determined in polynomial time. Utilization bound for Rate Monotonic scheduling is presented using the ratio of the maximum and the minimum period. To overcome the small utilization bound, we also presented a method to enhance the utilization bound by limiting the each task's utilization. It was shown that by keeping the period ratio small and dividing a task into tasks with the same period and smaller execution time, we can achieve relatively higher utilization. The optimal fixed-priority scheduling policy for non-preemptive tasks with relative deadline smaller than period, and schedulability of concrete periodic tasks will be our future work.

References

1. N.C. Audsley. Optimal priority assignment and feasibility of static priority tasks with arbitrary start times. Technical report, Department of Computer Science, University of York, 1991.
2. N.C. Audsley, A. Burns, M. Richardson, and A. Wellings. Hard real-time scheduling: The deadline monotonic approach. In *Proceedings of IEEE Workshop on Real-Time Operating Systems and Software*, pages 133–137, May 1991.
3. T.P. Baker. A stack-based allocation policy for real-time processes. In *Proceedings of IEEE Real-Time Systems Symposium*, pages 191–200, December 1990.
4. L. George, N. Riviere, and M. Spuri. Preemptive and non-preemptive real-time uniprocessor scheduling. Technical report, INRIA, 1996.
5. K. Jeffay, D.F. Stanat, and C.U. Martel. On non-preemptive scheduling of periodic and sporadic tasks. In *Proceedings of IEEE Real-Time Systems Symposium*, pages 129–139, December 1991.
6. S. Lauzac and D. Mossé R. Melhem. An improved rate-monotonic admission control and its applications. *IEEE Transactions on Computers*, 52(3):337–350, 2003.
7. J.P. Lehoczky. Fixed priority scheduling of periodic task sets with arbitrary deadlines. In *Proceedings of IEEE Real-Time Systems Symposium*, pages 201–209, 1990.
8. C. Liu and J. Layland. Scheduling algorithms for multiprogramming in a hard real-time environment. *Journal of ACM*, 20(1):46–61, 1973.
9. W.-C. Lu, H.-W. Wei, and K.-J. Lin. Rate monotonic schedulability conditions using relative period ratios. In *Proceedings of IEEE International Conference on Embedded and Real-Time Computing Systems and Applications*, pages 3–9, August 2006.
10. T. M. Parks and E. A. Lee. Non-preemptive real-time scheduling of dataflow systems. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 3225–3238, May 1995.
11. L. Sha, R. Rajkumar, and J. Lehoczky. Priority inheritance protocols: An approach to real-time synchronization. *IEEE Transactions on Computers*, 39(9):1175–1185, 1990.
12. M. Spuri. Analysis of deadline scheduled real-time systems. Technical report, INRIA, 1996.

A New Mobile Payment Method for Embedded Systems Using Light Signal*

Hoyoung Hwang[1], Moonhaeng Huh[2], Siwoo Byun[2], and Sungsoo Lim[3]

[1] Department of Multimedia Engineering, Hansung University, Korea
hyhwang@hansung.ac.kr
[2] Department of Digital Media Engineering, Anyang University, Korea
{moonh,swbyun}@anyang.ac.kr
[3] Department of Computer Science, Kookmin University, Korea
sslim@kookmin.ac.kr

Abstract. This paper presents a new type of digital payment method and algorithm for embedded systems including mobile phone and other handheld devices. In the conventional mobile payment system, a payment signal transmitting device and a wireless internet connection device are separately provided to mobile systems. Therefore, the production cost of the mobile devices increases, and the complexity and the size of the mobile devices grow. In order to solve the problems, a new digital payment method using light signal containing payment-related information is suggested for mobile phone and embedded devices. The new method provides an economic and efficient solution to mobile payment system by updating embedded software without need of additional hardware modules to mobile devices.

1 Introductions

The trend of digital and mobile convergence in IT technology brings multi-functionality to mobile embedded devices including phones and PDAs. As a result, various functions such as digital camera, camcoder, mp3 player, which were produced and purchased as independent devices before are now integrated in a single phone or embedded device. One of those functionalities is digital cash or payment system.

In recent years, several mobile payment methods or systems have been developed in order to avoid the inconvenience of carrying cash or credit card [1]-[7]. Those payment systems are mainly used for mobile phone devices and mostly for small-valued payment. However, in conventional mobile payment systems, a payment signal transmitting device and a wireless internet connection device are separately provided to the mobile phone or embedded devices. Therefore, the production cost of the mobile phone or devices increases, and the complexity and size of the mobile devices grow. In addition, there is another critical problem in that the users who want to use the mobile payment system may need to purchase a new phone or devices with additional hardware functions.

In order to solve the aforementioned problems, a new mobile payment method is proposed. The method is implemented by updating embedded software of a phone

* This research was financially supported by Hansung University in the year of 2007.

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 889–896, 2007.
© Springer-Verlag Berlin Heidelberg 2007

without the need of additional hardware modules to the existing mobile phone or devices. In the proposed method, a *virtual mobile card* containing payment-related information is loaded to the virtual machine in the mobile phone or embedded system OS. The virtual mobile card can comprise various types of payment-related information such as credit card number, membership card number, discount coupon image, online ticket, etc., which are downloaded from a control server and stored in the mobile phone or devices. The virtual mobile card information is transmitted by image by blinking the backlight of the LCD/LED panel of the mobile devices, and the transmitted information is decoded by the photo receiver which is designed for the proposed mobile payment system. The data transmission function is implemented by embedded software that controls the lightning device of a phone or handheld systems, which can be applied to the most mobile phone already in use.

In the following Section 2, the characteristics of existing mobile payment systems are presented and compared with the proposed method. Section 3 describes the proposed mobile payment system model and its detailed subsystems; the mobile phone, the photo receiver, and the control server. Section 4 presents the payment procedure of the proposed method, and Section 5 summarizes the paper.

2 Mobile Phone Payment Systems

In Korea, it is believed that normally 4 or 5 credit cards are owned and carried per adult person in addition to transportation card, debit card, membership card, ID card, and etc. [8] There is a problem in that, when the cards are lost or stolen, the owner may suffer a great deal of damages. In order to solve the inconvenience of carrying various cards, several approaches have been tried to put those card information into a phone or other embedded devices. The approaches include barcode-based solution, and IR/RF-based solution as presented in Table 1.

Some telecom companies adapted barcode type membership card into mobile phones and used it as mobile payment and membership authentication[9]. This method is easy to implement and widely applicable to membership management applications. The weak point of this method is that the information being contained by a barcode is very limited. Moreover, if conventional laser type barcode reader devices are used, many transmission errors occur because of the reflection problem on the surface of the display device of a phone. Therefore, CCD type reader device is needed which increase the cost of system. In addition, barcode type information can be easily copied thus incurs security problem.

More widely adapted mobile payment method is using IrDA (Infrared Data Association) or RFID (Radio Frequency Identification) technology to integrate payment related information into a phone or embedded devices[10]. This method shows high ability of processing large volume of data and multimedia applications, and provides high level of security management functions. However, the shortcoming of this method is that it requires additional hardware modules. Therefore, users should have specialized phones or devices with the additional hardware module instead of existing devices already in use, which makes widespread use of this method difficult.

Table 1. Characteristics of mobile payment methods

	o o o	B o	
i i n	. i ing ig n i	. i g .	. i i
i ing i	ig . i ing n i	ig . i ing n i	. n i
S i y	. i y n	. i y i	ig . ig i y n g
i	. i n	ig . i y n i	ig . i in
	. i n n	ig . i n n n n	. y n
in S i i y	ig . n i i i n n	. i i n i n in i n	. i i n i n . , g in i i ig i y.

A new mobile payment method is proposed to solve the problems of existing payment methods. This method can be applied to the most mobile phones or devices already in use by updating embedded software without additional hardware modules.

3 Light Signal Based Mobile Payment System

The proposed mobile payment method transmits payment related information using the backlight of devices which is adopted in most mobile phones and embedded devices, thus eliminates the need of supplying new customized devices. The payment information using backlight signal is called virtual mobile card. The virtual mobile card based payment method has the following characteristics.

- The software application program is downloaded to the virtual machine of a phone or embedded devices.
- The optical source such as LCS backlight or keypad backlight of existing phones or devices is used for information transmission including software download and payment.
- Various types of security management solutions can be implemented by software.
- Large volume of card information data can be adopted as long as the embedded memory size allows.
- The card information can be downloaded by wireless connection immediately after the subscription.

- CRM (Customer Relationship Management) and other applications can be easily applicable by coordination with phone number information.
- The cost of optical reader device is relatively cheap compared with CCD type readers or IR/RF type readers.

The virtual mobile card is downloaded from control server. The control server first sends the Callback URL which contains WAP address in the SMS page, and the users of phones or embedded devices access the address and download the application program and information. The mobile payment procedure is similar to the mobile game play procedure. The display device of a phone is attached to the photo receiver and the request and reply data is transmitted between the PCs in the branch office and the server in the main office as shown in Fig. 1.

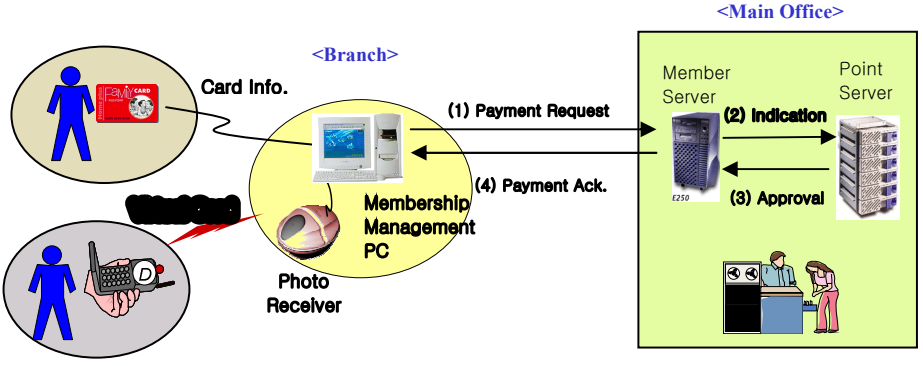


Fig. 1. Use of virtual mobile card for mobile payment

Our mobile payment system using a light signal consists of three parts: a mobile phone for generating a predetermined light signal; a photo receiver for converting the received light signal to an electrical signal, encrypting the electrical signal, and generating the encrypted electrical signal as an output signal; and the control server for authenticating payment by using the output signal of the photo receiver.

Fig. 2 is a block diagram showing a mobile phone payment system. The system comprises a mobile phone 111, a photo receiver 121, and a control server 131.

In general, the mobile phone 111 has a display unit such as an LCD, an EL, and an organic EL illuminated with a backlight. Some types of the mobile phone 111 may have a keypad illuminated with a keypad-dedicated backlight such as a blue LED. In the mobile phone payment system, a light signal is generated in a form of pulses by turning on and off the backlight of the mobile phone based on payment-related information. In other words, the light signal contains the payment-related information. Therefore, the user of the mobile phone can make payments by using his or her mobile phone.

In order to use the mobile phone payment system 101, the user accesses the control server 131 to subscribe to the mobile phone payment service by transmitting his or her personal information to the control server 131. The personal information includes identification (ID) information, a credit card number, an E-mail address, E-money

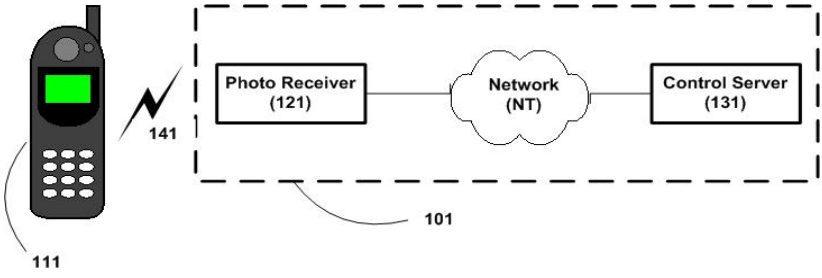


Fig. 2. Block diagram showing a mobile payment system

information, and a phone number of the user. The control server 131 stores the personal information in a data storage unit. The user downloads a program used for the mobile phone payment service from the control server 131 and installs the program in the mobile phone 111.

The user can pay a charge by using the mobile phone 111. The charge includes, for example, a charge for a game machine and a roller coaster, a charge for a gambling machine, an admission fee of a theater, a toll, a public vehicle fare, and a parking fee. In addition, the mobile phone 111 may be used to pay a price of gasoline, a vending-machine item. In addition, mobile phone 111 may be also used to transfer E-money in ATM.

A photo receiver 121 decodes a light signal from a mobile phone, converting the light signal to an electrical signal, and transmitting the electrical signal to a control server. The photo receiver comprising: a photo sensor for receiving a light signal from a mobile phone, and converting the light signal into an electrical signal; a data error determination unit for receiving the electrical signal of the photo sensor and determining an error of the electrical signal; a control unit for transmitting the electrical signal of the photo sensor to the data error determination unit, generating a control signal based on the result of the error determination by data error determination unit, and generating the electrical signal if the electrical signal has no error; a display unit for displaying presence or absence of the error of the light signal in response to the control signal; and a communication unit for transmitting the electrical signal from the control unit to the control server.

The photo receiver may further comprise an encryption unit for encrypting the electrical signal transmitted from the control unit, and a communication unit for transmitting the encrypted signal by the encryption unit to the control server.

The display unit may comprise at least one of a third lighting device which is turned on, if a signal indicating that the payment-related information transmitted from the mobile phone is correct is transmitted from the control server, and a forth lighting device is turned on, if not, and the control unit may receive a predetermined signal from the control server via the communication unit and control the operation of the display unit in accordance with the received signal.

Fig. 3 is a waveform diagram showing pulses of a light signal 141 in Fig. 2, generated by a backlight of a mobile phone 111 used for a mobile payment system. The light signal 141 is partitioned into a starting frame t1, an information frame t2, and an

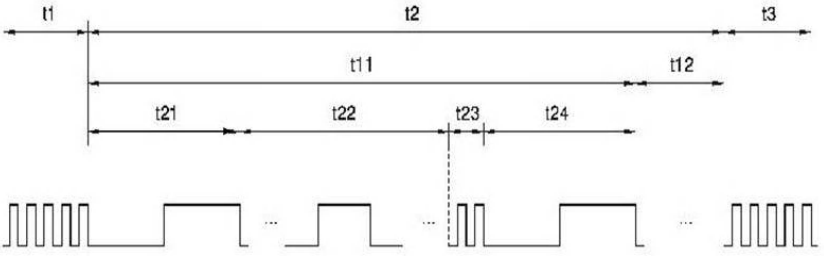


Fig. 3. Waveform diagram showing pulses of a signal generated by a backlight

ending frame t3. The starting frame t1 comprises a plurality of short pulses to indicate the starting point of the light signal 141. The information frame t2 comprises a plurality of pulses generated based on the payment-related information. The ending frame t3 comprises a plurality of pulses to indicate the ending point of the light signal 141. The photo receiver may further comprise an encryption unit for encrypting the electrical signal transmitted from the control unit, and a communication unit for transmitting the encrypted signal by the encryption unit to the control server. Fig. 4 shows the block diagram and implemented prototype of photo receiver.

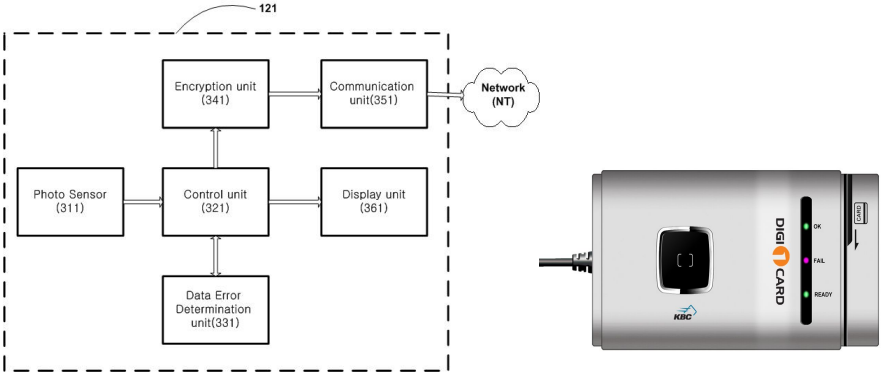


Fig. 4. Block diagram and prototype of a photo receiver

Fig. 5 is a flowchart showing a mobile phone payment method.

The mobile payment method is comprised of the following steps: (a) downloading the payment-related information of the user from the control server to the mobile phone and storing the payment-related information in the mobile phone; (b) converting the payment-related information into a binary code data in response to the user's request; (c) generating a light signal having a series of pulses by turning on and off a backlight of the mobile phone based on the binary code data and transmitting the light signal to the photo receiver; (d) receiving the light signal in the photo receiver, converting the received light signal into an electrical signal and transmitting the electrical signal from the photo receiver to the control server; (e) receiving the electrical signal,

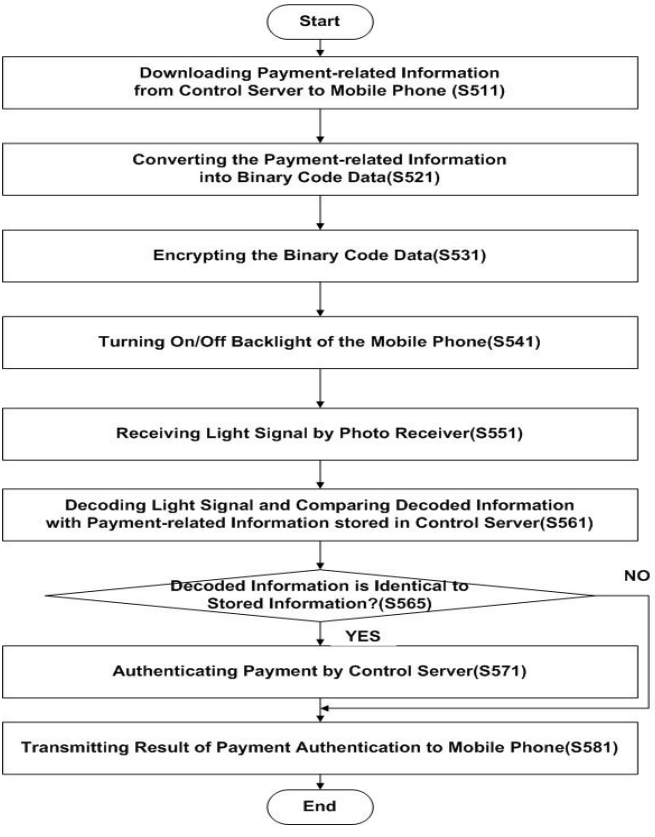


Fig. 5. Flowchart showing a mobile payment procedure

comparing payment-related information contained in the electrical signal with the payment-related information of the user stored in the control server and determining whether or not the transmitted payment-related information is correct; (f) if the transmitted payment-related information is correct, authenticating the user's payment; and (g) transmitting the result of the determination of the step (e) from the control server to at least one of the mobile phone and the photo receiver.

The step (b) may further comprise a step of encrypting the binary code data, and the step (e) may further comprise a step of decoding the received electrical signal.

4 Conclusion

A mobile payment system was proposed for mobile phone or embedded devices. The system consists of mobile device, photo receiver, and control server. The payment related information is downloaded from the control server to the user device. The new method provides an economic and efficient solution to mobile payment system using light signal. It is possible to make payments by using a mobile payment system being implemented by updating embedded software of the mobile phone or devices without

additional hardware modules. Therefore, it is possible to reduce production costs of mobile phones and mobile payment systems. The measurement and comparison of performance such as recognition speed, error rate, and response time with other systems will be our future work, as well as in-depth study for security and privacy problems incurred by mobile payment systems. This work is subject to patent pending [11].

References

1. Zheng, X., Chen, D.: Study of Mobile Payments System, Proceedings of International Conference on E-Commerce, (2003) 24-27
2. Mallat, N., Tuunainen, V.K.: Merchant Adoption of Mobile Payment Systems, Proceedings of International Conference on Mobile Business, (2005) 347-353
3. Pousttchi, K., Zenker, M.: Current Mobile Payment Procedures on the German Market from the View of Customer Requirements, Proceedings of Workshop on Database and Expert Systems and Applications, (2003) 870-874
4. Delic, N., Vukasinovic, A.: Mobile Payment Solution – Symbiosis between Banks, Application Service Providers and Mobile Network Operators, Proceedings of International conference on Onformation Technology: New Generations, (2006) 346-350
5. Lee, O.: Sound-based Mobile Payment System, Proceedings of International Conference on Web Services, (2004) 820-821
6. Gao, J., Edunuru, K., Cai, J., Shim, S.: A Peer-to-Peer Wireless Payment System, Proceedings of International Conference on Mobile Commerce and Services, (2005) 102-111
7. Antovski, L., Gusev, M.: M-payments, Proceedings of International Conference on Information Technology Interfaces, (2003) 95-100
8. Korea Cyber Police Agency, http://www.police.go.kr, (2005)
9. http://www.ktfmembers.com, http://www.pheonixpark.co.kr
10. http://www.monetacard.co.kr, http://www.irda.org
11. Mobile Phone Payment Method and System, US patent pending, No. 10/893836

Bounding Demand Paging Costs in Fixed Priority Real-Time Systems

Young-Ho Lee[1], Hoyoung Hwang[2], Kanghee Kim[3], and Sung-Soo Lim[1]

[1] School of Computer Engineering, Kookmin University,
Jungrung 3-dong, Seongbuk-gu,
Seoul, Korea 136-702
{buriburi,sslim}@kookmin.ac.kr
[2] Department of Multimdia Engineering, Hansung University,
Samsun-Dong 3Ga 389, Sungbook-Gu,
Seoul, Korea 136-792
hyhwang@hansung.ac.kr
[3] Mobile Communication Division,
Telecommunication Network Business,
Samsung Electronics Co., LTD.
Kang.hee.kim@samsung.com

Abstract. In real-time systems, demand paging has not been used since worst case performance analysis technique for demand paging is not available. In this paper, we propose a technique to obtain the worst case bound on the demand paging costs and devise a worst case response time analysis framework considering the demand paging costs. We divide the technique into two different methods, DP-Pessimistic and DP-Accurate, depending on the accuracy and the complexity of analysis. DP-Accurate considers the page reuses among task instances and thus gives more accurate analysis results. We evaluate the accuracy of the proposed techniques comparing the analysis results from DP-Accurate and DP-Pessimistic.

Keywords: WCET analysis, WCRT analysis, worst case response time, demand paging, flash memory.

1 Introduction

Demand paging has long been the key feature of virtual memory systems as full-featured operating systems are increasingly used in various computing devices. Until recently, the demand paging technique has been mainly used in desktop or server systems rather than real-time embedded systems. This is due to the fact that the applications in real-time embedded systems have not been such heavy-weighted ones that the demand paging do not have to be adapted in the systems. Therefore, rather than using demand paging, the whole program image is copied into DRAM before execution (we call it *shadowing*). The shadowing has not been considered as preferable because of its DRAM consumption.

With the advent of multi-functional and high performance real-time embedded systems, there is no choice but to use demand paging in the operating systems for

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 897–904, 2007.
© Springer-Verlag Berlin Heidelberg 2007

real-time embedded systems. One major factor that makes adapting the demand paging in real-time embedded systems difficult is the unpredictable performance of demand paging. The unpredictability stems from the history-sensitive feature of demand paging that the page fault delay would be highly dependent upon the past memory access patterns. Though such difficulty exists, we would need any framework to analyze the demand paging performance if the demand paging is to be used in real-time system design.

In this paper, we devise a worst case response time (WCRT) analysis technique considering the demand paging cost in fixed priority scheduling real-time systems. The basic idea is that the conventional WCRT analysis framework is augmented with the calculation of additional cost caused by demand paging. One simple solution for the analysis would be deriving the worst case page fault cost for a whole task and add the cost to the original response time of the task. Such naïve solution would yield too loose bound on the WCRT considering demand paging cost. Therefore, we enhance the accuracy of the analysis by considering the possibility of page reuses among the task instances. We propose graph theoretical formulation for the analysis considering page reuses among the task instances. We show that the proposed enhanced technique produces more accurate bounds on the worst case demand paging cost than the simple and naïve technique up to 57% by simulation. The rest of this paper is organized as follows. Section 2 describes our WCRT analysis method. Section 3 shows the simulation results for our WCRT analysis method with a simple simulator and virtual task sets. Finally, we conclude in section 4.

2 Bounding Demand Paging Cost

In this section, we provide how we extend the original WCRT analysis to consider the additional costs of demand paging. Our technique is based on the conventional WCRT analysis frameworks[1][2][3][4][5][6] and an extension of the original framework to consider cache-related preemption costs[5]. In [5], the WCRT of task τ_i is defined as follows:

$$R_i = C_i + \sum_{j \in hp(i)} \lceil \frac{R_i}{T_j} \rceil C_j + PC_i(R_i). \quad (1)$$

where $PC_i(R_i)$ is the total preemption costs of task τ_i. In this paper, we only focus on how the page fault handling cost could be considered in WCRT analysis leaving the consideration of remaining factors affecting the accuracy of the analysis for future work.

For simplicity of our analysis, we make the following assumptions: First, we assume the page fault handling cost for each page is identical in all cases. Second, we assume that our target system has infinitive memory and thus our target system does not suffer from memory limitation causing page replacement. The second assumption is to exclude the worst case page replacement analysis issue from the scope of the paper while the replacement policy should ultimately be considered for completeness.

The key issue of the WCRT analysis for demand paging is how to obtain the demand paging cost for each task and how to take into account it in computing the worst case response time of a task. In order to obtain the demand paging cost for each task,

we introduce two analysis steps: demand paging cost analysis and page reuses aspect analysis. In the demand paging cost analysis step, the demand paging costs for every execution path is calculated. For an execution path-wise analysis, we use control-flow based approach. After the analysis, we obtain the demand paging cost and pages to be loaded for every execution path of the task. In the page reuses aspect analysis step, the page reuses among task instances are analyzed and we obtain the WCETs of each task instance considering the demand paging cost and the page reuses. We use graph-theoretical formulation to derive the worst case bound on the page fault cost of a series of task instances.

2.1 Demand Paging Cost Analysis

The objective of the demand paging cost analysis is to obtain the demand paging cost for every execution path. Because the demand paging cost is highly dependent on the execution paths of the task, we use control flow-based approach to obtain the demand paging cost depending on the execution paths. The first step of building the control flow graph for our analysis is to make the correspondence between each basic block and each memory page. All the basic block nodes in the control flow graph are mapped to the corresponding memory pages. Since the sizes of basic blocks are various, the page boundaries do not always match basic block boundaries. Figure 1 shows an example of basic block and page mapping for a sample control flow graph. The control flow graph has four different execution paths and each basic block is mapped to one of the six different pages.

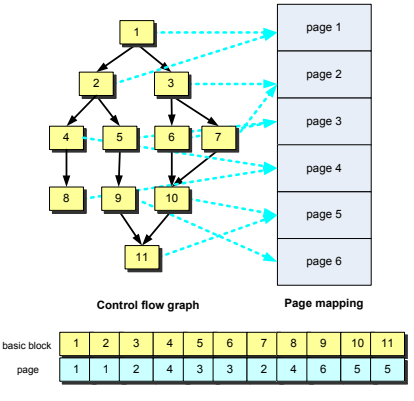


Fig. 1. Basic block and page mapping for a control flow graph

We use the following notations for formulation of our analysis: $C_{i,j}$ denotes the WCET of *j*-th execution path of task τ_i. Note that $C_{i,j}$ is the pure WCET which does not consider the demand paging cost. $P_{i,j}$ denotes the set of pages for *i*-th execution path of task τ_i, and $n(P_{i,j})$ denotes the number of pages in $P_{i,j}$. π denotes the worst case page fault handling cost. Due to the assumption in the previous section, π is the same in all cases. The demand paging cost for each execution path in task τ_i is:

$$demand\ paging\ cost_{i,j} = \pi \times n(P_{i,j}). \tag{2}$$

The WCET including demand paging cost for each execution path could be calculated as follows:

$$C_i'' = max \left\{ \begin{array}{c} C_{i,0} + demand\ paging\ cost_{i,0} \\ C_{i,1} + demand\ paging\ cost_{i,1} \\ ... \\ C_{i,n-1} + demand\ paging\ cost_{i,n-1} \end{array} \right\}. \tag{3}$$

2.2 Page Reuses Aspect Analysis

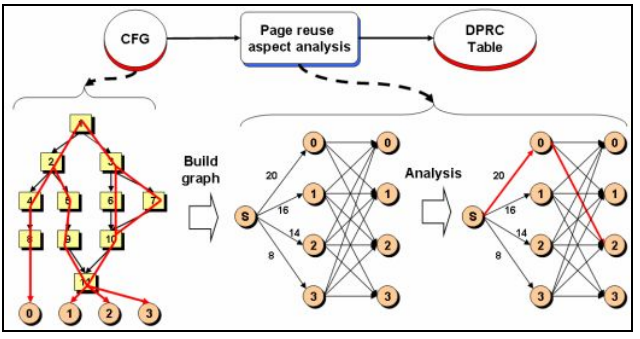
The demand paging cost could be reduced among task instances since a task instance reuse the pages already loaded by preceding task instances. The objective of the page reuses aspect analysis is to obtain a safe bound on the demand paging cost considering the page reuses among task instances. For the analysis, we use graph-theoretical formulation that the execution time of each task instance (*i.e.*, execution path) considering the page reload cost is represented. By traversing the graph from the start node until the last task instance to obtain the maximum distance, we obtain the WCRT of each task instance and the worst case scenario of task schedule. Figure 2(a) shows the overall page reuses aspect analysis in brief.

For the page reuses aspect analysis, we consider a directed graph $G = (N, E)$: *N* is the set of nodes that represent instances of each task and *E* is the set of edges among the nodes whose weights are the execution times between two task instances. The number of elements of *N* is the number of execution paths multiplied by the number of instances of each task. The weight of each edge, $w_{p,q,r}$, in *E* represents the execution time of execution path *q* of *r*'th task instance of when the immediately preceding path is *p*. For our formulation, we define $c_{p,q}$ as the WCET considering the demand paging cost of the execution path *q* when the immediately preceding execution path was *p*. In addition, we define $\theta_{p,q,r}$ as the page reloads cost of the execution path *q* for *r*'th task instance when the immediately preceding execution path was *p*. $w_{p,q,r}$ can be defined as $c_{p,q} - \theta_{p,q,r}$.

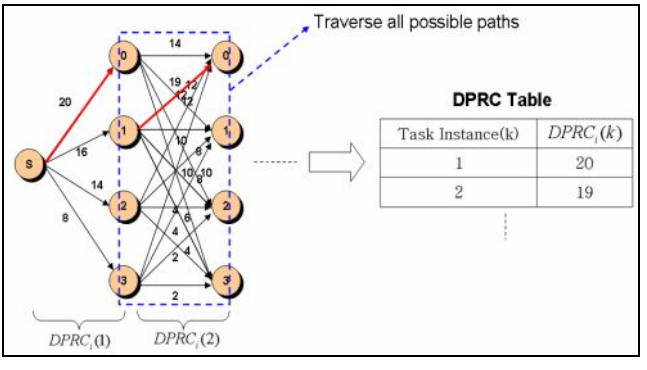
The page reuses aspect analysis is to determine the maximum sum of the WCETs considering the demand paging cost and page reuses for each task instances. The following formulation shows a problem definition.

$$\text{Maximize } \sum_{r=1}^{n} (c_{p,q} - \theta_{p,q,r}). \tag{4}$$

where $(p, q) \in N$. Since the pure WCET of the task could not vary, $c_{p,q}$ is always the same in all cases. But $\theta_{p,q,r}$ ranges from 0 to the maximum page fault delay assuming that all the pages of execution path q cause faults. This is because the page reload cost varies according to the pages loaded in the previous task instances.



(a)



(b)

Fig. 2. The page reuses aspect analysis overview

The key point in the formulation is that in order to obtain the safe bound on the WCRT considering page fault cost, we choose the maximum weight for each pair of task instances. For describing the problem more formally, we introduce the following new notations, *DPRC* (Demand Paging Related Cost), and *PLOAD*.

$DPRC_i(j)$: the WCET of task instance $\tau_{i,j}$ considering demand paging costs and the page reuses

$PLOAD_i(j)$: the set of all the combinations of pages loaded until task instance $\tau_{i,j}$.

For example, the WCETs considering demand paging costs for task instances $\tau_{0,0}$, $\tau_{0,1}$ and $\tau_{0,2}$ are $DPRC_0(1)$, $DPRC_0(2)$ and $DPRC_0(3)$, respectively. In the above graph, *DPRC* is the same as $c_{p,q} - \theta_{p,q,r}$ and *PLOAD* determines $\theta_{p,q,r}$.

Figure 2(b) shows how to calculate each DPRC values of the task. In the figure, the first task instance has the maximum WCET when the task executes an execution path 0 and in this case, $DPRC_i(1)$ is 20. In order to calculate $DPRC_i(2)$, all the possibilities of page loads for the task instance $\tau_{i,1}$ should be re-considered. As a result,

$DPRC_i(2)$ is 19 when the task executes the execution path 0. This would produce safe bound on the WCET of the second task instance.

2.3 Calculation of the WCRT

In this paper, we propose two different WCRT analysis methods: DP-Pessimistic and DP-Accurate, depending on the accuracy and the complexity of the analysis. In DP-Pessimistic method, the WCRT R_i of task τ_i is calculated only by using the demand paging cost analysis and could be formulated as follows:

$$R_i = C_i^{''} + \sum_{j \in hp(i)} \lceil \frac{R_i}{T_j} \rceil C_j^{''}. \tag{5}$$

where $C_i^{''}$ is the new WCET including demand paging cost recalculated from C_i. In DP-Accurate method, the WCRT R_i of task τ_i is calculated using the demand paging cost analysis and page reuses aspect analysis and could be obtained as follows:

$$R_i = DPRC_i(1) + \sum_{j \in hp(i)} \sum_{k=1}^{\lceil \frac{R_i}{T_j} \rceil} DPRC_j(k). \tag{6}$$

where $DPRC_i(1)$ and $DPRC_j(k)$ are the WCETs of each task instance considering the demand paging cost and page reuses.

3 Evaluations

To evaluate the analysis accuracy of DP-Pessimistic and DP-Accurate, we simulate the task executions with demand paging using manually composed sample task set with various task parameters. Our simulator performs RM (rate monotonic) scheduling simulation with three different demand paging cost analysis: shadowing, DP-Pessimistic, and DP-Accurate. For our evaluation experiments, we carefully select the task parameters to show the effect of DP-accurate compared to DP-pessimistic as much as possible.

Table 1 shows the sample task set and parameters. Each task set contains same number of tasks with the same pure WCETs when demand paging costs are not considered. The only differences among the task sets are the page access behaviors of the tasks to show the accuracy differences of DP-Pessimistic and DP-Accurate methods. Task set 2 and Task set 3 have execution paths whose demand paging costs are significantly larger than the costs of other paths.

Figure 3 shows the WCRT analysis results for the lowest priority task of each task set. The execution times are given at *ms* (milliseconds) and the ratio of DP-Accurate over DP-Pessimistic is given. The differences between DP-Accurate analysis and DP-Pessimistic are significant for task sets 2 and 3 since the task sets have execution paths whose page access costs are significantly larger than other paths. This reveals that DP-Pessimistic analysis produces too loose bounds on the WCRTs compared to DP-Accurate analysis. Such loose bound affects the decision on the schedulability of the task sets. For example, with DP-Accurate, all the task sets satisfy the deadlines,

Table 1. Task set parameters

Task set	T_i	# of paths	$C_{i,j}$	$P_{i,j}$
1	5	1	1	{45}
	15	1	2	{46}
	60	5	4	{1, 2}
			3	{8, 9}
			5	{15, 16}
			4	{24, 25, 26, 27, 28, 29}
			5	{33, 34, 38, 39}
	240	1	60	{47}
2	5	1	1	{45}
	15	1	2	{46}
	60	5	4	{1, 2}
			3	{8, 9}
			5	{15, 16}
			4	{24, 25, 26, 27, 28, 29, 30, 31}
			5	{38, 39}
	240	1	60	{47}
3	5	1	1	{45}
	15	1	2	{46}
	60	5	4	{1, 2}
			3	{8, 9}
			5	{15, 16}
			4	{24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34}
			5	{38, 39}
	240	1	60	{47}
4	5	1	1	{45}
	15	1	2	{46}
	60	5	4	{1, 2, 3}
			3	{8, 9, 10}
			5	{15, 16}
			4	{24, 25, 26, 27}
			5	{38, 39, 40}
	240	1	60	{47}

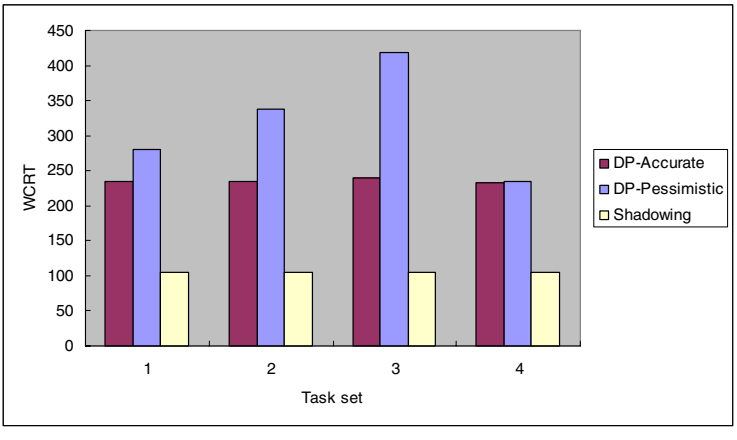


Fig. 3. Evaluation results

but with DP-Pessimistic, the tasks for Task sets 1, 2, and 3 do not satisfy the deadlines from the analysis. Task set 4 shows similar worst case response times between DP-Accurate and DP-Pessimistic since the page access costs are similar for all execution paths.

4 Conclusion

In this paper, we have proposed a WCRT analysis method considering demand paging costs. Depending on the accuracy and the complexity of the analysis, two different analysis methods, DP-Pessimistic and DP-Accurate, are devised. DP-Accurate considers the possibility of page reuses among different task instances while DP-Pessimistic does not consider the page reuses.

Evaluation results show that DP-Accurate method gives more accurate analysis results than DP-Pessimistic method especially for the tasks consisting of task instances whose pages are reused by succeeding task instances. The current analysis method is the basis for future extensions that aim at more practical modeling of demand-paging for real-time systems.

Acknowledgments. This work was supported in part by the MIC & IITA under IT Leading R&D Support Project 2006, in part by No. 379 from the Basic Research Program of KOSEF, in part by the MIC, Korea under the ITRC (Information Technology Research Center) support program supervised by the IITA (IITA-2006-C1090-0603-0045), and in part by Kookmin University under Research Center Incubation Program (UICRC).

References

1. K. Tindell.: Adding Time-Offsets to Schedulability Analysis. Technical Report YCS 221, Dept. of Computer Science, University of York, England (1994)
2. Palencia, J. C. and Gonzlez Harbour, M.: Schedulability Analysis for Tasks with Static and Dynamic Offsets. In Proceedings of the IEEE Real-Time Systems Symposium (1998)
3. N.Audsley et al.: Applying new scheduling theory to static priority preemptive scheduling. Software Engineering Journal. (1993) 284-292
4. J. V. Busquets-Mataix, J. J. Serrano-Martin, R. Ors, P. Gil, and A. Wellings.: Adding Instruction Cache Effect to Schedulability Analysis of Preemptive Real-Time Systems. Proceedings of the 2nd Real-Time Technology and Applications Symposium (1996)
5. C. Lee, J. Hahn, Y. Seo, S. Min, R. Ha, S. Hong, C. Park, M. Lee, and C. Kim.: Analysis of Cache-related Preemption Delay in Fixed-Priority Preemptive Scheduling. IEEE Transactions on Software Engineering (1996) 264-274
6. K. Tindell, A. Burns, and A. Wellings.: An Extendible Approach for Analysing Fixed-Priority Hard Real-Time Tasks. Journal of Real-Time Systems, 6(2). (1994) 133-151
7. Scott F. Kaplan, Lyle A. McGeoch, Megan F. Cole.: Adaptive caching for demand prepaging. Proceedings of the 3rd international symposium on Memory management. Berlin, Germany (2002)

OTL: On-Demand Thread Stack Allocation Scheme for Real-Time Sensor Operating Systems*

Sangho Yi[1], Seungwoo Lee[1], Yookun Cho[1], and Jiman Hong[2,**]

[1] System Software Research Laboratory
School of Computer Science and Engineering, Seoul National University
Tel.: +82-2-872-7431, Fax.: +82-2-875-7726
{shyi,solee,cho}@ssrnet.snu.ac.kr
[2] School of Computing, Soongsil University
jiman@ssu.ac.kr

Abstract. In wireless sensor networks, each sensor node has severe resource constraints in terms of energy, computing device, and memory space. Especially, the memory space of the platform hardware is much smaller than that of the other computing systems. In this paper, we propose a OTL, which is an on-demand thread stack allocation scheme for MMU-less real-time sensor operating systems. The OTL enables to adaptively adjust the stack size by allocating stack frame based on the amount of each function's stack usage. The amount of the function's stack usage is checked at compile-time, and the adaptive adjustment of the stack occurs at run-time. Our experimental results show that the OTL significantly minimizes the spatial overhead of the threads' stacks with tolerable time overhead compared with fixed stack allocation mechanism of the existing sensor operating systems.

1 Introduction

Many kinds of existing operating systems can be cartegorized into two classes according to how they operate and design kernel architecture. Historically, name of the two rough classes are defined as follows: event-driven(or message-oriented), and multi-threaded(or procedure-oriented). For the past several decades, many discussions have been made on these two canonical models of the operating systems[1,2,3,4,5,6]. In 1978, Lauer and Needham attempted to finish the discussion on the two models by showing duality between event-driven and multi-threaded operating systems[2]. As a result, they concluded that the two canonical models are "duals" in terms of characteristics, program operation, and system performance.

The above conclusions are correct theoretically. In the real implementations, however, the two models show significant difference on performance and efficiency of systems and applications. There are pros and cons of the event-driven

* This Research was supported by the Soongsil University Research Fund.
** Corresponding author.

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 905–912, 2007.
© Springer-Verlag Berlin Heidelberg 2007

and the multi-threaded system models on designing a sensor operating system. The event-driven system design gives us smaller context switching latency and memory usage based on single stack management. However, response time and preemptivity are much poorer than that of the multi-threaded systems[3]. The multi-threaded design enables preemption, but necessary memory space for the thread's stack is larger than that of the event-driven[5].

In this paper, we propose a OTL, which is an on-demand thread stack allocation scheme for MMU-less real-time sensor operating systems. The OTL enables to adaptively adjust stack size by allocating or releasing additional stack frame based on the amount of each function's stack usage. The information is calculated at compile-time, and adaptive adjustment of the stack occurs at run-time. Our experimental results show that OTL significantly minimizes the spatial overhead of the threads' stacks with tolerable time overhead compared with fixed stack allocation mechanism.

The rest of this paper is summarized as follows. In Section 2, we present some related works on sensor operating systems. Section 3 describes characteristics of the wireless sensor nodes, and explains the design and implementation of OTL in detail. Section 4 presents and evaluates performance of OTL compared with fixed stack allocation mechanism on the existing multi-threaded sensor operating systems. Finally, some conclusions and future work are presented in Section 5.

2 Related Works

In this section, we briefly introduce previous works relating to our proposing mechanism. Many kinds of research efforts[3,4,5,6,7,8] have been done on designing and implementing sensor operating systems and stack allocation mechanisms.

In [4], Adya et al. presented a "sweet spot" hybrid task management model of the two existing models(event-driven and multi-threaded). It can be implemented based on continuation and stack ripping mechanism, and the "sweet spot" enables cooperative task processing without manual stack management.

In [3], Ousterhout argued disadvantages of the multi-threaded task model with many usages and examples of the two existing models. As a result, the multi-threaded task model is better than event-driven if a target system is susceptible to the performance of response time or preemptivity of the task execution.

In [5], Behren et al. compared performance of the multi-threaded task model with event-driven. In addition, they insisted that the performance of the multi-threaded operating systems can be improved by compiler-assisted techniques.

In [9], Torgerson proposed an automatic thread stack management scheme on the MANTIS operating system based on compile-time calculation of the upper-bound of the thread stack size. The proposed scheme mitigates trade-off relation between stack overflow and stack space efficiency on the multi-threaded task model. However, this scheme cannot remove the space inefficiency which caused by static allocation of each thread stack area.

In [6], Gustafsson briefly introduced concepts of the stackless implementation of C/C++ programs to minimize memory usage. For example, C compiler can

predict maximum required stack space, the compiler would allocate space for function arguments and local variables on the heap instead of a stack.

In [10], Dunkels et al. proposed Protothread, which is a stackless threads in C programmming to minimize memory usage on implementing multi-threaded applications for wireless sensor networks. However, the Protothread is not a method of multi-threaded programming, but a programming sugar of event-driven programming. In addition, it does not support automatic state management and preemption between multi-threaded tasks.

3 OTL: On-Demand Thread Stack Allocation Scheme

In this section, we describe requirements for wireless sensor networks including platform hardware of the sensor nodes with some existing constraints. Then, we present design and implementation of OTL in detail.

3.1 Requirements for Wireless Sensor Networks

In wireless sensor networks, it is very important to minimize production cost to enhance networks efficiency. Such restriction limits computing power, memory space, and batteries of each sensor node. For example, Berkeley's MICA motes have only 8-bit processor, 4 KB memory space, and 2xAA batteries[11]. Thus, developers of the sensor operating systems must consider these severe constraints on the sensor hardware platform. In addition, sensor operating systems need to take the multi-threaded task model to support real-time, concurrent processing of the various tasks such as communication, sensing, analog-to-digital conversion, data aggregation, and etc. For this reason, operating systems that run on the sensor nodes need a space-efficient stack allocation mechanism.

The following are the essential requirements when designing a new stack allocation mechanism for the multi-threaded task model.

- ◇ **Minimization of allocated stack space.** In general purpose operating systems, each thread has a relatively large stack area(4 KB in Linux), but the memory space of a sensor node is very limited(about 1∼4 KB).
- ◇ **Minimization of time overhead.** Many existing software virtual memory or MMU mechanisms have large overhead on execution time, but computing power of a sensor node is much smaller than the other computing systems.

3.2 Design and Implementation of OTL

The major objective of the OTL is minimizing stack usage on multi-threaded operating systems. In order to efficiently utilize memory usage, it provides adaptive and on-demand stack allocation for each thread based on the calculation method of the amount of stack used for each function call.

Table 1 compares OTL with fixed stack allocation mechanism. In fixed(or static) stack allocation mechanism, thread stack allocation is performed only

Table 1. Comparison between fixed stack allocation mechanism and OTL

Feature	Fixed stack allocation	OTL
Allocation point	on creating a thread	on calling a function
Freeing point	on exiting a thread	on returning from a function
Stack frame size	fixed(or static)	adaptive(or dynamic)
Timing overhead	single allocation	multiple allocation

when a new thread is created, and the allocated space will be released only when the threads exits. In this mechanism, the allocated stack space is belong to a thread until the thread exits. If a thread stack's upper bound is significantly larger than average stack usage, the system will be inefficiently operated in terms of memory space utilization. On the otherhand, OTL adaptively allocates the necessary thread stack space on demand. Therefore, in OTL, the space efficiency will be significantly improved, but the several times of stack allocation may increase total execution time.

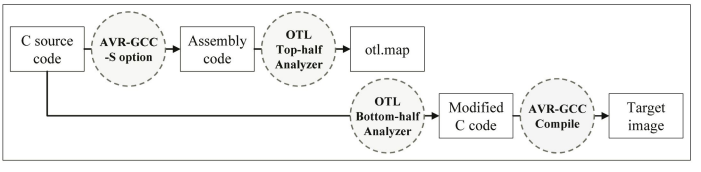


Fig. 1. Building process of sensor operating systems when using OTL

Figure 1 shows the building process of sensor operating systems and a pplications when using OTL. First of all, "OTL Library" is used for allocating and releasing thread stack space at thread's run-time. Second, the "Top-half Analyzer" is used for calculating the stack space usage of each function. AVR-GCC[12] is a C compiler for the AVR processor. The AVR-GCC with "-S" compiles a C source file(*.c), and it gives an assembly code file(*.s) as a output. The "Top-half Analyzer" parses and analyzes the output assembly code, and calculates the amount of stack space used for each function. The calculated results are stored at the file "otl.map". The third one is the "Bottom-half Analyzer". It is used for assigning the amount of stack space for each function call. The "Bottom-half Analyzer" parses the C source file(*.c), and then for each C function call, it finds the stack space usage of each function call from the file "otl.map". Then, the allocating and releasing OTL library routine for each function call is inserted to the before and after the call respectively. The modified C source files(*.c) is compiled. Finally, the kernel image of the operating system is built for wireless sensor nodes.

Figure 2 shows comparison between existing fixed stack allocation mechanism with the proposed scheme. This is an example of the same "pthread-net4: sensor110" application on the Nano-Qplus[8] sensor operating system. In case of

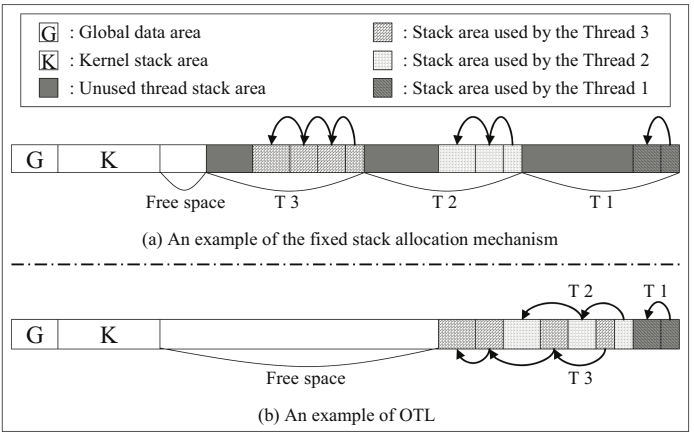


Fig. 2. An example of stack space of fixed stack allocation mechanism and the OTL

the fixed stack allocation, the stack space is managed as a single frame for each thread. However, in case of the OTL, the stack space is similar to the linked list structure, and the only necessary stack space is allocated for the threads. In the OTL, the maximum available number of threads are not affected by the fixed stack size, but the actual stack space used on the threads. Therefore, if the operating system uses the OTL, then the system can support more threads than existing fixed stack allocation mechanism.

Table 2. List of functions and variables used in this paper

Notation	Description
$allocMemory(m_{size})$	returning address of an allocated m_{size} memory block
$releaseMemory(m_{addr})$	releasing a memory block of an address m_{addr}
$saveSp(m_{addr})$	saving stack pointer to an address m_{addr}
$loadSp(m_{addr})$	loading stack pointer from an address m_{addr}
$changeSp(m_{value})$	changing stack pointer to a value m_{value}
$func().stackAddress$	stack frame address of a $func()$
$func().stackNecessary$	necessary stack frame size of a $func()$
$func().arguments$	arguments of a $func()$

We used the above notations to present algorithms of the proposed OTL. Some functions and variables that are used in protocol descriptions are defined in Table 2. The following shows the main algorithm of the proposed scheme.

In Algorithm 1, necessary stack allocation and saving stack pointer are performed before the function call. After that, the OTL moves the thread's stack pointer to the newly allocated stack area. Then, it pushes C function's arguments and returning address, sets the local variables, and runs the function's internal

Algorithm 1. Dynamic stack management algorithm of OTL

— when a thread calls a func(), do the following: —
$func().stackAddress \leftarrow allocMemory(func().stackNecessary)$
$saveSp(func().stackAddress)$
$changeSp(func().stackAddress + func().stackNecessary - 1)$
$push\ func().arguments$
$call\ func()$
$loadSp(func().stackAddress)$
$releaseMemory(func().stackAddress)$

routine. The OTL restores the thread's stack pointer from the saved data, and finally, the allocated stack space is released from the thread.

4 Performance Evaluation

In our performance evaluation, we compared the actual performance of the above mechanisms by making experiments on the real sensor platform. We used Octacomm's Nano-24 wireless sensor platform[13], which is similar to the Berkeley's MICAZ sensor platform. The proposed scheme is implemented on the Nano-Qplus(ver.1.6.0e). This operating system supports a real-time memory allocation scheme with preemptive task scheduler.

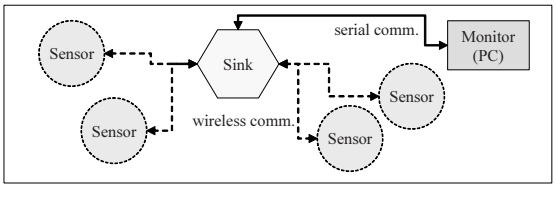
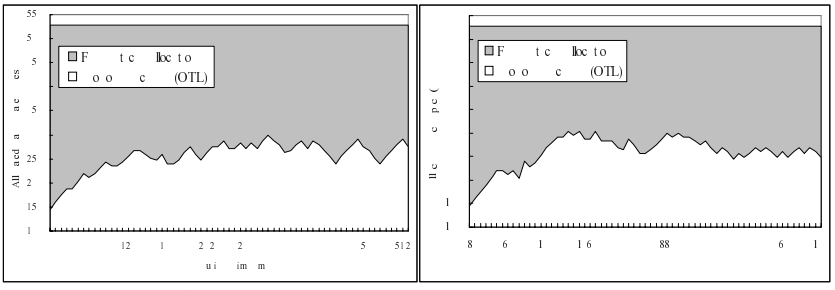


Fig. 3. Organization of the pthread-net4 sensor application

We used two kind of sensor applications which presented in the above, and the average memory usage and the average execution time are shown in Fig. 4 and 5, respectively.

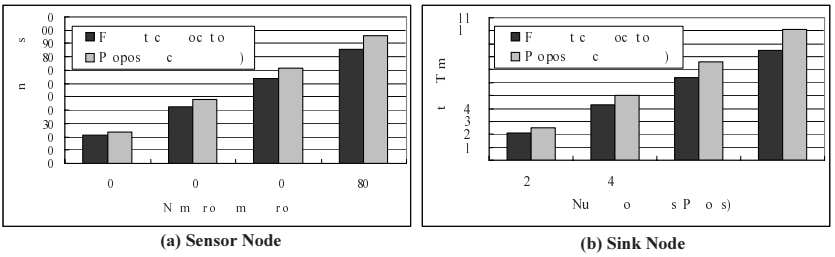
Figure 4 shows performance of the fixed allocation mechanism and the proposed scheme in terms of both time and space. In case of OTL, each function call needs additional 4 bytes to save previous stack pointer register. But it is very negligible space overhead, since it can save significant stack space compared with the fixed allocation mechanism.

Figure 5 shows the execution time when using the proposed scheme or not. The proposed scheme has a small timing overhead(about 10 percent) compared with fixed stack allocation mechanism. However, it can improve significantly the space-efficiency of the stack usage.



(a) Sensor Node (b) Sink Node

Fig. 4. Run-time stack space usage of the sensor applications



(a) Sensor Node (b) Sink Node

Fig. 5. Execution time of the sensor applications

Based on these results, we can convince that the proposed mechanisms significantly improves the space-efficiency of the threads' stack space allocation compared with the existing fixed stack allocation mechanism.

5 Conclusions

Wireless sensor networks are composed of tiny sensor nodes that usually have severe resource constraints in terms of energy, computing device and memory space. Sensor operating systems that run on tiny sensor nodes are the key to the performance of the wireless sensor networks. In order to process many tasks efficiently with the sensor nodes, the sensor operating systems need a space-efficient multi-threaded task model. In this paper, we propose a OTL, which is an on-demand thread stack allocation scheme for MMU-less real-time sensor operating systems. The OTL enables to adaptively adjust the stack size by allocating or releasing additional stack frame based on the amount of each function's stack usage information. The information of the function's stack usage is checked at compile-time, and the adaptive adjustment of the stack occurs at run-time. Our experimental results showed that the OTL can significantly minimize the spatial overhead of the threads' stacks with tolerable time overhead compared with fixed stack allocation mechanism on the existing multi-threaded sensor operating systems.

References

1. Hoare, C.A.R.: Monitors: An operating system structuring concept. Communications of the ACM **17** (1977) 549–557
2. Lauer, H.C., Needham, R.M.: On the duality of operating system structures. In: Second International Symposium on Operating Systems, IRIA. (1978)
3. Ousterhout, J.K.: Why threads are a bad idea (for most purposes). Presentation given at the 1996 Usenix Annual Technical Conference (1996)
4. Adya, A., Howell, J., Theimer, M., Bolosky, W.J., Douceur, J.R.: Cooperative task management without manual stack management. In: Proceedings of the 2002 USENIX Annual Technical Conference. (2002)
5. von Behren, R., Condit, J., Brewer, E.: Why events are a bad idea (for high-concurrency servers). In: HotOS IX: The 9th Workshop on Hot Topic in Operating Systems. (2003) 19–24
6. Gustafsson, A.: Threads without the pain. ACM Queue **3** (2005) 34–41
7. Yannakopoulos, J., Bilas, A.: Cormos: A communication-oriented runtime system for sensor networks. In: Proceedings of the 2nd IEEE European Workshop on Wireless Sensor Networks (EWSN '05). (2005)
8. Lee, K., Shin, Y., Choi, H., Park, S.: A design of sensor network system based on scalable and reconfigurable nano-os platform. In: IT-Soc International Conference. (2004)
9. Torgerson, A.: Automatic thread stack management for resource-constrained sensor operating systems. (2005)
10. Dunkels, A., Schmidt, O., Voigt, T.: Using protothreads for sensor node programming. In: Proceedings of the REALWSN'05 Workshop on Real-World Wireless Sensor Networks. (2005)
11. Crossbow: http://www.xbow.com/. (website)
12. AVR-GCC: http://www.avrfreaks.net/AVRGCC/. (website)
13. Octacomm: http://www.octacomm.net/. (website)

EF-Greedy: A Novel Garbage Collection Policy for Flash Memory Based Embedded Systems

Ohhoon Kwon, Jaewoo Lee, and Kern Koh

School of Computer Science and Engineering, Seoul National University
{ohkwon,jwlee,kernkoh}@oslab.snu.ac.kr

Abstract. Flash memory is becoming increasingly important for embedded systems because it has many attractive features such as small size, fast access speeds, shock resistance, and light weight. Although flash memory has attractive features, it should perform garbage collection, which includes erase operations. The erase operations are very slow, and usually decrease the performance of the system. Besides, the number of the erase operations allowed to each block is also limited. To minimize the garbage collection time and evenly wear out, our proposed garbage collection policy focuses on minimizing the garbage collection time and wear-leveling. Trace-driven simulations show that the proposed policy performs better than existing garbage collection policies in terms of the garbage collection time and the endurance of flash memory. Specifically, we have shown that the performance improvement of our proposed policy against the greedy policy in terms of the endurance of flash memory is as much as 90.6%.

Keywords: Flash memory, Garbage collection, Embedded systems.

1 Introduction

Flash memory is a non-volatile solid state memory, which is becoming increasingly important for digital computing devices because it has many attractive features such as small size, fast access speeds, shock resistance, high reliability, and light weight. Because of these features, flash memory will be widely used in various computing systems such as embedded systems, mobile computers, and consumer electronics.

Although flash memory has attractive features, it has a critical drawback, which is an inefficiency of in-place-update operation. When we update data, we can not write new data directly at same address due to physical characteristics of flash memory. All data in the block must first be copied to a system buffer and then updated. Then, after the block has been erased, all data must be written back from system buffer to the block. Therefore, updating even one byte data requires one slow erase and several write operations. Besides, if the block is a hot spot, it will soon be worn out.

To address the problem of the in-place-update operation, the out-place-update operation is exploited in many flash memory based systems [6-8]. When the data is updated, the out-place-update operation writes new date at new place, and then the obsolete data are left as garbage. When there are not enough free spaces in flash memory, we should collect the garbage space and translate a free space. This operation

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 913–920, 2007.
© Springer-Verlag Berlin Heidelberg 2007

Table 1. Performance of NAND flash memory [11]

	Page Read (2K bytes)	Page Write (2K bytes)	Block Erase (128K bytes)
Performance (μs)	25(Max.)	200(Typ.)	2000(Typ.)

is a garbage collection, which consists of the write operations and the erase operations. In the flash memory, the erase operations are very slow, and usually decrease the performance of the system. Besides, the number of the erase operations allowed to each block is also limited. In order to minimize the garbage collection time and evenly wear out flash memory, we propose a novel garbage collection policy in this paper. Our proposed garbage collection policy focuses on minimizing the garbage collection time as well as reducing the number of the erase operations and wear-leveling. Trace-driven simulations with HP traces show that our proposed policy performs better than the greedy, the Cost-Benefit (CB), and the Cost Age Time (CAT) policies in terms of the garbage collection time, the number of erase operations, and the endurance of flash memory.

The remainder of this paper is organized as follows. We review characteristics of flash memory and existing works on garbage collection in Section 2. Section 3 presents a new garbage collection policy for flash memory. We evaluate the performance of the proposed policy in Section 4. Finally, we conclude this paper in Section 5.

2 Related Works

In this section, we describe characteristics of flash memory and existing works on garbage collection.

2.1 Characteristics of Flash Memory

Flash memory is a non-volatile solid state memory, which has many attractive features such as small size, fast access speeds, shock resistance, high reliability, and light weight. Furthermore, its density and I/O performance have improved to a level at which it can be used as an auxiliary storage for mobile computing devices such as PDA and laptop computer. Flash memory is partitioned into blocks and each block has a fixed number of pages. Unlike hard disks, flash memory has three kinds of operations: page read, page write, and block erase operations. They have difference performances, and the performances of three kinds of operations are summarized in Table 1.

As aforementioned, flash memory has many attractive features. Flash memory, however, has two drawbacks. First, blocks of flash memory need to be erased before they are rewritten. The erase operation needs more time than read or write operation. The second drawback is that the number of erase operations allowed to each block is limited. This drawback becomes an obstacle to developing a reliable flash memory-based embedded system. Due to this drawback, the flash memory based embedded systems are required to wear down all blocks as evenly as possible, which is called wear-leveling.

2.2 Existing Works on Garbage Collection

Rosenblum et al. proposed the Log-Structured File System (LFS) and garbage collection policies have long been discussed in log-based disk storage systems [1-4]. Fortunately, the Log-Structured File System can be applied to flash memory based storage systems and the garbage collection policies in log-based disk storage also can be applied to flash memory based storage systems. Wu et al. proposed the greedy policy for garbage collection [5]. The greedy policy considers only valid data pages in blocks to reduce write cost and chooses the block with the least utilization. However it dose not consider wear-leveling. Therefore, it was shown to perform well for random localities of reference, but it was shown to perform poorly for high localities of reference. Kawaguchi et al. proposed the cost-benefit policy [6]. The cost-benefit policy evaluates the cost benefit of all blocks in flash memory using $((a*(1-u))/2u)$ method, where a is the elapsed time from the last data invalidation on the block, and u is the percentage of fullness of the block. After evaluating the all blocks, it chooses the victim block that has a maximum cost benefit value. Chiang et al. proposed the Cost Age Time (CAT) policy [7]. The CAT policy focuses on reducing the number of the erase operation. To reduce the number of the erase operations, they use a data redistribution method that uses a fine-grained method to separate cold and hot data. The method is similar to the cost-benefit policy but operates at the granularity of pages. Furthermore, the CAT policy considers wear-leveling. To perform even-leveling, the CAT chooses the victim block according to cleaning cost, ages of data in blocks, and the number of the erase operations. Kim et al. proposed the cleaning cost policy, which focuses on lowering cleaning cost and evenly utilizing flash memory blocks. In this policy, they dynamically separates cold data and hot data and periodically move valid data among blocks so that blocks have more even life times [9]. Chang et al. proposed the real-time garbage collection policy, which provides a guaranteed performance for hard real-time systems [10]. They also resolved the endurance problem by the wear-leveling method.

3 EF-Greedy: A Novel Garbage Collection Policy

In flash memory, the erase operation is even slower than the read and write operation. Thus, the erase operation is dominant to the performance of the flash memory based system. As mentioned in Section 2, to improve the performance, existing works for garbage collection tried to reduce the number of the erase operations. They also considered the wear-leveling for the endurance of flash memory.

In this paper, we propose the new garbage collection policy, which extends the greedy policy and focuses on minimizing the garbage collection time as well as wear-leveling. Thus, our proposed garbage collection policy is named 'EF-Greedy', which stands for '*Endurant and Fast Greedy*'.

3.1 Block Recycling Scheme

In flash memory, if there are not enough free blocks, the system should perform garbage collection. During garbage collection, we should wait and do not perform any operations such as read and write operations until the garbage collection finishes. To improve the performance of the flash memory based system, we should minimize the

garbage collection time. In this paper, we use the greedy policy to make a decision which block should be erased during garbage collection. Because the greedy policy considers only valid pages in blocks and chooses the block with the least utilization, we can minimize the garbage collection time. However it dose not consider wear-leveling and was shown perform poorly for high localities of reference. To address the problems of the greedy policy, we extend the greedy policy by considering the different update interval of the pages in the blocks and the number of the erase operation of the blocks.

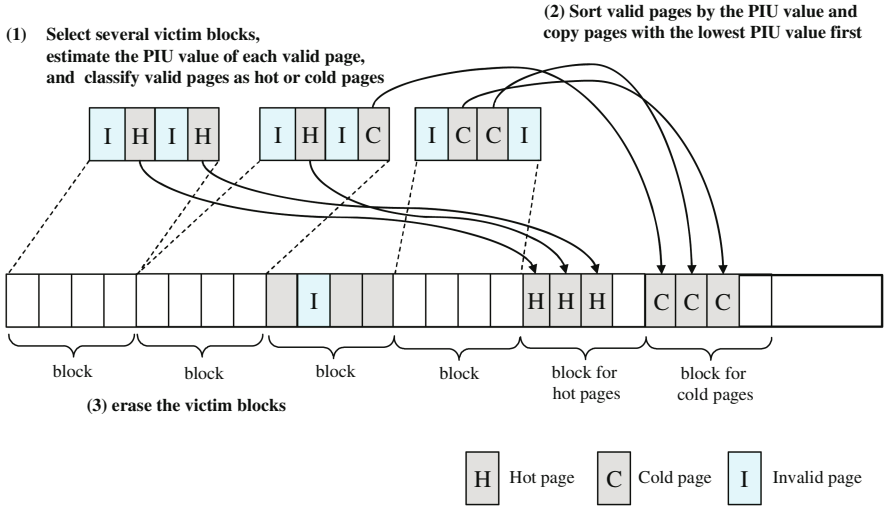


Fig. 1. The classification of the pages and the redistribution of the valid pages

When we perform garbage collection, we select several victim blocks with the least utilization, and then copy valid pages in the victim blocks to the free block before we clean the block. For the redistribution of valid pages, we should classify data pages as hot and cold pages. Hot valid pages are updated frequently at recent time, and cold valid pages are not updated frequently. During redistributing, hot valid pages are redistributed to hot blocks, while cold valid pages are redistributed to cold blocks. To classify data pages, we estimate the update interval of each page based on past update behavior, and then use this update interval information in deciding whether the page is hot or cold in the EF-Greedy policy. Expression (1) represents the calculation of the *predicted inter-update time (PIU)* based on past update times. The *predicted inter-update time* denotes the interval between future update time and latest update time on an identical page. Let I_k be the (k)th inter-update time, I_{k-1} be the ($k–1$)th inter-update time, and I_{k-2} be the ($k–2$)th inter-update time. Then, the kth predicted inter-update time PIU_k is computed as

$$PIU_k = \frac{\sum_{j=k-n+1}^{k} I_j}{n} \tag{1}$$

where n is the number of the past inter-update time. We, hence, just consider the last inter-update time if n is 1, and consider the whole inter-update times if n is k. EF-Greedy sets the default values of n as 3. If the *PIU* value of the page is greater than the average *PIU* value of all pages, the page is classified the cold page, otherwise the page is classified the hot page. Fig. 1 shows the classification of the pages and the redistribution of the valid pages.

3.2 Efficient Free Block Lists Management Scheme

Flash memory should be controlled to evenly wear out all blocks since wearing out specific blocks could limit the usefulness of the whole flash memory. Thus, most of the existing works considered wear-leveling of flash memory when the victim block is selected. In contrast, our proposed policy does not consider wear-leveling similar to the greedy policy when the victim block is selected. In order to guarantee the long endurance of flash memory, we propose an efficient free block lists management scheme for wear-leveling on flash memory. In our proposed policy, we use two free lists: hot free block list and cold free block list. After cleaning the victim blocks, if the number of the erase operation of the block is greater than the average number of the

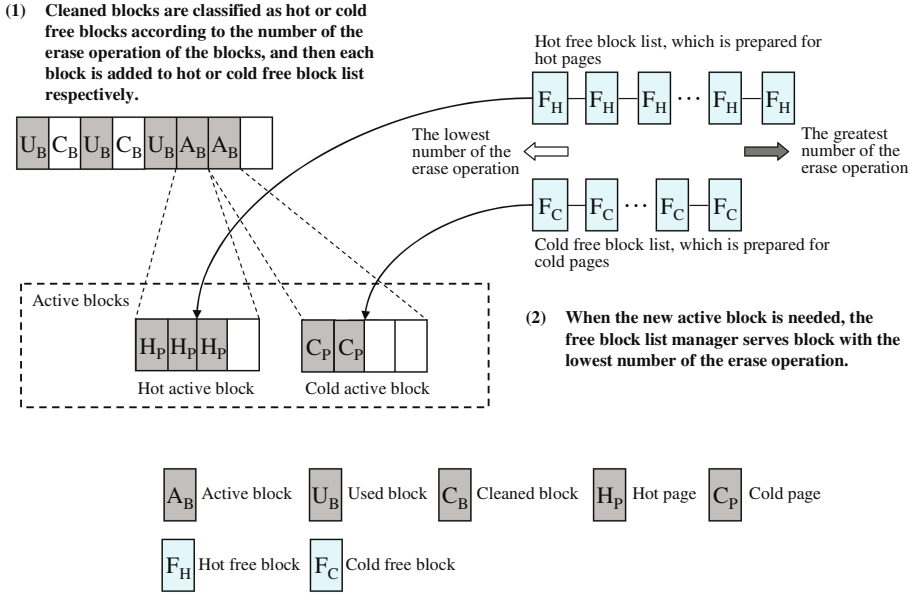


Fig. 2. The efficient two free block lists management scheme for wear-leveling

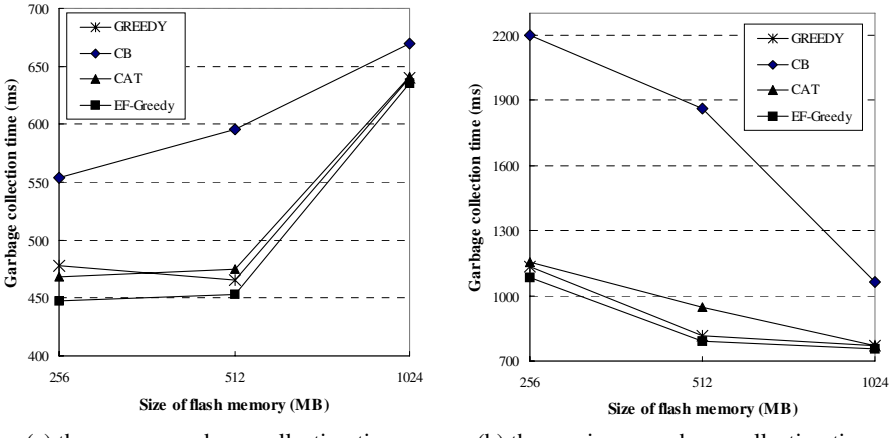
erase operation of all free blocks, the block is added to the cold free block list. Otherwise, the block is added to the hot free block list. And then the free blocks in each free block list are sorted by the number of the erase operation of the block. Hence, during copying out, we could allocate the block with the minimum number of the erase operations to valid pages, and could evenly wear out. Fig. 2 shows the efficient free block lists management scheme for wear-leveling.

4 Performance Evaluation

In this section, we present the performance evaluation results for various garbage collection policies to assess the effectiveness of our proposed policy. We conducted trace-driven simulations with the HP traces to compare the performance of our proposed policy with those of the greedy, the Cost-benefit (CB), and the Cost Age Time (CAT) policies.

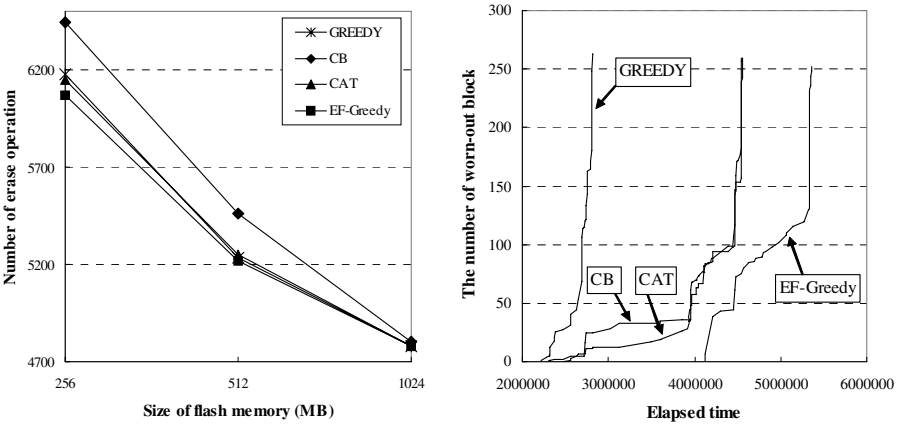
The HP traces are disk-level traces of personal workstation (hplajw) at Hewlett Packard laboratories, which were used for document editing and electronic mail. They also exhibit high locality of reference that 71.2% of writes were to metadata [12]. Since the size of the disks used in hplajw does not equal to the size of the flash memory used in simulations, the traces were preprocessed to map flash memory spaces before simulation. To evaluate the performance, when the size of free block is fewer than 5% of the total size of flash memory, garbage collection is started. And garbage collection is stopped when the size of free block is larger than 10% of the total size of flash memory. Fig. 3 shows the garbage collection time for the four policies as a function of the flash memory size. Our proposed policy shows better performance in terms of the average and maximum garbage collection time. This is because the EF-Greedy policy just considers the utilization of each block to minimize the garbage collection time unlike other three policies. Furthermore, our proposed policy performs better than the original greedy policy because it estimates the predicted inter-update time (PIU) of each page and exploits the PIU value to redistribute pages.

Fig. 4 shows the performance results of the number of erase operation and worn-out blocks. In these results, our proposed policy shows the best performance in terms of the number of worn-out blocks due to the efficient free block lists management scheme. This result means that our proposed policy guarantees the long endurance of flash memory. Specifically, the performance improvement of EF-Greedy against the greedy policy is as much as 90.6% in terms of the endurance of flash memory. The result of the number of erase operation also shows better performance in our proposed policy.



(a) the average garbage collection time (b) the maximum garbage collection time

Fig. 3. The performance results of the garbage collection time



(a) the number of erase operation (b) the number of worn-out blocks

Fig. 4. The performance results of the number of erase operation and worn-out blocks

5 Conclusion

In this paper, we presented the novel garbage collection policy for flash memory based embedded systems. To minimize the garbage collection time and guarantee the endurance of flash memory, we extended the greedy by considering the different update interval of the pages and the efficient free block lists management scheme. As a result, EF-Greedy performs better than other existing garbage collection policies in terms of the garbage collection time, the number of erase operations, and the endurance of flash memory. Specifically, we have shown that the performance improvement of EF-Greedy policy against the greedy policy in terms of the endurance of flash memory is as much as 90.6%.

Acknowledgments. The authors thank Hewlett Packard Laboratories for making their I/O traces available.

References

1. Rosenblum, M., Ousterhout, J. K.: The Design and Implementation of a Log-Structured FileSystem. ACM Transactions on Computer Systems, Vol. 10, No. 1 (1992)
2. Blackwell, T., Harris, J., Seltzer, M.: Heuristic Cleaning Algorithms in Log-Structured File Systems. Proceedings of the 1995 USENIX Technical Conference, Jan. (1995)
3. Matthews, J. N., Roselli, D., Costello, A. M., Wang, R. Y., Anderson, T. E.: Improving the Performance of Log-Structured File Systems with Adaptive Methods. Proceedings of the Sixteenth ACM Symposium on Operating System Principles (1997)
4. Seltzer, M., Bostic, K., McKusick, M. K., Staelin, C.: An Implementation of a Log-Structured File System for UNIX. Proceedings of the 1993 Winter USENIX (1993).
5. Wu, M., Zwaenepoel, W.: eNVy: A Non-Volatile, Main Memory Storage System. Proceedings of the 6th International Conference on Architectural Support for Programming Languages and Operating Systems (1994)

6. Kawaguchi, A., Nishioka, S., and Motoda, H.: A Flash-Memory Based File System. Proceedings of USENIX Technical Conference (1995)
7. Mei-Ling Chiang, Paul C. H. Lee, Ruei-Chuan Chang: Cleaning policies in mobile computers using flash memory. Journal of Systems and Software, Vol. 48, (1999)
8. Torelli, P.: The Microsoft Flash File System. Dr. Dobb's Journal, Feb. (1995)
9. Hanjoon Kim, Sanggoo Lee, S. G.: A new flash memory management for flash storage system. Proceedings of the Computer Software and Applications Conference (1999)
10. Li-Pin Chang, Tei-Wei Kuo, Shi-Wu Lo: Real-time garbage collection for flash-memory storage systems of real-time embedded systems. ACM Transactions on Embedded Computing Systems, Vol. 3. (2004)
11. Samsung Electronics: 128M x 8 Bit NAND Flash Memory. http://www.samsung.com
12. Ruemmler, C., Wilkes, J.: UNIX Disk Access Patterns. Proceedings of the 1993 Winter USENIX (1993)

Power-Directed Software Prefetching Algorithm with Dynamic Voltage Scaling*

Juan Chen, Yong Dong, Huizhan Yi, and Xuejun Yang

School of Computer, National University of Defense Technology, P. R. China
{juanchen,yongdong,huizhanyi,xjyang}@nudt.edu.cn

Abstract. We first demonstrate software prefetching provides an average 66.28% performance enhancement with much higher average power on six memory-intensive benchmarks. Then we propose a power-directed software prefetching algorithm with dynamic voltage scaling (*PDP-DVS*) that monitors a system's power and adapts the voltage level accordingly to guarantee no power increase while maintaining good performance boost. Our *PDP-DVS* algorithm achieves a 35.75% performance gain with only 1.19% power increase.

1 Introduction

High power consumption has become an important limiting factor in developing designs for battery-operated embedded systems due to exorbitant cooling, packing and power costs. Unfortunately, some traditional compiler optimization techniques are only aimed at improving program performance, which causes significant power increase. For example, software prefetching[1] improves the performance by overlapping CPU computing and memory access operation. However, inserting prefetching instruction and overlapping computing operation and memory access increase the probability of processor unit utility, which leads to power increase as Fig. 2 and Fig. 3 show. In this paper, we propose a power-directed software prefetching algorithm with dynamic voltage scaling (*PDP-DVS*), which eliminates power increase due to software prefetching while obtaining significant performance enhancement.

Agarwal et al. [2] presented a similar work, but their objective is to reduce energy consumption without performance loss. Furthermore, our *PDP-DVS* algorithm uses selective prefetching method besides DVS method.

2 Power-Directed Software Prefetching

We use the algorithm developed by Mowry[1] for prefetching affine array accesses and indexed array references. Let P_p and P_{np} be the average powers to run one part of codes with and without prefetching, respectively. To reduce power from P_p to P_{np}, the

* This work was supported by the Program of Nature Science Fund under Grant No. 60633050 and was supported by the National High Technology Development 863 Program of China under Grant No. 2002AA1Z2101 and No. 2004AA1Z2210.

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 921–924, 2007.
© Springer-Verlag Berlin Heidelberg 2007

optimal way to is to reduce voltage to v^* from v such that $\frac{P_p}{P_{np}} = \frac{v(v-v_t)^\alpha}{v^*(v^*-v_t)^\alpha}$ with respect to $P \propto Cv^2 f$ and $f \propto (v-v_t)^\alpha / v$.

```
1. ALGORITHM: PDP-DVS
2. INPUT: A program with prefetching, power without prefetching
          (objpower)
3. OUTPUT: A program with real time voltage scaling
4. v=vmax;
5. repeat for each N instructions till the completion of the application {
6.     execute M instructions with prefetching at v;
7.     Pp=average power for the execution of these M instructions;
8.     execute the next M instructions without prefetching at v;
9.     Pnp=average power for the execution of these M instructions;
10.    Tnp=execution time for the execution of these M instructions;
11.    ratio=Pp/Pnp;
12.    if (ratio <=1)
13.        execute the rest instructions with prefetching at
           the current voltage v;
14.    else {
15.        calculate new voltage v* such that ratio = v(v-vt)^α / v*(v*-vt)^α
16.        execute the next M instructions with prefetching at v*;
17.        Plp=average power for the execution of these M instructions;
18.        Tlp=execution time for the execution of these M instructions;
19.        if (Tlp>Tnp)
20.            execute the rest instructions without prefetching at v;
21.        else
22.            execute the rest instructions with prefetching at v*;
23.    }
24.    if (avepower > objpower)
25.        voltage_down();
26.    else
27.        voltage_up();
28. }
```

Fig. 1. Pseudo code for the profiling DVS algorithm *PDP-DVS*

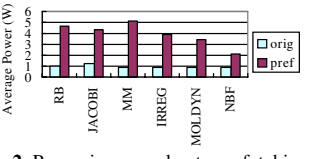


Fig. 2. Power increase due to prefetching

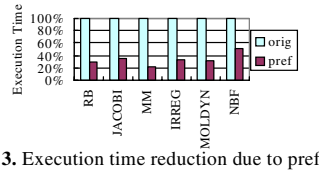


Fig. 3. Execution time reduction due to prefetching

Bechmark	Power Increase by DVS	Time Reduction by DVS	Power Increase by SVS	Time Reduction by SVS
RB	4.92%	37.61%	25.81%	44.66%
JACOBI	4.74%	43.16%	26.91%	49.60%
MATMULT	0.53%	43.89%	28.39%	52.80%
IRREG	0.37%	30.93%	22.64%	39.61%
MOLDYN	-0.24%	36.47%	30.75%	46.22%
NBF	-3.16%	22.44%	14.94%	30.29%
Average	1.19%	35.75%	24.91%	43.86%

Table 1. Power and Performance Gain Achieved by DVS and SVS

Fig. 1 illustrates our power-directed software prefetching algorithm with DVS (*PDP-DVS*). This algorithm periodically conducts periodic profiling to estimate the average power increase due to prefetching and eliminate such power increase by DVS and selective prefetching.

For each N instructions (called **repetitive period**), we execute the first M instructions with prefetching and the next M instructions without prefetching. The latter can be achieved by treating prefetch instructions as NOPs. Assume that they take (P_p) and (P_{np}) of average power dissipation, respectively and take T_{np} of execution time without prefetching. If the ratio P_p to P_{np} is no more than 1, prefetching does not bring power increase, and we remain the voltage unchanged to execute the remaining instructions with prefetching (steps 12-13). We call executing these remaining instructions as **profile-guided period**, which is equal to (N-2*M) instructions and we call the previous two M instruction as **profiling period**. That is, each repetitive period consists of profiling period and profile-guided period.

When we identify power increase, we execute the next M instructions with prefetching at lower voltage v^*. Note in this case, profiling period includes 3*M instructions instead of 2*M instructions. Assume the average power for the third M instructions is P_{lp} and execution time is T_{lp}. $T_{lp} > T_{np}$ represents prefetching cannot bring any performance boost. In this case, non-prefetching is better than prefetching and we will give up prefetching for the remaining instructions while remaining the previous operating voltage v (steps 19-20). We scale voltage up only when cumulative

average power (*avepower*) is lower than the objective power (*objpower*) (steps 26-27). Here objective power is the power for non-prefetching version. We use the following parameters during our simulation: *N*=100k instructions (repetitive period), *M*=5k (profiling period) and *objpower* (power for non-prefetching version).

Our experiment uses the six benchmarks, among which the first three applications perform affine array accesses and the others perform indexed array accesses. The description for all the benchmarks can be found in [3].

We use Wattch[4] to implement our *PDP-DVS* algorithm. Perfetch instruction was added to the ISA of the processor model and our simulation is based on 0.1 μm process technology parameters with ideal clock gating. We use 800MHz and 1.65V as the baseline frequency and voltage. And the relationship between frequency *f* and voltage *v* meets: $f \propto (v - v_t)^{\alpha} / v$, where α =2.

A suitable voltage level for the whole program according to the average power increase ratio is commonly not exact enough because each part of the program assumes different power consumption. From Table 1, the power dissipation after such static voltage scaling is still 24.91% larger than the original version. In contrast, *PDP-DVS* algorithm obtains 35.75% performance gain (time reduction) with only 1.19% power increase in average. Power increase ratio and time reduction ratio both take the non-prefetching version as a basis. For RB and JACOBI, our online *PDP-DVS* algorithm causes more power increase than other applications because RB and JACOBI have much less repetitive periods (only 62) than other applications.

Fig. 4 gives the dynamic voltage scaling results throughout a complete run of two applications. The others have the similar curves. During our online *PDP-DVS*, there isn't a steady optimal voltage at which system power can eliminate power increase completely. Instead, the optimal voltage steadily fluctuates within a range as Fig. 4 shows. In this Figure, horizontal axis represents number of repetitive periods. Vertical axis represents the optimal operating voltage for each profile-guided period.

In theory, ideal static operating voltage should lie in the midst of the peak value and lowest value of wave. However, the voltage setting by SVS is always higher this ideal value. We can see this from Fig. 4, where a red straight line represents static voltage scaling results. For MOLDYN, static voltage setting is almost at the peak value of wave so that its power increase ratio by SVS is the biggest in Table 1.

Take RB as an example, we notice each repetitive period almost includes the same number of prefetch instructions as Fig. 5 shows. Fig. 6 shows the average power of each repetitive period before applying *PDP-DVS* algorithm. One can see the average power of each repetitive period has periodic varying trend, which is different from cumulative power. **Cumulative power** at one point is the ratio of current total energy to the whole execution time as Fig. 7 shows. At the end of application execution, cumulative power is equal to the average power of the whole program. After using *PDP-DVS* algorithm, cumulative power reduces greatly at first, and fluctuates around 1W as Fig. 8 shows.

In our algorithm *PDP-DVS*, selective prefetching is used (steps 15-22 in *PDP-DVS*). Fig. 9 gives the selective prefetching profiles for RB. Due to space limitations, we omit the other applications' profiles, which have similar results. In Fig. 9, vertical axis represents selecting prefetching or not for each repetitive period, where "1" represents prefetching and "0" represents non-prefetching. Take RB as an example, there are 35 repetitive periods selecting non-prefetching among all 62 repetitive periods. Selective prefetching result for RB is simply denoted as 35/62. Similarly, 28/62 for

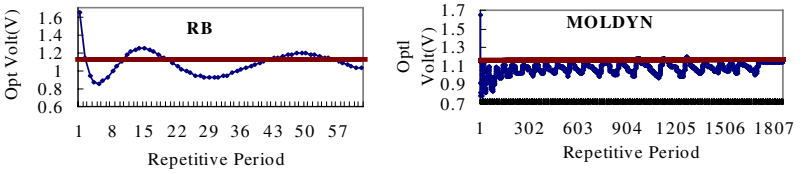


Fig. 4. Dynamic voltage scaling results by *PDP-DVS* algorithm and static voltage scaling results

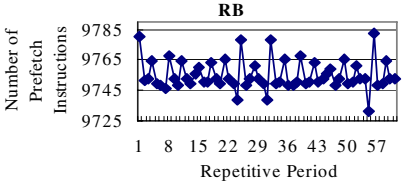


Fig. 5. The number of prefetch instructions for each repetitive period for RB

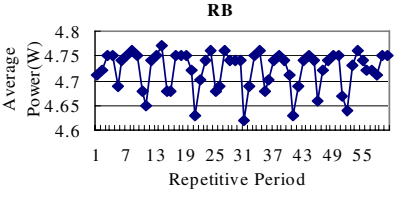


Fig. 6. Average power of each repetitive period before using *PDP-DVS* algorithm for RB

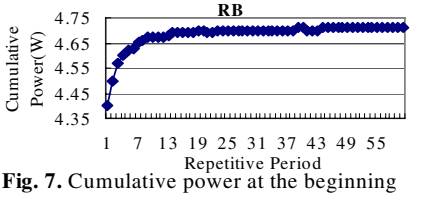


Fig. 7. Cumulative power at the beginning point of each repetitive period before using *PDP-DVS* algorithm for RB

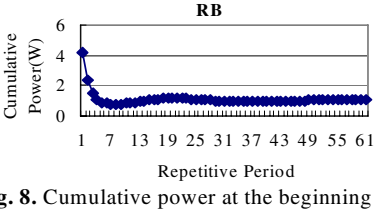


Fig. 8. Cumulative power at the beginning point of each repetitive period after using *PDP-DVS* algorithm for RB

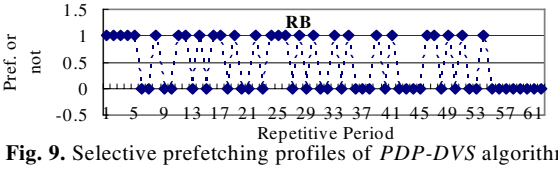


Fig. 9. Selective prefetching profiles of *PDP-DVS* algorithm

JACOBI, 351/669 for MATMULT, 299/556 for IRREG, 1018/1858 for MOLDYN and 231/413 for NBF.

References

[1] Todd C. Mowry. Tolerating Latency Through Software-Controlled Data Prefetching. Ph. D. thesis. Stanford University. Computer System Laboratory. March 1994.

[2] Deepak N. Agarwal, Sumitkumar N. Pamnani, Gang Qu and Donald Yeung. Transferring Performance Gain from Software Prefetching to Energy Reduction. In *Proceedings of the 2004 International Symposium on Circuits and Systems*. Vancouver, Canada. May 2004.

[3] Abdel-Hameed Badawy, Aneesh Aggarwal et al. The Efficacy of Software Prefetching and Locality Optimizations on Future Memory Systems. *Journal of Instruction-Level Parallelism*, .2004.

[4] D. Brooks, V. Tiwari and M. Martonosi. Wattch: A framework for architectural-level power analysis and optimizations. In *Proceedings of 27th International Symposium on Computer Architecture*. June 2000. p83-94.

An Efficient Bandwidth Reclaim Scheme for the Integrated Transmission of Real-Time and Non-Real-Time Messages on the WLAN

Junghoon Lee[1], In-Hye Shin[1], Gyung-Leen Park[1], Wang-Cheol Song[2], Jinhwan Kim[3], Pankoo Kim[4], and Jiman Hong[5,⋆]

[1] Dept. of Computer Science and Statistics, Cheju National University
[2] Dept. of Computer Engineering, Cheju National University
[3] Dept. of Multimedia Engineering, Hansung University
[4] Dept. of Computer Engineering, Chosun University
[5] School of Computing, Soongsil University
jhlee@cheju.ac.kr, kimjh@hansung.ac.kr, pkkim@chosun.ac.kr, gman@ssu.ac.kr

Abstract. This paper proposes and analyzes bandwidth reclaim scheme for IEEE 802.11 WLAN, which may suffer from severe bandwidth waste resulting from not only the variation of transmission rate and message length but also the overallocation to the real-time traffic in compensating for the delay due to the intervention of non-real-time messages. Built on top of the weighted round robin scheduling policy, we address that the polling order rearrangement according to the degree of overallocation can enhance reclaimability of unused network time and that the rearranegable slot has its message pending at the rearranging time. The simulation results show that the proposed scheme is able to reclaim up to 52.3 % of bandwidth waste when the number of streams is 2 and that it also provides stable throughput for utilization of 0.5 through 0.8.

1 Introduction

[1] According to the expansion of WLAN (Wireless Local Area Network), real-time and non-real-time messages coexist in the wireless media. Real-time traffic such as video and sensor data requires bounded delay, but is usually tolerant of some packet losses. As contrast, non-real-time traffic requires loss-free transmission without demanding bounded delay[1]. The IEEE 802.11 was developed as a MAC (Medium Access Control) standard for WLAN and this standard consists of both an essential DCF (Distributed Coordination Function) and an optional PCF (Point Coordination Function)[2]. The DCF exploits collision-based CSMA/CA (Carrier Sense Multiple Access with Collision Avoidance) protocol for non-real-time messages, aiming at enhancing their average delivery time as well as overall

⋆ Corresponding author.

[1] This research was supported by the MIC, Korea, under the ITRC support program supervised by the IITA (IITA-2006-C1090-0603-0040).

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 925–932, 2007.
© Springer-Verlag Berlin Heidelberg 2007

network throughput. On the other hand, the collision-free PCF can provide a real-time guarantee by developing a deterministic access schedule[3].

Network management consists of two parts on real-time communications, namely, static bandwidth allocation and dynamic adaptation parts, respectively [4]. Based on the static informations that do not change for a long time, for example, period and maximum transmission time of each stream, the bandwidth allocation procedure determines the network access schedule for the given set of active streams. However, the dynamic change in network condition needs additional management that can cope with such situations as non-real-time traffic load oscillation, channel status change, and so on. One of challenging problems in WLAN is a bandwidth reclaim scheme that reassigns the network time reserved but unused to another node[3]. In particular, the auto-selection mechanism can create much of such unused bandwidth, as it chooses the rate to be used for each packet that is submitted to the physical layer[5]. If a stream meets better rate than expected for a time interval, it can complete its transmission early.

The reclaim scheme is very crucial to the network throughput, as hard real-time guarantee inevitably brings bandwidth overallocation resulted from a pessimistic assumption that the stream has the worst case available time at each period. Moreover, a phenomenon called as a *deferred beacon problem*, which will be discussed in Section 3.1, further deteriorates the worst case available time for the real-time stream, increasing the amount of overbooking. It is desirable that as much unused bandwidth as possible should be reclaimed and allocated to another stream to minimize bandwidth waste. To this end, this paper is to propose and analyze a bandwidth reclaim scheme for WLAN that strictly obeys the IEEE 802.11 standard, assuming that PCF operates according to the weighted round robin schedule. We can improve the amount of reclaimed bandwidth by adjusting the polling order.

This paper is organized as follows: After issuing the problem in Section 1, Section 2 introduces the related works on both network scheduling and bandwidth reclaim schemes. With the description on network and message models along with a bandwidth allocation scheme in Section 3, Section 4 proposes a reclaim procedure. Section 5 discusses the performance measurement results and then Section 6 finally concludes this paper with a brief summarization and the description of future works.

2 Related Works

Based on the attribute that most real-time traffic is periodic, several MAC protocols have been proposed to support the hard real-time communication over a wireless channel[6]. However, they cannot be easily exploited to the IEEE 802.11 WLAN standard, as they ignored the CSMA/CA part defined as mandatory in the WLAN standard, or just aimed to enhance the ratio of timely delivery for soft multimedia applications[4]. For example, Choi and Shin suggested a unified protocol for real-time and non-real-time communications in wireless networks[1]. Based on frame-structured access mechanism, a BS (Base Station) polls every

station, be it a real-time or a non-real-time one, according to the corresponding policy. Though unpredictability due to message collision is eliminated, this scheme is neither compatible with the standard CSMA/CA protocol, nor takes into account any resource reclaim scheme.

Most works that conform to the IEEE standard are aiming at enhancing the ratio of timely delivery for soft multimedia applications, rather than providing a hard real-time guarantee. DBASE (Distributed Bandwidth Allocation/Sharing/Extension) is a protocol that supports multimedia traffics over IEEE 802.11 *ad hoc* WLAN[4]. The basic concept is that each time real-time station transmits its packet it will also declare and reserve the bandwidth needed at the next CFP. Every station respectively collects this information and then calculates its actual bandwidth at the next cycle. Though the per-packet reservation makes a resource reclaim scheme unnecessary, it does not only increase the runtime burden on member stations, but also demands all stations receive all the transmitted packets.

M. Caccamo and his colleagues have proposed a MAC capable of supporting deterministic real-time scheduling by implementing TDMA[3]. Referred as *implicit contention*, their scheme makes every station respectively run the common real-time scheduling algorithm to determine which message can access the medium. Accompanied with the implicit contention, FRASH (FRAame SHaring) can reclaim the unused bandwidth. Whenever the transmission of the current dispatched message is over and it does not use all the reserved frames, its identifier is put in a field in the header of the last data packet of the current message. However, the identifier should be correctly received by all stations in the network to reach a global agreement. FRASH can perform properly only for TDMA protocols that operate on fixed size slots. Moreover, this scheme cannot be implemented in the current 802.11 WLAN standard without the addition of new management frames and thus causes additional overhead.

3 Backgrounds

3.1 Network and Message Models

In BSS (Basic Service Set), the time axis of WLAN is divided into a series of superframes, each of which alternately operates CP (Collision Period) and CFP (Collision Free Period) phases, which are mapped to PCF and DCF, respectively. PC (Point Coordinator) node, typically AP (Access Point), sequentially polls each station during CFP according to the polling schedule determined by a specific policy such as EDF (Earliest Deadline First). All stations including the ones even in the polling list contend in the CP to send a management frame or a control frame. Even in the *ad hoc* mode, it is possible to designate a specific node to play a role of PC in a target group.

The phase of network operation is managed by the exchange of control packets which have higher priority than other ordinary packets. The PC attempts to initiate CFP by broadcasting a *Beacon* at regular intervals derived from a network parameter of *CFPRate*. Round robin is one of the commonly used polling

policies for CFP, in which every node is polled once a polling round. A polling round may be completed within one superframe, or spread over more than one superframe. In case the CFP terminates before all stations have been completely polled, the polling list is resumed at the next node in the ensuing CFP cycle. The polled node transmits its message for up to a predefined time interval, and always responds to a poll immediately whether it has a pending message or not.

To prevent starvation of stations that are not allowed to send during the CFP, a superframe is forced to include a CP of minimum length that allows at least one data packet delivery under DCF[2]. Hence, a non-real-time packet may occupy the network when the coordinator is to send the beacon frame. However, only after the medium is idle the coordinator will get the higher priority due to the shorter IFS (InterFrame Space). Thus, the delivery of a beacon frame can get delayed, resulting in the *deferred beacon problem*, possibly invalidating the network schedule determined for real-time messages. The maximum amount of deferment coincides with the maximum length of a non-real-time packet.

The real-time traffic is typically *isochronous* (or synchronous), consisting of message streams that are generated by their sources on a continuing basis and delivered to their respective destinations also on a continuing basis[6]. For example, a sensor node periodically reports the collected sensor data to a remote server. Accordingly, the general real-time message model consists of n streams, namely, S_1, S_2, ..., S_n, and for each S_i, a message sized to at most C_i arrives at the beginning of its period, P_i, and it must be transmitted by P_i. M_i is the transmission time of C_i, estimated with a reference transmission rate, R_i, which can be set empirically or analytically. Small value of R_i increases the probability that the actual transmission rate is above R_i, bringing more unused bandwidth. Finally, the destination of message can be either within a cell or outside a cell, and the outbound messages are first sent to the AP and then forwarded to the final destination, or vice versa. In case of a change in the stream set, bandwidth should be reallocated[1].

3.2 Bandwidth Allocation

This subsection briefly describes the allocation scheme of Lee's work on which this paper is built. For detailed description of bandwidth allocation, refer to [7]. To begin with, by allocation, we mean the procedure of determining capacity vector, $\{H_i\}$, for the given superframe time, F, and message stream set described as $\{S_i(P_i, M_i)\}$. As shown in Fig. 1, H_i denotes the time amount during which

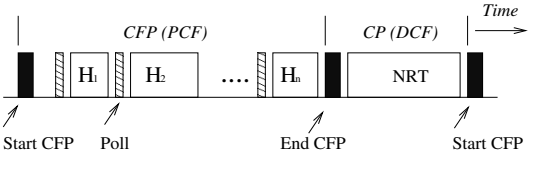


Fig. 1. Polling procedure and capacity vector

S_i can send its message when it is polled. A stream can timely send C_i only if its average transmission rate is over R_i during the P_i.

Let δ denote the total overhead of a superframe including polling latency, IFS, exchange of beacon frame, and the like, while D_{max} the maximum length of a non-real-time data packet. In addition, P_{min} denotes the smallest element of set $\{P_i\}$. Then the requirement for the superframe time, F, can be summarized as in Ineq. (1). Within this range, the scheme can select F and modify some of P_i's such that they are harmonic[8].

$$\sum H_i + \delta + 2 \cdot D_{max} \leq F \leq P_{min} \tag{1}$$

In addition, the least bound of H_i that can meet the time constraint of S_i is calculated as in Eq. (2).

$$\begin{array}{ll} H_i = \frac{M_i}{(\lfloor \frac{P_i}{F} \rfloor - 1)} & if(P_i - \lfloor \frac{P_i}{F} \rfloor \cdot F) \leq D_{max} \\ H_i = \frac{M_i}{\lfloor \frac{P_i}{F} \rfloor} & Otherwise \end{array} \tag{2}$$

The allocation vector calculated by Eq. (2) is a feasible schedule if the vector meets Ineq. (1). Finally, we can determine the length of CFP (T_{CFP}) and that of CP (T_{CP}) as follows:

$$T_{CFP} = \sum H_i + \delta, \quad T_{CP} = F - T_{CFP} \geq D_{max} \tag{3}$$

4 Bandwidth Reclaim Scheme

4.1 Reclaim Test

Hard real-time guarantee is given by the worst case available transmission time which is calculated with a pessimistic assumption that the transmission rate of S_i is just R_i while the size of message is always C_i. So a stream can meet extra slots in some periods if any of above conditions are not met. As a result, a node

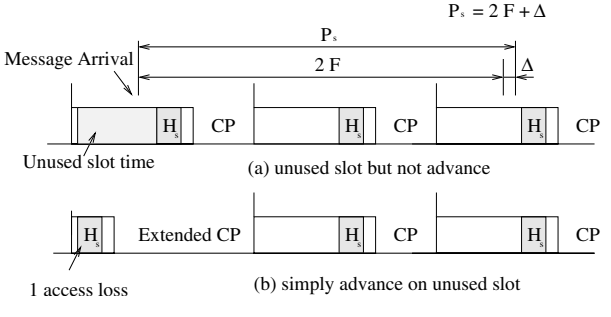


Fig. 2. Bandwidth reclaiming

may have no pending message when it receives a poll, in which case it responds with a null frame containing no payload. How to cope with this unused slot is critical to the network throughput. The first step to reclaim the bandwidth is to determine whether to advance the rest of the polling or leave the slot unused.

To begin with, let's assume that if a slot is unused, AP simply moves ahead every subsequent poll. Fig. 2 shows the example in which the predecessors of S_s generate unused slots. Fig. 2(b) illustrates that the unused slots are reclaimed, CFP terminates earlier than scheduled to extend the CP for non-real-time message transmission. However, this method may deprive S_s of one access and the real-time guarantee can be broken. If we let $P_s = k \cdot F + \Delta$, where Δ is a value from 0 to F, then the least bound of network access within P_s is k or $k-1$, as noted in Eq. (2). (The figure shows the case of $k = 2$). If AP simply advances H_s, S_s loses one scheduled access as shown in Fig. 2(b) provided that the new arrival of message falls in between new and original polling instants. As contrast, that access can be preserved if the AP does not reclaim the unused bandwidth as shown in Fig. 2(a).

The main idea of proposed scheme is that the rest of polling schedule can be advanced if all the subsequent streams have their messages to send, that is, if none of them are waiting for a new message arrival. As the PC can finish the polling schedule of that round earlier than the original CFP duration, CP can be extended to transmit more non-real-time messages. In addition, as the AP receives all the informations on period and transmission time before bandwidth allocation, it can estimate the status of each stream, namely, whether its transmission buffer is empty or not[3]. Finally, in case the slot cannot be reclaimed, it can be used for error control of that stream.

4.2 Runtime Operation

Polling order is important not only in deciding whether a stream will be affected by a deferred beacon but also in improving the probability of reclaim. The more a stream generates unused slots, it would be better to put the stream in the latter place, as small number of successor increases the probability of being reclaimed. How much a stream generates unused slot depends on the amount of overallocated bandwidth. This amount consists of static and dynamic factors. The static factor does not change during the whole life time of a stream, calculated by subtracting the actual bandwidth requirement, $\frac{C_i}{P_i}$, from allocated bandwidth, $\frac{H_i}{F}$. On the other hand, the dynamic factor keeps changing period by period, as it is decided by the current transmission rate. As a result, the overallocation, O_i, is calculated as in Eq. (4).

$$O_i = (\frac{H_i}{F} - \frac{\bar{C}_i}{P_i}) + max\{0, (1 - \frac{A_i}{R_i}) \cdot M_i\} \tag{4}$$

where A_i is the actual transmission rate S_i is now experiencing for this period. With this information, the polling order should be decided such that the larger O_i, the latter the stream is polled. The order is rearranged for each beginning of a superframe, taking into account the current transmission rates of respective streams.

Finally, the stream which has higher error rate brings more unused slots, so it seems better to place such a stream to the latter part. However, the error dynamics, conforming to Guilbert model, are so unpredictable that the average behavior cannot provide meaningful criteria[1]. If we are to consider the error characteristics, the channel probing mechanism should be reinforced to the reclaim scheme.

5 Performance Measurements

This section measures the performance of the proposed reclaim scheme via simulation using NS-2 event scheduler[9]. The experiments focus on measuring the achievable throughput to demonstrate the effectiveness of reclaiming scheme. We define *achievable throughput* as the virtual throughput for a given stream set without any collision even in CP. This can be estimated as the sum of both utilization of real-time message streams and ratio of average length of CP to F.

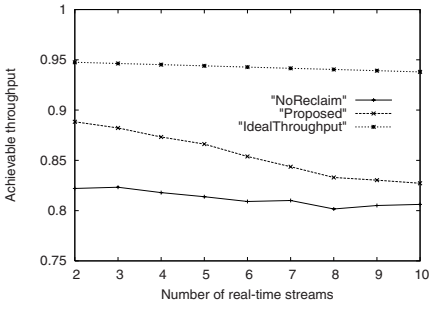


Fig. 3. Bandwidth reclaiming

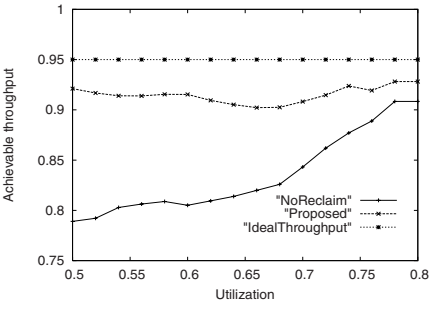


Fig. 4. Bandwidth reclaiming

Fig. 3 plots achievable bandwidth according to the average number of streams on the superframe to evaluate the performance of reclaiming scheme. Without overallocation caused by the hard real-time guarantee, the only waste is polling overhead, but overallocation makes the throughput much less than ideal. However, the resource reclaiming scheme can narrow the gap between those two curves, that is, considerably relieves the problem of poor utilization of PCF operation, as shown in Fig. 4. The amount of overallocation does not depend on the number of streams but how much F is harmonic with each P_i. For the experiment, 200 stream sets are generated for each number of streams ranging from 2 to 20 with utilization between 0.64 and 0.65. At last, it is certain that the improvement increases when the number of streams is small, and the 52.3 % of waste was recovered. As shown in the figure, the improvement gets smaller as the number of streams increases. This is due to the fact that the reclaimed portion gets smaller and the probability of reclaim decreases.

Fig. 4 plots the reclaimed throughput measured by changing the utilization of stream set from 0.5 to 0.8, while the number of streams randomly distributes

from 2 to 10. If only a stream set has a feasible schedule, the throughput goes high as utilization increases. On the contrary, reclaimed scheme provides stable throughput throughout the given utilization range. When the utilization is from 0.5 to 0.65, about 31.3 % of bandwidth waste was reclaimed.

6 Conclusion and Future Work

In this paper, we have proposed and analyzed a bandwidth reclaim scheme that can overcome poor utilization problem of PCF for real-time communication in WLAN. When an unused slot occurs, AP tests whether it can be reclaimed by checking all of its successors have messages to send. This test confirms that the other time-sensitive traffics are not affected by the early termination of polling round. The reclaimed bandwidth is reassigned to CP to improve the response time of connection management, error control, and other non-real-time messages. The simulation results show that the proposed scheme is able to reclaim up to 52.3 % of bandwidth waste when the number of streams is 2 and that it also provides stable throughput throughout the utilization from 0.5 to 0.8.

Finally, we are to apply the bandwidth reclaim scheme proposed in this paper to the EDF style polling framework. In addition, we are also planning to develop a bandwidth reclaim scheme combined with an error control mechanism.

References

1. Choi, S., Shin, K.: A unified wireless LAN architecture for real-time and non-real-time communication services. IEEE/ACM Trans. on Networking (2000) 44-59
2. IEEE 802.11-1999: Part 11 - Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications (1999) also available at *http://standards.ieee.org/getieee802*
3. Caccamo, M., Zhang, L., Sha, L., Buttazzo, G.: An implicit prioritized access protocol for wireless sensor networks. Proc. IEEE Real-Time Systems Symposium (2002)
4. Sheu, S., Sheu, T.: A bandwidth allocation/sharing/extension protocol for multimedia over IEEE 802.11 ad hoc wireless LANS. IEEE Journal on Selected Areas in Communications, Vol. 19 (2001) 2065-2080
5. Khattab, A. and Elsayed, K.: Channel-quality dependent earliest deadline due fair scheduling schemes for wireless multimedia networks. Proc. of MSWIM. (2004) 31-38
6. Adamou, M., Khanna, S., Lee, I., Shin, I., Zhou, S.: Fair real-time traffic scheduling over a wireless LAN. Proc. IEEE Real-Time Systems Symposium (2001) 279-288
7. Lee, J., Kang, M., Jin, Y., Kim, H., Kim, J.: An efficient bandwidth management scheme for a hard real-time fuzzy control system based on the wireless LAN. Lecture Notes in Artificial Intelligence, Vol. 3642. Springer-Verlag, Berlin Heidelberg New York (2005) 644-659
8. Carley, T., Ba, M., Barua, R., Stewart, D.: Contention-free periodic message scheduler medium access control in wireless sensor/actuator networks. Proc. IEEE Real-Time Systems Symposium (2003) 298-307
9. Fall, K., Varadhan, K.: Ns notes and documentation. Technical Report. VINT project. UC-Berkeley and LBNL (1997)

A Fast Real Time Link Adaptation Scheme for Wireless Communication Systems*

Hyukjun Oh[1], Jiman Hong[2], and Yongseok Kim[3,**]

[1] Kwangwoon University, Seoul, Korea
hj_oh@kw.ac.kr
[2] Soongsil University, Seoul, Korea
jiman@ssu.ac.kr
[3] Samsung Electronics, Suwon, Korea
yongseokkimasi@samsung.com

Abstract. In this paper, a fast real time link adaptation scheme for wireless communication systems is proposed. The proposed scheme is employing multi-stage adaptation controls based on channel state information (CSI). The optimal link adaptation scheme is known to have high complexity to implement, and it shoud be run iteratively to deal with time varying wireless channels in a short time. The proposed method determines the speed of channel variation using CSI's, then it applies the most appropriate link adaptation methodology or algorithm to the current time varying rate of channels. For examples, a simple up/down transmit power adjustment is used for the very fast time varying channel condition. On the other hand, the optimal link adaptation scheme is directly applied in the very slow varying channel state. The proposed multi-stage link adaptation scheme can be easily implemented in real time wirelss communication systems because of its capability of selecting an appropriate real time link adaptation scheme adaptively to each channel variation rate. The design example on the selected platform shows that the proposed scheme is very efficient in real time applications of the link adaptation in wireless communications, while its performance is maintained close to the optimal.

1 Introduction

Efficient utilization and allcation of the resources over the propagation channel is certainly one of the major challenges in wireless communication system design [1]. Especially, a modern day wireless communications system is required to operate over channels that experience fading and multipath. Given the ever-growing demand for wireless communication, a higher efficiency, higher performance wireless communications system is desirable. In order to improve the efficiency and decrease the complexity of the system, CSI's can be transmitted

* This research work has been supported by Seoul R&BD Program in 2007.
** Corresponding author.

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 933–940, 2007.
© Springer-Verlag Berlin Heidelberg 2007

back to the transmitter unit in order to precondition the signal before transmission. These preconditioning or transmit coordination works are called *link adaptations*.

One such communications system is an orthogonal frequency division multiplex (OFDM) system with CSI feedbacks. Its multicarrier nature with CSI feedbacks allows the use of link adaptation to enhance its performance significantly [1], [2]. The idea of link adaptation in OFDM is as follows. Based on the channel characteristic given by CSI's, the number of bits to be transmitted, the level of modulation and the transmission power in each sub-carrier are selected in order to increase the transmission bit rate [3], [4] or reduce the required transmit power [5].

For example, in a standard TDMA system employing OFDM with link adaptation, each user is allocated a fixed time slot and only that user can transmit in that time slot. These unused sub-carriers, as a result of link adaptation, are wasted and cannot be used by other users during that time slot. However, the sub-carriers that are in deep fade over the link between the base station (BS) and the designated mobile user may not be in deep fade over the links between the BS and other mobile users. This motivates us to consider the link adapation schemes that users share the downlink transmission by adaptively using different sub-carriers, instead of using different time slots as in a TDMA system. This approach will allow all the sub-carriers to be used effectively. However, its algorithm is not simple to implement like one proposed in [6] that is not suitable for the time varying channel as the required computation makes real-time implementation impractical.

The optimal link adaptation strategy is to perform link adaptation schemes like power allocation, sub-carrier allocation, bit allocation, user allocation, modulation and coding allocation every time the propagation channel condition is changed. Unfortunately, however, it is quite difficult to perform the optimal link adaptation strategy in the fast time varying channel condition, because its complexity is pretty high to be run in real time.

In this work, we propose a link adaptation scheme that is simple to implement. Link adaptation strategy is developed which results in minimum transmit power. Also real-time heuristic allocation algorithm which is close to the optimum while having much fewer computations is also proposed. The proposed scheme is employing multi-stage adaptation controls as per channel variation rate based on channel state information (CSI). The proposed method determines the speed of channel variation using CSI's, then it applies the most appropriate link adaptation methodology or algorithm to the current time varying rate of channels. For examples, a simple up/down transmit power adjustment is used for the very fast time varying channel condition. On the other hand, the optimal link adaptation scheme is directly applied in the very slow varying channel state. The proposed multi-stage link adaptation scheme can be easily implemented in real time wirelss communication systems because of its capability of selecting an appropriate real time link adaptation scheme adaptively to each channel variation rate. We compare the performance and complexity of these algorithms with existing ones to

show the effectiveness and efficiency of the proposed scheme. Such design examples on the selected platform shows that the proposed scheme is very efficient in real time applications of the link adaptation in wireless communications, while its performance is maintained close to the optimal.

2 System Model

The structure of a multiuser OFDM system using link adaptation is shown in Fig. 1.

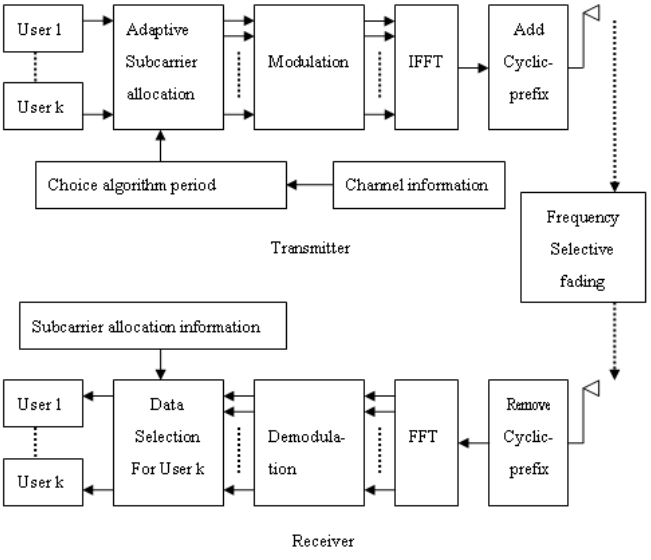


Fig. 1. System model of OFDM system with link adaptation

Let K be the number of users supported by the system and N be the number of sub-carriers. Also, let S_k denote the number of sub-carriers required by user k. Assume each sub-carrier has a bandwidth that is much smaller than the coherence bandwidth of the channel and the instantaneous channel gain of each sub-carrier for every user is known to the base station. In the transmitter side, the link adaptation algorithm assigns sub-carriers to each user according to the channel gains. The bit stream of each user is then transmitted out using the assigned sub-carriers. In each sub-carrier, the assigned bits are modulated with a power level that can overcome the fading of the channel. In addition to the power required to transmit the bits in the AWGN channel, power proportional to the inverse of the corresponding sub-carrier channel gain a_k," is needed. The modulated signals of all the sub-carriers are then transformed into time domain using IFFT and cyclic prefix is inserted to eliminate the ISI. Dedicated sub-carriers are reserved for the broadcasting of the sub-carrier allocation or link adaptation

information of which each receiver can retrieve to determine the corresponding sub-carriers for receiving the data. In the receiver side, after removing the guard interval and transforming the time domain samples into the frequency domain symbol, the user demodulates the sub-carriers prescribed by the link adaptation of allocation information to recover the original bit stream. The goal of the link adaptation is to minimize the total transmit power of all the users while satisfying the data transmission requirement for each user. Given P is the transmit power of a sub-carrier in the AWGN channel and $p_{k,n}$ is a binary assignment variable which is equal to 1 if sub-carrier n is assigned to user k or equal to 0 otherwise, the total transmission power is given by

$$P_T = \min \sum_{n=1}^{N} \sum_{k=1}^{K} p_{k,n} \frac{P}{\alpha_{k,n}^2} \tag{1}$$

The objective is to find the values of the assignment variables, $p_{k,n}$, to minimize p_T while satisfying the following constraints,

$$\sum_{n=1}^{N} p_{k,n} = S_k, \qquad k \in \{1, 2, ..., K\} \tag{2}$$

$$\sum_{n=1}^{N} p_{k,n} = 1, \qquad n \in \{1, 2, ..., N\}. \tag{3}$$

Constraint (2) specifies that the total number of sub-carriers allocated to user k is S_k and constraint (3) specifies that each sub-carrier can only be allocated to one user.

3 Optimal Link Adaptation

The above optimization problem is similar to the classical assignment problem in the area of linear programming. Therefore, we can solve this problem by first transforming it into an equivalent assignment problem. Here each user is expanded to S_k sub-users and for every sub-user, only one sub-carrier will be allocated. The channel gain of the sub-user is equal to that of the original user. The optimization problem then becomes Eq. (1) with $K = N$. The constraints (2) are also transformed to $\sum_{n=1}^{N} p_{k,n} = 1$ with $K = N$.

Numerous methods have been proposed to solve this classical assignment problem. One of them is the famous Hungarian method proposed in [7] of which the computation complexity of the algorithm is $O(n^4)$. In a time varying channel, the channel characteristic of users changes frequently. To cope with this situation, we need a link adaptation algorithm like a sub-carrier allocation above that is fast enough to allocate the subcarrier within the coherence time of the channel. Unless it is fast enough, the actual link adaptation operation cannot be run in real time while keeping achieving the required performance. The optimal

link adaptation strategy is to run the optimal link adaptation algorithm like the sub-carrier allocation given in above continuously as per channel variations. Ideally, the link adaptation should be done before the channel varies. It is quite difficult goal to achieve and it is even impossible for fast varying channels with high speed mobiles. That is, the complexity of the above optimal algorithm is too large that an optimal solution may not be generated within the coherence time, especially for a fast-fading channel. Therefore, we propose a heuristic link adaptation scheme that satisfies the real time requirement.

4 Proposed Multi-stage Link Adaptation Strategy

The optimal link adaptation algorithm of sub-carrier and transmit power allocations addressed in the previous section is not appropriate for the real time operation due to its high complexity and duty cycle limitation in fast varying channel condition. In this section, we propose using CSI's to determine the current channel variation rate and propose to use the computationally efficient link adaptation alogorithm optimized for the estimated speed of channel variations.

The channel variation rate is not required to be very accurate because the characteristics of the propogation channel is not changing rapidly as per the channel variation rate. Its characteristics are more dependent on the trend of channel variation speed rather than the exact rate or speed of the variations. In this paper, we propose to classfy propagation channel conditions into three categories as per the estimated channel variation rate: slow, medium, and high speed. These three categories reflect the trend of channel variation. Slow varying channel means that noticeable changes in their characteristics are observed after relatively long time is elapsed. High speed categoty means that significant changes in their channel properties are observed in very short time. That is, the acutal rate of channel variations is not important. The more relevent information to reflect the changes in their characteristics is how much they have been chaged actually. In this paper, we propose to use such relevent information based on CSI's intead of estimating the rate of channel variations directly. In other words, the trend of channel variations are divided to three groups of microscopic, mediumscopic, and macroscopic channel changes.

The proposed multi-stage link adaptation strategy is very simple. First, three stages of link adaptations are set up: microscopic, medium scopic, and macroscopic channel changes. For each stage, there exists the link adaptation algorithm optimized for the characteristic of each stage. For example, the link adaptation algorithm of the first stage for microscopic change in channel should be simple enough to be run fast. A simple up/down power control would be a good candidate for the link adaptation algorithm of this stage. In Stage 3, the optimal link adaptation algorithm of allocatioins in Section 3 can be used because macroscopic changes in channel are infrequent usually. Therefore, periodic runnings of the optimal allocation algorithm should be sufficient for Stage 3. Because the required duty cycle is pretty long in this case, there is no problem to run the optimal algorithm in real time. On the other hand, Stage 2 must compensate

for the possible performance loss when the noticeable channel changes happen in a time shorter than Stage 3 running period. For this, the stage 2 kicks in and run the link adaptation algorithm optimized for this stage asynchronously if the accumulated channel variation from CSI's is larger than a threshold. Stage 1 keeps running during the call. On the other hand, the running period of Stage 3 is determined by the threshold Th. Fig. 2 shows the only simplified version of the flow chart of the proposed link adaptation scheme due to the space limitation.

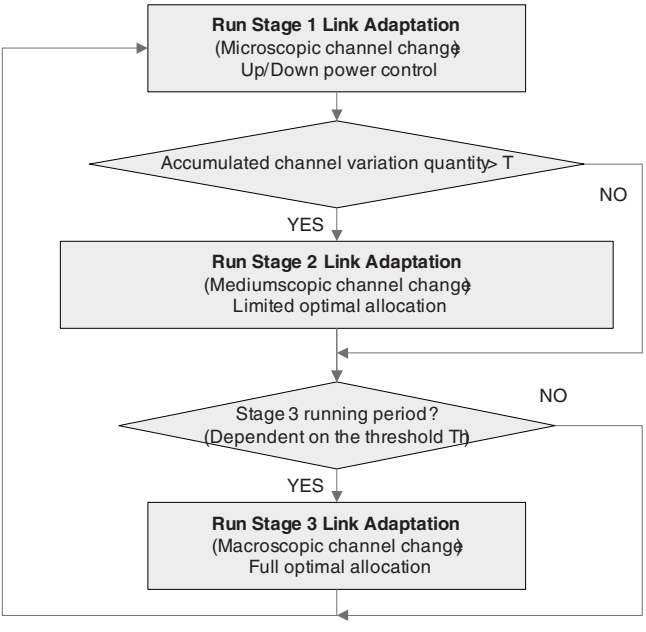


Fig. 2. The simplified flow chart of the proposed multi-stage link adaptation

5 Simulation Results

To show the effectiveness of the proposed multi-stage link adaptation scheme, the performance and complexity of the proposed method is compared with the optimal and sub-optimal ones in [6] and [8]. Here we consider an OFDM system with several multi-stage switching rates and thresholds. Several mobile speeds of 0km/h, 3km/h, 30km/h, 60km/h, 90km/h, and 120km/h are considered to simulate the various rates of channel variations. The basic unit of performing the link adaptation is multiple of frames.

The performance of the proposed scheme is shown in Fig. 3 together with the optimal one that is performing the link adaptation every frame. In practice, the optimal scheme is hardly finished in a frame, so that it cannot be used in the real time application. In addition, sub-optimal scheme of running the link adaptation over two frames is also shown in the same plot for the comparison purpose. Four

different link adaptation duty cyle threshold values of 1, 0.1, 0.01, and 0.001 are simulated to show the effectiveness of the proposed multi-stage link adaptation scheme. As the threshold value increases, the duty cycle of Stage 3 is increased so that the total number of full link adapation operations is reduced. It results in much less complexity and sufficient cycle time for the real time operation. Note that Stage 1 of simple power control for microscopic channel variations and Stage 2 link adaptation for mediumscopic channel variations are running asynchronously regardless of the threshold values for Stage 3 duty cycle. Table 1 summarizes the total number of the full link adaptation operations to compare the complexity of the proposed scheme with the optimal one.

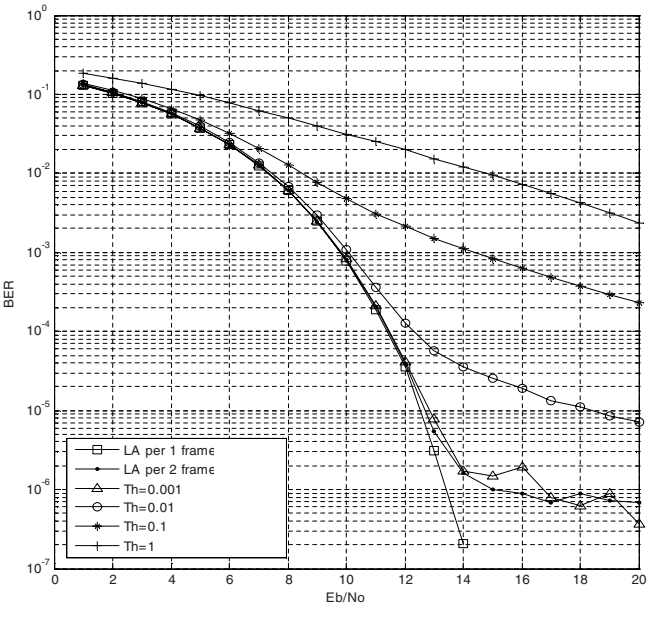


Fig. 3. The performance of the proposed method

Fig. 3 shows that the performance of the proposed multi-stage link adaptation scheme is close to the performance of the optimal method while considerable reduction in the total number of full link adaption operations is achieved as shown in Table 1.

It is obvious that the proposed multi-stage link adaptation scheme can provide the real time operations in practical wirelss communication systems even without noticeable performance degradation. The performance gap between the proposed method and the optimal one will becomes noticeable eventually when the rate of channel variation is very high. In this case, however, correct CSI's cannot be available at BS side at the right timing because of CSI feedback loop delay. Even optimal method also suffers from it.

Table 1. Simulation Parameters

Speed	Duty cycle in frames	
	$Th = 0.01$	$Th = 0.001$
3km/h	132	42
30km/h	22	7
60km/h	14	5
90km/h	10	3
120km/h	8	2

6 Conclusion

In this paper, we considered the real time link adaptation strategy for downlink OFDM transmission. To satisfy the time variation of fading channel, a real-time heuristic multi-stage link adaptation strategy based on the determined channel variation speed using available CSI's was proposed. Switchig between optimized link adaptation stages were controlled adaptively. Simulation results show that the proposed multi-stage link adaptation scheme provides a solution that is considerably simpler to implement than the optimal method, and that is appropriate for the real time applications, while its performance is very close to the optimal solution.

References

1. Rohling, H., Grunheid, R.: Performance of an OFDM-TDMA Mobile communication system. Proc. IEEE Vehicular Technology Conference (1996) 1589-1593
2. Czylwik, R.: Adaptive OFDM for wideband radio channels. Proc. IEEE Globecom conference (1996) 713-718
3. Rhee, W., Cioffi, J.: Increase in capacity of multiuser OFDM system using dynamic subchannel allocation. Proc. IEEE Vehicular Technology Conference (2000) 1058-1089
4. Yin, H., Liu, H.: An Efficient Multiuser Loading Algorithm for OFDM-based Broadband Wireless Systems. Proc. IEEE Globecom conference (2000) 103-107
5. Chen, Y., Chen, J., Li P.: A fast suboptimal subcarrier, bit, and power allocation algorithm for multiuser OFDM-based systems. Proc. IEEE International Conference on Communications (2004) 3212-3216
6. Wong, C., Cheng, R., Letaief, K., Murch, R.: Mutliuser Sub-carrier Allocation for OFDM Transmission using Adaptive Modulation. Proc. IEEE Vehicular Technology Conference (1999) 479-483
7. Khun, H.: The Hungarian Method for the Assignment Problem. Naval Research Logistics Quarterly 2 (1955) 83-97
8. Wong, C., Tsui, C., Cheng, R., Letaief, K.: A real-time sub-carrier allocation scheme for multiple access downlink OFDM transmission. Proc. IEEE Vehicular Technology Conference (1999) 1124-1128

EAR: An Energy-Aware Block Reallocation Framework for Energy Efficiency*

Woo Hyun Ahn

Department of Computer Science, Kwangwoon University
447-1 Wolgye-Dong, Nowon-Gu, Seoul, Korea
whahn@cs.kw.ac.kr

Abstract. File systems, which are embedded in low-power computer systems, have continuously improved energy saving features. However, the inevitable aging of the file system gives disks less chances to stay idle, thus increasing the energy consumption along with the decrease in performance. As a solution to the problem, we propose the energy-aware block reallocation framework (EAR), which is a software framework with a reallocation algorithm applicable to file systems. EAR dynamically reallocates fragmented data with the same relationship to disk locations where less disk accesses and seeks are achieved. The optimized file data layout improves the energy saving on aged disks.

Keywords: Embedded systems, energy saving, file systems, disks, data layout.

1 Introduction

Today, disks are increasingly adapted as primary storage systems in low-power embedded systems such as mobile devices. However, disks consume more power than other computer components due to their mechanical movements. Earlier researches [2][6] achieve the energy saving by stretching idle periods for disks. Less disk I/Os and seeks have important roles on keeping disks in the idle state as well as increasing performance of file systems[4][9]. Especially, optimizing data layout of small files (< 64 KB), the major file usages[3][8], has a large impact on disk I/Os and seeks.

Earlier file systems, FFS[5][7], C-FFS[3], EEFS[4], and DFFS[1], have focused on how to transfer large data at one disk I/O since the multiple data transfers can increase the disk burst ratio, thus granting disks chances to stay idle and thus consume less power. FFS attempts to cluster logically sequential blocks of a file on physically contiguous disk blocks. Multiple block transfers can be used to access the clustered blocks, thus reducing disk I/Os and improving the energy saving. FFS divides the disk into *cylinder groups,* each of which is a set of consecutive cylinders. A cylinder group co-locates related data in a directory. C-FFS allocates small files in a directory to contiguous disk locations, where they form a group whose size is limited by 64 KB. It attempts to cluster a created file into an existing group associated with the same directory. When a file of a group is read, C-FFS reads all files in the group at one time.

* The present Research has been conducted by the Research Grant of Kwangwoon University in 2006.

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 941–948, 2007.
© Springer-Verlag Berlin Heidelberg 2007

This multiple file transfers reduce disk I/Os, improving the energy saving. Hence, C-FFS is suitable to low-power computers. EEFS divides disk into fix-sized clusters, each of which stores a set of files with locality. A group with a large set of files spans several contiguous clusters. When a file of a group is accessed, all files of the group are prefetched at one time. This multiple file accesses increase the disk burstiness, improving the energy saving.

On heavily aged disks, however, FFS cannot cluster blocks of a created file due to the shortage of free clusters and thus the efficiency of energy saving is reduced. This file fragmentation is defined as *InterA-file Fragmentation(IAF)*. Also, if once related files are placed across cylinders, their accesses persistently incur disk seeks because they can be no longer reallocated. In C-FFS, the file system aging splits existing groups into smaller groups, thus decreasing the number of files transferred at a single disk I/O. This separation is defined as *IntEr-file Fragmentation(IEF).* In EEFS, the file system aging incurs IEF in existing groups, draining free clusters needed to group related files. This causes the cleaning activity frequently executed to make rooms for free clusters. Unfortunately, the cleaning needs much CPU utilization as well as many disk accesses, increasing the energy consumption. Moreover, DFFS does not take the energy saving into consideration when reallocating fragmented data. Instead, it only focuses on improving small file fragmentation.

This paper proposes a new framework, called *Energy-Aware block Reallocation (EAR)* framework, which reallocates fragmented data in consideration of the energy saving on aged disks. EAR consists of the two components: first, reallocation algorithms that can be applied to the previous file systems for energy saving. The second is a filesystem-independent module that can be installed in operating systems. Related files with temporal locality are likely to be accessed and cached in memory together. The filesystem-independent module builds up the relation of the in-cache files in respective to relationship such as directory name and group. By using the relation information, EAR reallocates fragmented data that can be sources of large power consumption at contiguous disk locations. The improvement of file data layout increases a possibility that disk drives can keep idle for the longer periods.

2 Energy-Aware Block Reallocation Framework

EAR framework is a reallocation framework that is designed to reduce both IAF and IEF of small files (2 - 12 blocks). Related files are likely to be together in file cache due to temporal locality. The file cache exploits the write back that delays writes for some period of time, and periodically flushing the modified data to disk. Each time a dirty (or modified) block flushed to disk, EAR reallocates all fragmented data in a fix-sized *target region* of 64 KB, which starts from the disk location of the dirty block regardless of their modification, then eliminating IAF and IEF of the fragmented data.

The EAR mechanism, called *group write*, consists of the four phases: the decision of whether a target region is to be reallocated, the buffer gathering, the re-mapping, and disk write. EAR examines if a target region meets the three conditions. The first is if it has any data with either IAF or IEF, the second is if its data are in memory and the third is if it has two or more free blocks. If so, fragmented data of the target region

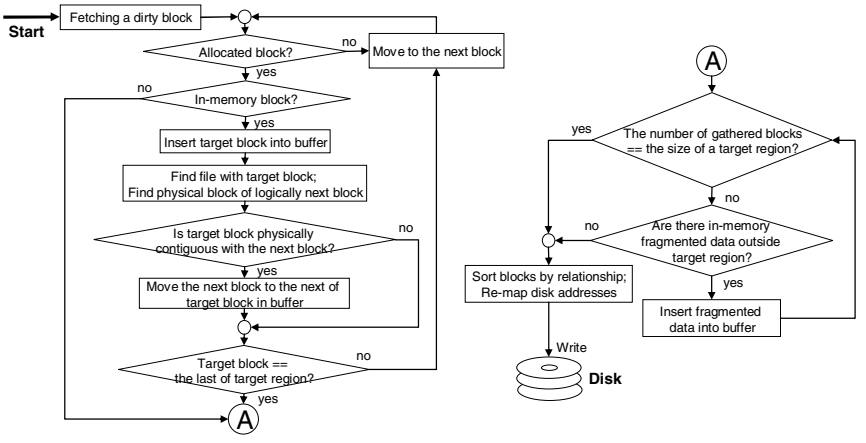


Fig. 1. EAR algorithm

are gathered into a buffer, which is called buffer gathering. Blocks addresses of the gathered data are re-mapped in order to eliminate their IAF and IEF. Finally, the re-mapped data are stored to the disk location indicated by the dirty block at one disk I/O.

Fig. 1 shows how blocks of a target region are gathered into a buffer and re-mapped during a group write. The process starts when a dirty block is stored to disk. The dirty block is first put at the head of the buffer. From the next of the dirty block to the last of the target region, EAR algorithm checks one by one if each block is a good candidate that meets conditions described below for a group write.

Any blocks of a target region may not be in memory under a buffer gathering. Two approaches can be taken to handle the uncached blocks; the first is to fetch the blocks into memory by using additional disk I/Os, but increases the time spent to do a group write. To avoid the cost, EAR algorithm uses the second approach in which the buffer gathering process stops when any uncached block is encountered. Then the reallocation is applied to the only gathered blocks because the uncached blocks can be lost on disk if some blocks being reallocated are overwritten to the disk locations of the blocks that were not fetched. After the stop, if free space still exits between the dirty block and the uncached block, it is used to reallocate fragmented data again.

EAR algorithm first checks if each disk block address is not only allocated, but also in memory since it reallocates the in-memory allocated block, called *target block*, to avoid additional disk I/Os, which is inserted to the tail of the buffer. EAR algorithm finds the file with the target block, then checking whether the logically next of the target block has physically contiguous disk address in reference to the target block. If not, the logically next block whose disk block may be outside of the target region is moved into the next of the target block. Then EAR algorithm re-assigns the two blocks newly physical block addresses in order to make them physically contiguous with each other. This process is progressed until the last block address of the target region is encountered. The process concentrates on eliminating IAF of files within the target region and thus decreasing disk I/Os needed to access fragmented files.

Some files gathered in a buffer may have the different relationships with the file with the dirty block. The mixture of the files with different relationships on disk

decreases the advantages of multiple data transfers. To solve the problem, EAR reallocates files irrelevant to the dirty block to the back of a target region, whereas it adjacently places files with the same relationship at the front.

There may be still some free blocks in a target region after the buffer gathering. By using the free blocks, EAR attempts to gather and re-map another fragmented data that have the same relationship, but are outside of the target region. The free blocks are preferentially filled up with a file with blocks that exist the inside and the outside of a target region. The purpose of this is to eliminate IAF of the files, which is one of sources that cause IEF of related files. In addition, if there are still free blocks in the target region, EAR algorithm scrapes up IEF-files that are outside of the target region into a larger group at the unit of relationship such as explicit group or directory.

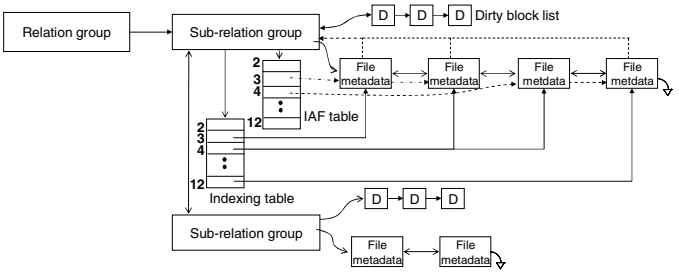


Fig. 2. A relation group and its sub-relation groups

In Fig. 2, a *relation group(RG)* is implemented in memory to make the relationship of in-memory files with temporal locality at the relationship unit. It represents a logical set of files that are contained not only in the same relationship, but also in the file cache. When a block is accessed, the block and its file metadata are registered at the RG associated with its relationship. Examples of block and file metadata are a pointer to a buffer holding the block and the in-memory i-node respectively. The metadata are linked in the RG to make the relation of related files.

A set of consecutive disk locations where related files are usually co-located is defined as *disk region,* into which the file systems physically divide the disk. A cylinder of FFS, a group of C-FFS, and a cluster of EFFS are the examples of the disk region. The summary of a disk region is implemented as in-memory metadata, called *sub-relation group(SRG)* for the fast lookup of a disk region with files that are to be reallocated. The summary includes the information of a disk region: the list of in-memory files out of files in the disk region, the total number of the in-memory files, the degree of IEF and IAF, etc. Its usage is explained in Section 3.

A SRG has two tables, *IAF table* and *indexing table*, to manage its in-memory files of 2 – 12 blocks. The IAF table is used to sort IAF-files of the SRG by sizes. A table entry indicated by an index points to a list of IAF-files with the same number of blocks as the index number. The IAF table is looked up to select fragmented files with specific sizes under the buffer gathering. The indexing table sorts files of the associated SRG by sizes regardless of their IAF states. A table entry indicated by an index links files with the same number of blocks as the index number. The indexing table is used to look up an in-memory file with a specific size in a disk region.

3 File Grouping Methodologies

EAR determines which fragmented data outside a target region are to be grouped into free space of the target region during a group write. The file grouping policies have different impacts on the efficiency of energy saving according to file system types.

In FFS, the effectiveness of multiple block transfers depends on how many blocks of a file are contiguously placed on disk. For the higher effectiveness, EAR first uses free blocks of a target region to reallocate as much of large files out of small files with IAF outside the target region. The files may be on either a cylinder with the target region or cylinders far away the target region. Out of the cylinders, EAR preferentially selects the small files of a cylinder that is at the longest distance from the current cylinder with the target region to reduce disk seek distances. This scheme reduces disk accesses and seeks that can be incurred by accessing the related files.

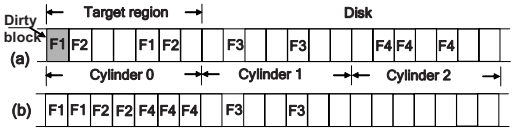


Fig. 3. File grouping of FFS. A file Fx is composed of blocks marked by Fx.

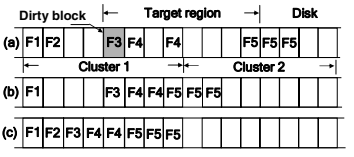


Fig. 4. File grouping of EEFS

A SRG manages metadata of related files on a cylinder in FFS while a RG does all in-memory files in a directory whose files are usually stored over several cylinders in Fig. 3(a). For free blocks of a target region, all SRGs of the RG are examined to find a cylinder that is at the longest distance from the current cylinder among cylinders with IAF-files like Fig. 3(b). For the found SRG, its IAF table is looked up to find an IAF-file of blocks that can be held in the free blocks. If there are not any files of the size, EAR again looks up the table entry indicated by the value subtracted by one from the current index number since EAR tries to scrape up related files scattered across outer cylinders into inner cylinders. This is repeated until the free space is filled up.

In EFFS, a group that is associated with a RG presents a set of files with locality while a cluster of a group is disk region associated with a SRG. When there are free blocks in a target region, EAR examines all SRGs of the group with the dirty block that triggers the group write. Among clusters with IEF, EAR selects a cluster with the largest amount of files, then reallocating its data to the free blocks. The cluster state is changed to empty so that EFFS can allocate new files. This sequence of operations is a cleaning activity achieved by a group write. Since clusters of large data becomes clean during group writes, EEFS does but clean clusters with small data which it takes less time to clean. Hence, EAR lifts a burden from the cluster cleaning activity.

The starting location of a target region should be aligned to the starting block address of a cluster with the dirty block, instead of the disk location of the dirty block. Moreover, the size of a target region is defined as a multiple size of a fix-sized cluster. In Fig. 4(a), the disk address of a dirty block may be at the center of a cluster instead of the front. If the related files that are gathered into a buffer are stored to the

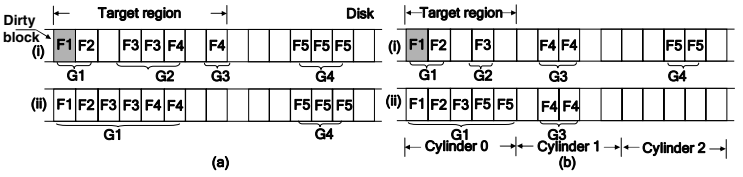


Fig. 5. File grouping of C-FFS

starting location of the dirty block, they will be still placed across the two clusters, suffering from another IEF in Fig. 4(b). To solve this problem, the starting block of a target region is aligned to that of a cluster in Fig. 4(c).

In C-FFS, a directory name space is associated to a RG like FFS, whereas a group is considered to a disk region and thus associated with a SRG. In Fig. 5(a)(i), the file system aging may make related files of a directory split into small groups. When a group write is executed like Fig. 5(a)(ii), EAR finds SRGs corresponded to one or more groups with enough many files that can be migrated into the free blocks, then reallocating their files to the free blocks. However, if there are some free blocks that can only include one part of a group (i.e., G4), the group are not reallocated into the free space because the group is again split as separate group after the reallocation.

Small groups in a directory can be put on several cylinders like Fig. 5(b)(i). Accessing files of the groups can incur many disk seeks. To reduce disk seeks, EAR selects groups in a cylinder that is at the longest distance from the current cylinder with the target region, reallocating their files into the free space like Fig. 5(b)(ii). It is because clustering the groups of cylinders that are at the long distance from each other can minimize the distance of disk seeks that are caused by group accesses.

4 Experimental Evaluation

We used a low-cost hardware computer with a 700 MHz Pentium processor and 128 MB of memory to perform all experiments in the environment similar to embedded systems. EAR framework was developed as a module in the OpenBSD kernel version 2.8. Since it takes a long time to newly implement EEFS and C-FFS in the kernel, FFS implemented in the OpenBSD is used for our experiments. More parameters are shown in Table 1. A benchmark is made to recreate the non-adjacency of on-disk placements shown in the previous research[8]. On an empty disk, the benchmark executes sequences of file operations determined by the probability distribution in Table 2. When the disk became extremely full (disk utilization of 75%), the ratio of

Table 1. Testing system configuration

Disk		File System	
Disk space	6.4 GB	Size	2 GB
Rotation speed	5400rpm	Block size	4 KB
Cylinder	13328	Cylinder group	283
Average seek	9.5 ms		

Table 2. Ratio of file requests

Requests	Before 75%	After 75%
Create	10%	10%
Delete	5%	10%
Read	65%	60%
Write	20%	20%

create/delete is changed to make the disk heavily fragmented. File sizes of the requests are determined by the distribution observed on our file server in Fig. 6. For directory locality, the benchmark makes 50 sub-directories, each of which is selected by a Poisson distribution, then selecting a random number within the range of 100-5000 for the number of file requests that will be performed in a sub-directory.



Fig. 6. Distribution of file sizes

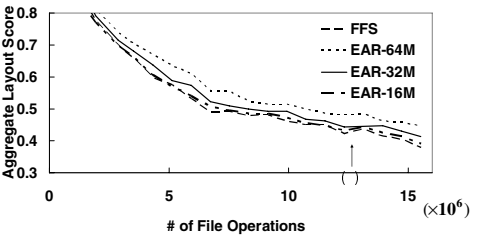
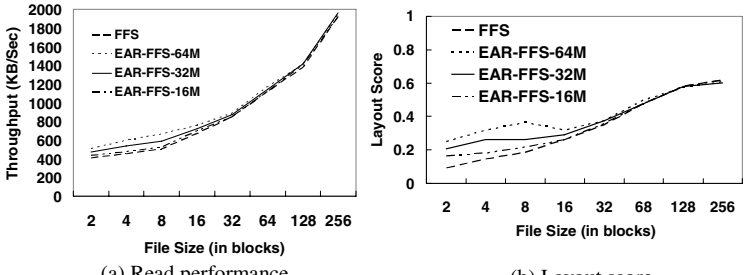


Fig. 7. Aggregate file data layouts

We used a *layout score*[8] to compare the efficiency of energy saving among file systems. Earlier researches[4][9] show that optimizing file data layout can improve the energy saving. File data layouts can be numerically presented as the layout score, which quantifies the degree of contiguous allocation of files. The layout score for an individual file is the ratio of the block physically contiguous with the previous block of the same file. For a file with a layout score of 1.00, all of its blocks are allocated contiguously. A file with a layout score of 0.00 has no contiguously allocated blocks. The aggregate layout score presents the average score of all files on one file system.



(a) Read performance

(b) Layout score

Fig. 8. Read performance and layout score

Fig. 7 shows the aggregate layout scores of EAR-FFS and FFS according to the number of file operations and file cache sizes. EAR-FFS represents FFS that cooperates with EAR framework, which uses 16 blocks as the size of a target region. EAR-FFS-64, 32 and 16 use 64 MB, 32 MB and 16 MB as the file cache size respectively. As file operations are executed for a long period of time, the difference in the layout score among the file systems increases. All EAR-FFSes outperform FFS in the layout score. Especially, the layout score of EAR-FFS-64 is larger than that of FFS by 14%.

The improvement in the file data layouts seriously decreases disk I/Os caused by related file accesses, giving disks more chances to keep in the idle state.

Fig. 8(a) and 8(b) respectively show the read performance of small files and the layout scores at the point (1) in Fig. 7, where the disks become heavily aged. For the measurement, we read files in each of the 50 directories sequentially. The read performance on disk that was aged by EAR-FFS-64 is much higher than that of FFS by 30%. The results allow us to make two points. First, disk I/Os are significantly reduced by the increase of multiple block transfers, which results from the improvement of EAR-FFS-64 in the layout score over FFS by 94%. Second, disk seeks go down because related files spread over long distant cylinders are reallocated into adjacent cylinders with free space that was made by the file system aging.

In Fig. 7 and 8, small file caches weaken the effectiveness of group writes. The decrease of the file cache sizes reduces the number of blocks that will be cached in memory, diminishing the number of blocks which a group write can gather into buffers for reallocation. The decrease of blocks that are to be reallocated gives EAR less chances to optimize the file data layout.

5 Conclusions

EAR optimizes file data layouts by using locality. The improvement of the file data layouts increases the efficiency of the energy saving for related file accesses. The measurements show that EAR improves the aggregate layout score, one of criteria for the efficiency of the energy saving as well as the read performance of small files.

References

1. Ahn, W.H., Park, D.: Mitigating Data Fragmentation for Small File Accesses. IEICE transactions on information and systems, Vol. E86-D, No. 6 (2003) 1126-1133
2. Helmbold, D.P., Long, D.D.E., Sherrod, B.: A Dynamic Disk Spin-down Technique for Mobile Computing. In Proceedings of the 2nd ACM MOBICOM, Rye NY (1996)
3. Ganger, G.R., Kasshoek, M.F.: Embedded Inodes and Explicit Grouping: Exploiting Disk Bandwidth for Small Files. In Proceedings of the USENIX Technical Conference (1997)
4. Li, D., Wang, J.: A Performance-oriented Energy Efficient File System. In Proceedings of International Workshop on Storage Network Architecture and Parallel I/Os (2004)
5. Mckusick, M., William, N.J., Leffler, S.J., Fabry, R.S.: A Fast File System for UNIX. ACM Transactions on Computer Systems, Vol. 2, Issues 3 (1983) 181- 197
6. Papathanasiou, A.E., Scott, M.L.: Energy Efficient Prefetching and Caching. In Proceedings of the USENIX Technical Conference (2004)
7. Seltzer, M., Smith, K.A., et. al.: File System Logging versus Clustering: a Performance Comparison. In Proceedings of the USENIX Technical Conference (1995)
8. Smith, K.A., Seltzer, M.: File System Aging – Increasing the Relevance of File System Benchmarks. In Proceedings of the 1997 ACM SIGMETRICS, Seattle WA (1997)
9. Zheng, F., et. al.: Considering the Energy Consumption of Mobile Storage Alternatives. In Proceedings of the 11th IEEE/ACM MASCOTS, Orlando Florida (2003)

Virtual Development Environment Based on SystemC for Embedded Systems*

Sang-Young Cho, Yoojin Chung, and Jung-Bae Lee

[1] Computer Science & Information Communications Engineering Division
Hankuk University of Foreign Studies, Yongin, Kyeonggi, Korea
{sycho,chungyj}@hufs.ac.kr
[2] Computer Information Department
Sunmoon University, Asan, Chungnam, Korea
jblee@sunmoon.ac.kr

Abstract. Virtual development environment increases efficiency of embedded system development because it enables developers to develop, execute, and verify an embedded system without real hardware. We implemented a virtual development environment that is based on SystemC, a system modeling language. This environment was implemented by linking the AxD debugger with a SystemC-based hardware simulation environment through the RDI interface. We minimized modification of SystemC simulation engine so that the environment can be easily changed or extended with various SystemC models. Also, by using RDI, any debugging controller that support RDI can be used to develop an embedded software on the simulation environment. We executed example applications on the developed environment to verify operations of our implemented models and debugging functions. Our environment targets in ARM cores that are widely used in commercial business.

Keywords: Virtual development environment, SystemC, Embedded system development, Remote debug interface, Hardware simulation.

1 Introduction

The biggest challenge of bringing an embedded system solution to market is delivering the solution on time and with complete functionality required because market is highly competitive and demands of consumers rapidly change. Most embedded systems are made up of a variety of Intellectual Property (IP) including hardware IP's (processors and peripherals) as well as software IP's (operating systems, device drivers and middlewares, like a protocol stack).

Virtual Development Environment(VDE) is an embedded system development environment that can verify a hardware prototype, develop a software without a real hardware, or be used for co-design of hardware and software.

* This research was supported by the MIC Korea under the ITRC support program supervised by the IITA.

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 949–956, 2007.
© Springer-Verlag Berlin Heidelberg 2007

This environment usually provides a hardware simulation model, simulation engine, and other tools that are useful in software development, and thus increases efficiency of embedded system development [1,2,3].

Virtual Platform[1] is Virtio's commercial virtual development environment. It supports many different types of processors such as ARM, X-Scale, and MIPS. Virtual Platform supports to develop software for a target system. MaxSim[2] is comprehensive SoC development environment provided by ARM. It consists of fast and easy modeling and simulation, and tools that provide debugging. Additionally, it enables system and hardware developers to compose the most suitable architecture quickly and accurately, and software developers to develop software before actual hardware comes out by providing VDE. Virtual ESC[3], made by Summit, is a set of 'Fast ISS' models and 'platform-based packages', which are composed of TLM-based busses, memory, and peripheral device models. According to its setting, it can comprise and run many different abstract-level systems, and debug related software and firmware. Also, it can create virtual prototypes for fast software development. The commercial VDE's described above integrate hardware simulation tools and software development tools tightly. Therefore it it limited that an embedded system developer uses various software tools and hardware simulation tools from different companies or organizations flexibly.

In this paper, we describe our design and implementation of a SystemC-based virtual development environment for developing embedded systems. We built virtual target hardware environment that is composed of various SystemC hardware models, and software development environment. We implemented the virtual hardware environment by implementing ARM processor core, its memory model, and other hardware IPs using SystemC. Also, we linked the hardware simulation environment with AxD(ARM eXtended Debugger), a debug controller provided by ARM's software development environment, through RDI(Remote Debug Interface). This system can be run with any debugger that implements RDI, and applied to any SystemC-modeled virtual hardware system.

The rest of our paper is organized as follows. In section 2, we describe related studies that are necessary for developing our environment. In section 3, we explain design and implementation of our virtual development environment for embedded system development. Finally, section 4 ends our paper with conclusion.

2 Related Studies

2.1 SystemC

SystemC is a system modeling language that can model and operate hardware at system-level. SystemC can easily express a complex SoC core at a high level while having all the merits of hardware description languages. It was developed using C++ classes. Hence, SystemC can be effectively used for simulation environment in checking not only hardware operation but also software one. Also, it supports TLM(Transaction-Level Modeling)[4,5].

We implemented the virtual development environment using SystemC version 2.0.1. Version 2.0.1 provides not only register-level transmission modeling, but also algorithm-and-function-level modeling. SystemC class libraries provide essential classes for modeling system structure. They supports hardware timing, concurrency, and reaction, which are not included in standard C++. SystemC allows developers to describe hardware and software, and their interface under C++ environment. Main parts of SystemC are as follows.

- Module: A container class that can include other modules or processes.
- Process: Used for modeling functionality and defined within a module.
- Ports and Signals: Signals connect modules through ports. (Modules have ports, through which they are connected to other modules.)
- Clock: SystemC's special signal (act as a system clock during simulation.)
- Cycle-based simulation: Supports an untimed model and includes high-level function model to RTL-level, which has clock cycle accuracy.

Figure 1 shows an example of system modeling in SystemC. A module can include processes or other modules due to its hierarchical structure. Processes run concurrently and do function modeling. Thus, they cannot have a hierarchical structure. Processes can communicate with each other through channels. Signals are modeled in the simplest form of a channel.

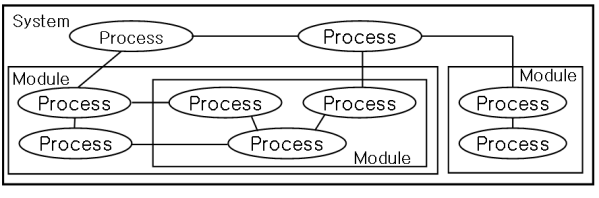


Fig. 1. A system modeling of SystemC

2.2 RDI (Remote Debug Interface)

RDI[6] is a C-based procedural interface that interfaces an ARM certified debug controller with a debug target. An RDI debug controller is a program that generates specific requests to a debug target through RDI. An RDI debug target then controls the received requests coming through RDI, so that the target hardware or software can perform debugging procedures. RDI consists of interface related to the controlling of the running of a debug target, and read/write operations of the states of the debug target. We used the RDI 1.5.1 that is an in-process interface and RID requests are handled through an RDI procedure vector in the form of windows DLL called WinRDI.

The primary purposes for WinRDI are as follows.

- To offer general mechanism to gain the RDI interface.

- To offer processes needed to adapt the versions of the debug controller and the debug target.
- To allow the setting of the debug target if necessary.

3 Design and Implementation

To build the development environment, we used SystemC to make the ARM7 core and memory. We also designed and implemented an interface using WinRDI so that the built core and memory can be connected with ARM's debug controller. Figure 2 shows the overall operation of the developed environment.

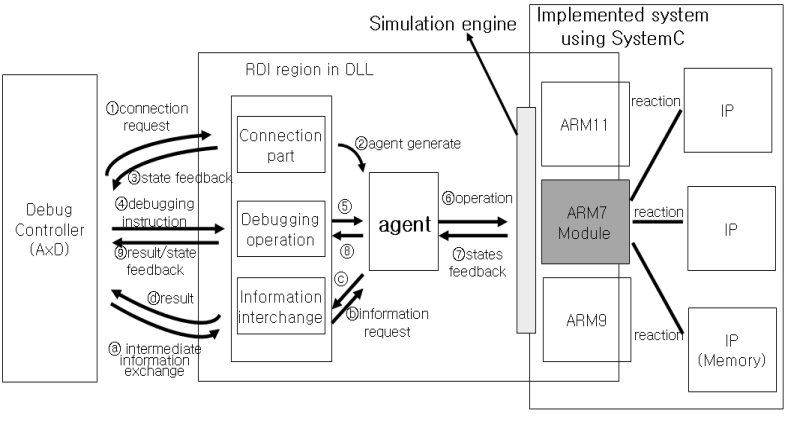


Fig. 2. The overall operation of the implemented VDE

If we set AxD (Debug controller)[7] to connect the implemented DLL, functions related to the connection are called and initializes the implemented interface. Next, AxD calls Info function on occasion and receives relevant information related to the modules and other peripherals that were requested at specific times. After the connection is set up, an agent is created and AxD recognizes SystemC-based modules and IPs according to its setting.

After this, if we run an application, the debugging becomes possible. When the debugging functions of AxD are called according to the debugging-related buttons that were pressed, the relevant functions of the agent are called, and the agent reads/sends the information that AxD want; AxD can set the values of registers and memory of the debug target.

RDI can be divided into three parts: the connection part, the debugging part, and the information exchange part. In the connection part, AxD sets the target according to RDI using the DLL. The connection part also enables AxD to connect to the target by processing the information of the target. At this point, we can create an agent and handle many targets. In our work, we connected only one target. In the debugging part, debugging functions are referenced

in Traget_ProVec of the agent so that appropriate debugging functions can be called. Finally, the information exchange part is called occasionally when AxD needs the state of the target.

3.1 Implementation of the Interface for the Connection to AxD

In this subsection, we describe interface functions that are necessary to connect debug controllers with connection procedures. The followings are the implemented entry points necessary for the connection with AxD.

- WinRDI_Valid_RDI_DLL: This function can be called at any time and checks the exact name of the implemented DLL.
- WinRDI_GetVersion: This function can be called at any point after the WinRDI_SetVersion is called. It is defined in winrdi.h and returns what type of RDI or WinRDI is supported in the DLL using a macro.
- WinRDI_Get_DLL_Description: This function shows the user the name of the DLL in null-terminated ASCII values.
- WinRDI_GetRDIProcVec: This can be called by the debug controller at any time. It creates the RDI entry point including struct RDI_ProVect for the debug target and returns the pointer.
- WinRDI_Register_Yield_Callback: The debug controller cannot enter into RDI_StepProc and RDI_ExecuteProc while the target is running. This function is used to solve that problem.
- WinRDI_Info: This function returns the information that the debug controller requests.
- WinRDI_SetVersion: This function returns the version that the debug controller demands from the target.
- WinRDI_Initialise: The DLL does the initialization required by this function and prepares RDI_OpenAgentProc.

To allow AxD to initialize the debug target, we made AxD to operate according to the procedure of RDI calls. (The termination procedure takes symmetrically to the initialization process.) At first, AxD initialize the agent for debug target and verify its handle. Then, AxD counts the number of modules (processors) of the debug target. Finally, for all the modules of the debug connection, AxD opens them and initializes each modules.

3.2 Modified Parts of SystemC.lib

SystemC generally uses SystemC.lib for modeling and simulation. Therefore, to connect with a debug controller, we analyzed the internal procedure of SystemC.lib and modified it according to our needs. To do this, we removed main() of SystemC.lib, which starts the simulation, and re-built SystemC.lib. Then we connected the starting function, sc_main(), of the SystemC.lib with the OpenAgentProc() of the implemented rdi.dll. Thus, the simulation can be controlled by the OpenAgentProc(). The class Csimul is implemented for the connection of RDI and SystemC simulation modules. Figure 3 shows the CSimul's behavior.

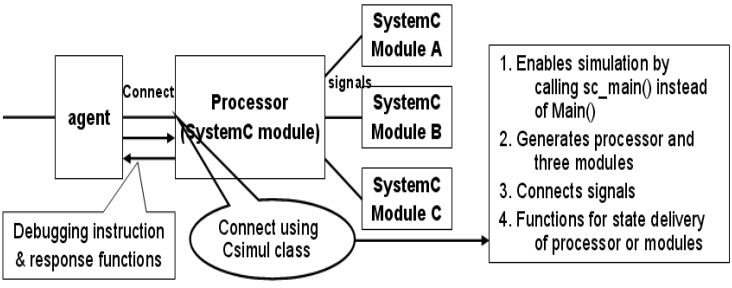


Fig. 3. CSimul class for the connection of RDI and SystemC simulation

The behavior of class CSimul is as follows.

1. Make sc_signal to control input/output wires of modules.
2. Connect signals after a core in SystemC is made.
3. Connect signals after memory in SystemC is made.
4. Create functions necessary for reading and writing of the state of the core.
5. Allows the result of simulation to be saved in the data waveform *vcd*.

To implement an ARM processor SystemC model, we used the ARM7 core C-model in GNU's GDB. The core was coded in C and provides simulation environment that, linked with GDB, runs assembly instructions. To change the C-model core into a SystemC model, we encapsulated the interface into SystemC while maintaining the internal part to run in C. Also, to make the control of debugging possible, we read internal information from pipelining functions, and the simulation runs in steps after saving the information. We also designed simple synchronized SystemC memory to run with SystemC core model.

3.3 SystemC Module Controlling Method for Debugging

In this subsection, we explain a method for the debug controller to debug a target while controlling the SystemC modules at the same time.

The simulation can start when all the necessary modules are created and the connection between signals and modules are validated. RDI calls CSimul's method, CSimul.init(), to control the starting of the simulation. This initialization process means the starting of the most top module of the simulation, and the actual simulation starts when sc_start() is called from the most top phase. The function sc_first() can have a double-type parameter and many time units as its value. If we want to run the simulation without a time limit, we can put a negative number for the parameter. All these functions run while the clock signal ticks for an assigned amount of units. When the time unit is all spent, the SystemC scheduler is called.

The simulation can be terminated at any time by calling sc_stop() with no parameters. But it is difficult to understand all the details above and to implement

the exact operations for each debugging step that the debug controller requires. To solve this problem, we used sc_initialize() function that starts the clock and controls the simulation, rather than sc_start(), which SystemC provides.

We initialize the SystemC scheduler using sc_initialize() function, not sc_start(). Then, we will be able to write values on the signals and simulate the result of the set values. This function takes a double-type parameter and can call the SystemC scheduler.

To implement one-step operation, we used sc_cycle() function and some internal variables such as F_pc, E_pc, and A_pc. The followings are implemented RDI debugging functions: RDI_ReadProc, RDI_WriteProc, RDI_CPUReadProc, RDU_CPUWriteProc, RDI_SetBreakProc, RDI_SetpProc, RDI_ExecuteProc, and etc.

3.4 Verification of the Developed Environment

The virtual development environment consists of the debug controller AxD, implemented interface, and SystemC modules. To verify the developed environment, we ran the developed environment and used vcd-file creation function of SystemC simulation environment to output the operation states and check its waves. We first checked the simulation of hardware models, the core and memory, through the vcd file. Next, using CodeWarrior of ARM Developer Suit v1.2, we had made many applications in assembly and C languages to verify wether the core module and the memory module correctly worked according to the debugging instructions (setbreak, getbreak, go, run, memory read/write, register, and read/write, etc.) generated from the debug controller through RDI. And thus, we checked whether the values of the core and memory changed and whether we could read/write the states. Figure 4 is the picture captured for checking the debug operation of an assembly and C programs.

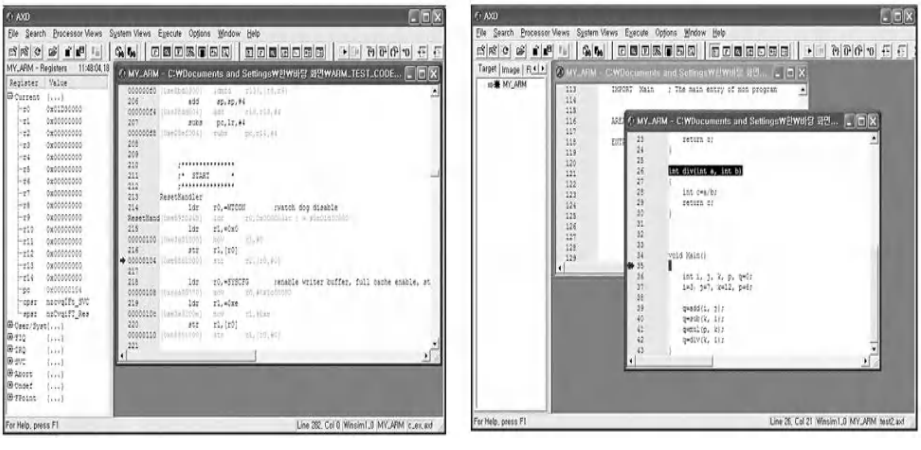


Fig. 4. An assembly program and a C program debugging test

4 Conclusion

In this paper, we described the design and implementation of a SystemC-based virtual development environment for developing embedded systems. The virtual development environment reduces the cost of the development of the embedded system by enabling engineers to write embedded software without real hardware. We implemented a virtual development environment that is based on SystemC. This environment was implemented by linking the AxD debugger with a SystemC-based hardware simulation environment through the RDI 1.5.1 interface. We minimized the modification of SystemC simulation engine so that the environment can be easily changed or extended with various SystemC models. The hardware simulation environment employed an ARM core that is the most commonly used one in commercial business. We implemented several SystemC-based hardware IPs for the virtual hardware environment.

For the verification of the developed environment, the core and memory were linked with the debug controller AxD. We ran example applications and verified the accuracy of our environment by checking the debug operations. The developed virtual development environment can be used in many phases of embedded software development such as developing a device driver, porting an OS, and developing an application.

The developed environment used the ARM processor core that is very popular in commercial business, so the environment is very useful in various application areas. Our environment uses the SystemC-based hardware simulation environment for the system-level design. So, the processor model, memory model, and IP model can be extended easily, and this environment can be run with many open SystemC models.

References

1. Virtio Corp. *VPDA295 Virtual Platform*, http://www.virtop.com/products/page/0,2573,33,00.html
2. ARM Corp. *Virtual Prototyping Solution*, http://www.arm.com/products/DevTools/MaxCore.html
3. Summit Design Corp. *Platform based Solutions*, http://www.summit-design.com/content.asp
4. Grotker, T. and Grotker, T., *System Design With SystemC*, Kluwer Academic Pub., (2002)
5. Ghenassia, F., *Transaction Level Modeling With SystemC*, Springer Verlag, (2006)
6. ARM Corp. *ARM RDI 1.5.1 RDI-0057-CUST-ESPC-B document*, (2003) p.206
7. ARM Corp. *ARM Develop Suite version 1.2 Debug Target Guide*, (2002)

Embedded Fault Diagnosis Expert System Based on CLIPS and ANN

Tan Dapeng, Li Peiyu, and Pan Xiaohong

College of Mechanical and Energy Engineering, Zhejiang University,
Hangzhou 310027, China
{tandapeng,lipeiyu,pan_xh}@zju.edu.cn

Abstract. Embedded fault diagnosis technology requires high pertinency and small occupation space. Traditional fault diagnosis system can not satisfy the demand above. Aiming at this problem, a kind of embedded fault diagnosis expert system (E-FDES) based on CLIPS and ANN was brought forward. FDES and its relative technology were discussed, developing environment and design tool chain for E-FDES was established in Linux environment. Using modularization theory principle, system re-configuration and expansion ability were guaranteed, and data process components and graphic user interface components could be reduced to meet the embedded system running requirement. By application program interface of CLIPS and Protégé plug-in, knowledge base and rule base were realized, that decreased the developing time of system inference engine and reduced source occupation space. Industry experiments prove that this system runs stably in ARM-Linux embedded environment, switch smoothly and can recognize some common faults for the devices monitored.

Keywords: embedded system, fault diagnosis, expert system, CLIPS, ANN.

1 Introduction

Expert system is an important research production of artificial intelligence, it can work efficiently in a small range field with difficult problems, and achieve the expert's ability. But the traditional expert system has heavy architecture, requires large scale data and space resource to keep its normal work, that can not adapt the embedded system environment demand such as small occupation space, high real-time performance, good configuration ability and so on. Aiming at the problem above, this paper introduces embedded system technology into the design process of FDES, make expert system applied in the embedded system environment[1].

2 System Architecture Design

According to the characteristic and special demand of FDES in embedded system environment, the developing chain of *C+CLIPS+Protégé+Eclipse+Qt* was adopted combined developing production and experience of relative depending project[2]. The realization method of FDES was described as follows: Firstly, bottom layer data

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 957–960, 2007.
© Springer-Verlag Berlin Heidelberg 2007

support component are realized by basic language C/C++, it will finish the work of real-time data collection, pre-process, calculation and extraction of symptom data etc; Then, system inference engine frame is established around by the C language integration product system (CLIPS), the management and development interactivity ability of monitoring objects system is strengthened by using Protégé, a visual ontology edit tool, and the process decision and inference are realized combined artificial neural network (ANN) technology; At last, graphic user interface (GUI) component is designed by mature graphic tool library *Qt/embedded*, and the application instance is developed integrated by and integrated developing environment tool-Eclipse.

3 System Crucial Technologies Design

3.1 Object Ontology System Based on CLIPS and Protégé

CLIPS was developed by NASA as an expert system developing tool, it has such advantages as convenience, low cost and easy to be integrated by outer system. Protégé is an open source ontology edit tool originated by Stanford university, it is a good choice to realize the building and ontology description of detecting target system and its components. According to the technical and running character of CLIPS, object ontology management system oriented to E-FDES is designed with followed steps: *1)* Every part of object system (factory, workshop, devices, sensors etc.) will be defined in the same mode, confirmed its class inheriting relation, instance and *slot* attribution in Protégé system frame; *2)* Using the CLIPS plug-in interface of Protégé, the defined ontology system is saved as the file with "*clp*" expansion name, and leaded into CLIPS frame to finish the ontology realization for embedded environment. *3)* In allusion to special requirement and application condition, CLIPS can modify the ontology file, then write back to Protégé frame. The above method integrates the merits of CLIPS and Protégé, resolves poor interactivity problem of CLIPS, reduces the developing time.

3.2 Inference Engine Based on CLIPS and ANN

Design of inference engine is the most important part of whole system realization. This paper adopts CLIPS and ANN to realize the diagnosis information integration, and obtain the diagnosis outcome. CLIPS can do fault inference by production rule, and has good rapidity. What's more, CLIPS has the ability of object-oriented, not only can realize the simply relational inference, but can process the diagnosis and recognition of complex interactive reticular fault network[3].

The application scope of expert system is very narrow, so E-FDES requires high specialty because of the space and calculation ability limitation of embedded environment. In the hand of inference effectivity, E-FDES requires long time industry fields train aiming at some kind of devices. Since ANN technology has the advantages for knowledge study, rule training and inference for random problem, this paper adopts it to realize the work above, and obtain the diagnosis outcome using information integration algorithm. ANN information integration algorithm combines the BP network and fuzzy set, then constructs a fuzzy neural network classifier, its input and output are confidence degree value with some semantic attribution. In the process of

diagnosis, the fault symptom data from RTDCS is looked as the input of BP network, the output is the confidence value of some kind of fault, then the deep decision is done according to relative decision rule, such as fault position, fault reason and fault measure, that require several ANN to work together, and the confidence degree is obtained by confidence function[4].

The running mechanism of system inference engine is shown in Fig.1. After system start, the monitoring object ontology and its corresponding instances are awaked, then the system parameters and diagnosis threshold are set according to special application condition, and the fault symptom data from RTDCS is transferred to match with CLIPS rules. If the matching confidence degree reach or exceed the diagnosis threshold, the diagnosis outcome can be determined directly; on the contrary, based on the matching outcome, normalization information integration is done combined ANN calculation outcome, and their confidence degree is compared to obtain the final diagnosis outcome.

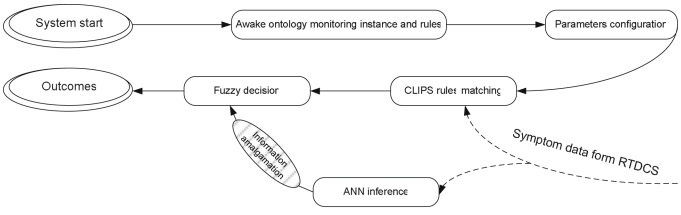


Fig. 1. System inference engine running mechanism

4 Application Instance Experiments

According to design steps and realization strategy, this paper has realized the E-FDES application system oriented to large-sized rotate machine. It can run stably in hand-held fault diagnosis instrument self-developed, and do large number of industry field experiments in some relative factory. As shown in Fig.2, operators are detecting the pour polymer pump. The pump requires long time continuous work, its key components valve apt to occur two kinds of typical fault: spring abruption (Fault A) and surface pit corrosion (Fault B). These faults occur at high randomness, their position is hard to judge. So the pump can but demount and inspect components one by one to find the fault diagnosis. By favor of operators, this system detected some devices which occur fault frequently, it can reach the recognition ratio of ninety percent, and can find the fault position correctly, the experiments outcome detail is shown in Fig.2.



Fault	E-times	Correct	Ratio
A	120	118	98.3%
B	120	106	88.3%

Fig. 2. Industry field application experiments

5 Conclusion

This paper introduces embedded system technology into the design process of FDES. On the base of stable and quick real-time data collect module, Protégé and CLIPS's advantage are combined to reduce the developing time and strengthen the system interactivity. By the integration of CLIPS and ANN, compact and effective inference engine is realized, and system application instance is developed by Eclipse, it can adapt the embedded system environment well. The industry experiment has proved its effectivity, and the future research work will be carried out around by real-time performance optimization and GUI rapidity.

References

1. Ranganathan, S., Beetner, D. G., and Wiese, R. H.: An expert system architecture to detect system-level automotive EMC problems. In Proceedings of IEEE International Symposium on Electromagnetic Compatibility (2002).
2. Vegh, J.: The "carbon contamination" rule set implemented in an "embedded expert system". Journal of Electron Spectroscopy and Related Phenomena, vol. 133, (2003) 87-101.
3. Guo, J., Liao, Y. H., and Pamula, R.: Extending eclipse to support object-oriented system verification. In Proceedings of IEEE International Conference on Information Reuse and Integration (2005).
4. Mweene, H. V.: Confessions of an eclipse consultant. Physics World, vol. 16, (2003) 56-59.

A Fault-Tolerant Real-Time Scheduling Algorithm in Software Fault-Tolerant Module

Dong Liu[1], Weiyan Xing[2], Rui Li[1], Chunyuan Zhang[1], and Haiyan Li[1]

[1] 6 yuan 7 dui, Department of Computer, National University of Defense Technology, Changsha, Hunan 410073, China
windleaf1980@163.com
[2] China Huayin Ordnance Test Center, Huayin 714200, China

Abstract. In software fault-tolerant module, the key issue that affects the performance of fault-tolerant scheduling algorithm is how to predict precisely whether a primary is executable. In order to improve the prediction precision, a new algorithm named DPA, Deep-Prediction based Algorithm, is put forward. DPA uses prediction-table to predict whether a primary can be completed before it is scheduled. Simulation results show that DPA provides more execution time for primaries and wastes less processor time than the well-known similar algorithms.

Keywords: scheduling algorithm, software fault-tolerance, real-time system.

1 Introduction

In addition to the function constraints and time constraints, real-time systems in life-critical applications should supply high reliability so that the systems would not halt in the presence of task failure. Therefore, various fault-tolerant modules were put forward to solve the problem. One of the widely used modules is software fault-tolerant module, which uses uniprocessor scheduling algorithms to schedule two versions for each real-time task, namely primary and alternate [1] [2] [3].

In software fault-tolerant module, real-time system has a set of n real-time periodic tasks $\tau = \{\tau_1, \tau_2, \ldots, \tau_n\}$. τ_i has a period T_i and two independent versions of computation program: the primary P_i and the alternate A_i. P_i has a computation time p_i, A_i has a computation time a_i, and, usually, $p_i \geq a_i$ for $1 \leq i \leq n$.

Let the planning cycle, T, be the least common multiple (LCM) of $T_1, T_2, \ldots, T_n$. We only need to consider all task invocations during the first planning cycle $[0, T]$. The primary and the alternate of the jth job, J_{ij}, of τ_i are denoted by P_{ij} and A_{ij}. We define $r_{ij} = (j - 1) T_i$ to be the release time and $d_{ij} = jT_i$ be the deadline of J_{ij}.

Since P_{ij} provides a better computation quality, we expect to complete as many primaries as possible while guaranteeing either the primary or the alternate of each task to be successfully completed by its deadline. If there are primaries pending for execution, alternate will not be scheduled until the latest possible time, called the notification time, denoted by v_{ij}. Therefore, we can say that v_{ij} is the *pre-deadline* of P_{ij}. FP is denoted as the probability that the primary fails after its execution.

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 961–964, 2007.
© Springer-Verlag Berlin Heidelberg 2007

[1] proposed BCE (Basic & CAT & EIT) algorithm to provide fault-tolerant scheme in software fault-tolerant module. BCE includes three sub-algorithms: Basic, CAT and EIT. Basic algorithm uses a fixed priority-driven preemptive scheduling scheme to pre-allocate time intervals to alternates as late as possible, and, at runtime, attempts to execute primaries before alternates. Since the prediction of an executable primary is the key issue that affects the scheduling performance, CAT (Check Available Time) algorithm is adopted to make prediction and abort the primary that has not sufficient time to finish. When processor becomes idle, EIT (Eliminate Idle Time) algorithm advances the scheduling of the alternate, which has the highest priority, to save time.

[2] and [3] enhanced BCE algorithm and introduced two algorithms: PKSA (Probing K-steps Algorithm) and CUBA (Changing Utilization-Based Algorithm). Both algorithms probe K steps, which are the probing deepness of the algorithms, before the executions of primaries, so that they can prevent early failure of one primary from triggering failures in subsequent tasks' executions. In this way, the algorithms increase the percentage of successfully completed primaries compared to BCE. Generally, PKSA works better than CUBA [2].

On the base of BCE and PKSA [1] [2], this paper proposes Deep-Prediction Based Algorithm (DPA) in order to precisely predict whether a primary is executable.

2 DPA Scheduling Algorithm

In this paper, we use a time-table, named prediction-table (PT), to make prediction for P_{ij}, named host primary. PT stores the execution information of P_{ij} and all its preemptive primaries P_{mn} and alternates A_{mn}.

Theorem 1. Let the current time be t, P_{ij} has been released, and the notification time of A_{ij} is v_{ij}. If the total time that has been allocated for primaries (not including P_{ij}) and alternates during $[t, v_{ij}]$ is I, then the available time of P_{ij} is $AT = v_{ij} - t - I$.

Proof. Since P_{ij} must be completed by its deadline v_{ij}, the maximum available time for P_{ij} to use will be $v_{ij} - t$. During $[t, v_{ij}]$, besides P_{ij}, there exist the primaries P_{mn}, which preempt P_{ij}, and the alternates, whose primaries have failed or been aborted. All of the primaries and the alternates preempt P_{ij} in PT, and the total time occupied by these tasks is I. Consequently, the available time, denoted by AT, for P_{ij} in PT will be $v_{ij} - t - I$. This is the relation we desired to prove.

The procedure of constructing PT for host primary P_{ij}, PT(P_{ij}), is described below:

Let t denote the current time. P_{ij} has been released.

(1) PT is initialized to be idle. Set the start-time and the end-time of PT to be t and v_{ij} respectively. Then, copy the time intervals that have been allocated to alternates by backward-RM scheduling algorithm [1] during $[t, v_{ij}]$ to PT. An alternate-list, containing all the alternates in PT, is created, and the alternates in the list are sorted ascendingly in order of their notification time.

(2) Let A_{mn} denote the first unchecked alternate, and P_{mn} be its primary. If P_{mn} has been aborted or failed, go to (2) to check the next alternate in alternate-list; otherwise, go to (3). If there is no unchecked alternate in alternate-list, go to (4).

(3) Judge whether the available time in PT is sufficient for P_{mn} using Theorem 1. If P_{mn} can meet its pre-deadline, then append P_{mn}'s time into PT and withdraw the time intervals occupied by A_{mn}. Otherwise, abort P_{mn}. Go to (2).

(4) Judge whether the available time in PT is sufficient for P_{ij} using Theorem 1. If P_{ij} can meet its pre-deadline, append P_{ij}'s time into PT, and return "successful". If P_{ij} can not finish, return "failed".

The following algorithm describes DPA in detail.

Procedure DPA( )

(1) Set the current time, t, to be 0; set the current scheduling state, S, to be "not PT".

(2) Use the backwards-RM algorithm to allocate time intervals for all alternates.

(3) The current time is t.

(3.1) If S = "not PT", then:

(3.1.1) If an alternate arrives, execute the alternate. Go to (4);

(3.1.2) If the processor is idle, do EIT algorithm. Go to (3)..

(3.1.3) If a primary arrives, let P_{ij} denote the primary whose alternate has the smallest notification time among all the released primaries. If PT(P_{ij}) = "successful", set S to "PT"; otherwise, abort P_{ij}. Go to (3).

(3.2) If S = "PT", let P_{ij} denote the host primary. Then:

(3.2.1) If it is time to execute P_{ij} in PT, schedule P_{ij}. If P_{ij} is completed, set S to "not PT". If there is no fault during P_{ij}'s execution, remove A_{ij}; otherwise, reserve A_{ij}.

(3.2.2) If it is time to execute P_{mn} ($P_{mn} \neq P_{ij}$) in PT, schedule P_{mn}. If P_{mn} is completed and failure occurred during its execution, set S to "not PT".

(3.2.3) it is time to execute A_{mn} in PT, schedule A_{mn}.

(4) Set the next time, and go to (3).

In the next section, we will use simulations to demonstrate the efficiency of DPA compared to PKSA. (Since PKSA is better than CUBA [2], we only compare PKSA and DPA.)

3 Simulation Results

We use two metrics: *PSP* (Percent of time used by Successful Primaries) and *W* (Wasted time units). *PSP* is the percentage of time used by successfully completed primaries. *W* is the time wasted by executing inexecutable (not failed) primaries.

In the simulation, we set *FP* to be 0.05 and set the probing deepness of PKSA to be 4, 8 and 16. We simulated five tasks with different T_i, p_i, a_i for 500 times. T_i of each task was between 10 and 150 time units and a_i was 0.1 times of T_i. By adjusting the proportion of primaries' computation time to their periods, we can control the utilization of primaries, U_P. We simulated 10000 time units in each simulation. Through this way, we got 500 results of *PSP* and *W*, respectively, and then the results were averaged. The averages are exactly the final results shown in Fig.1.

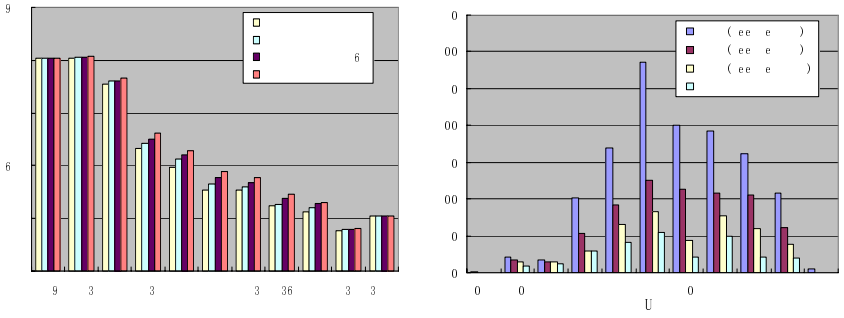


Fig. 1. Comparison between DPA and PKSA with different U_p

When $U_P < 1.1$, primaries' computation time is short, and both algorithms can predict that most of primaries have sufficient available time to execute, so two algorithms get similar scheduling performances. When $1.1 < U_p < 3$, DPA wastes less processor time and provides more time to successful primaries compared to PKSA. When $U_p > 3$, the performances of two algorithms get close again, since most primaries can not be completed before notification time.

4 Conclusion

In this paper, we proposed DPA algorithm for software fault-tolerant module. The algorithm constructs prediction-table, PT, to predict exactly whether a primary is executable. Simulation results show that DPA gets more execution time for primary and less wasted processor time compared to the well-known similar algorithms so far. The algorithm is being used in a RT-Linux based system and has realized fault-tolerant function well.

References

1. Han C.C, Shin K.G , Wu J: A Fault-Tolerant Scheduling Algorithm for Real-Time Periodic Tasks with Possible Software Faults. IEEE Transactions on computers, Vol. 52 (2003) 362-372
2. Han J.J, Li Q.H, Abbas A. Essa: A dynamic Real-Time Scheduling Algorithm with Software Fault-Tolerance. Journal of Computer Research and Development, Vol. 42 (2005) 315-321
3. Li QH, Han J.J, Essa A.A., Zhang W.: Dynamic scheduling algorithms with software fault-tolerance in hard real-time systems. Journal of Software, Vol. 16 (2005) 101-107

An Energy-Efficient Scheduling Algorithm for Real-Time Tasks

Youlin Ruan[1,2], Gan Liu[3], Jianjun Han[3], and Qinghua Li[3]

[1] School of Information Engineering, Wuhan University of Technology, 430070 Wuhan, P.R. China
[2] State Key Laboratory for Novel Software Technology, Nanjing University, 210093 Nanjing, P.R. China
[3] Department of Computer Science and Technology, Huazhong University of Science and Technology, 430074 Wuhan, P.R. China
ruanyl@126.com

Abstract. Based on maximal slack first, this paper proposes a novel energy-efficient scheduling algorithm for periodic real time tasks. The scheduling solution combines static priority and dynamic speed adjustment mechanism to save energy. Simulation results show that the proposed algorithm outperforms other major scheduling schemes.

Keywords: energy-efficient, maximal slack first, energy consumption.

1 Introduction

In recent years, several software techniques have been proposed to adjust the supply voltage. Krishna and Lee propose a power-aware scheduling technique using slack reclamation, but only in the context of systems with two voltage levels [1]. Mosse *et al.* propose and analyze several techniques to dynamically adjust processor speed with slack reclamation [2]. Aydin proposes a power-aware scheduling of periodic tasks to reduce CPU energy consumption in hard real-time systems through a static solution to compute optimal speed and an online speed reduction mechanism to reclaim energy[3].

By dynamic voltage scaling, Chen proposes an optimal real-time task scheduling algorithm for multiprocessor environments with the allowance of task migration[4].Zhu *et al.* present two algorithms GSSR based on the concept of slack sharing for single task sets with and without precedence constraints. These scheduling techniques are based on longest task first, which reclaim the time unused by a task to reduce the execution speed of future tasks, and thus reduce the total energy consumption of the system[5]. However, Han proposed the opposite strategy STFBA1 that is based on the policy of shortest task first and combine with other efficient techniques[6], which believe same slack used by longer tasks can save more energy.

We aim to the strategy of tasks assignment, and how to dynamic adjust speed to save energy consumption. The rest of the paper is organized as follows. Section 2 describes the task model, energy model and power management schemes. A dynamic power aware scheduling with maximal slack first for periodic tasks is addressed in section 3. Simulation and comparison are given and analyzed in section 4.

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 965–968, 2007.
© Springer-Verlag Berlin Heidelberg 2007

2 System Models and Schemes

In this paper, we assume that $E = C \cdot C_{ef} \cdot \frac{S^2}{k^2}$ holds as defined in literature for simplicity[10].We assume a frame based real-time system in which a frame of length D is executed repeatedly. A set of tasks $\Gamma = \{ T_1, \ , T_N \}$ to execute within each frame and is to complete before the end of the frame. Because of the schedule's periodicity, we consider only the problem of scheduling Γ in a single frame with deadline D.

In specifying the execution of a task T_i, we use the three tuple ($WCET_i$,$ACET_i$, AET_i), where $WCET_i$ is the estimated worst case execution time (WCET), $ACET_i$ is the estimated average case execution time ($ACET$) and AET_i is the actual execution time (AET), which are all based on maximal processor speed. We assume that for a task T_i, the value of $WCET_i$ and $ACET_i$ are known before execution, while AET_i is determined at run time. To get maximal energy savings, we combine static priority assignment and dynamic voltage/speed adjustment. Thus, we assume that canonical execution is first checked to see whether a task set can finish before D or not. If not, the task set is rejected; otherwise, our algorithm then apply static priority assignment and dynamic voltage/speed adjustment to save energy.

3 Maximal Slack First-Based Algorithm

Zhu propose global scheduling with longer tasks high priority, which believes longer tasks can generate more dynamic slack during execution that can be used by shorter tasks, thereby, it will reduce the energy consumption. However, it is not always true. Thus, Han propose an opposite strategy that is based on shortest task first, which believe same slack used by longer tasks can save more energy. In fact, such two strategies are not always efficient under all conditions, which fit for different aspects respectively. In the following, we will point out that both strategies are not always the most efficient heuristic to save energy consumption.

Lemma 1. Assuming that there exist two tasks T_i and T_j, the *WCET* and *ACET* of T_i are x and x-a respectively, the *WCET* and *ACET* of T_j are y and x-a respectively, and $x>y$. The energy consumption generated by the execution order in which T_j is prior to T_i is higher than that generated by the reverse execution order.

Proof. We assume that the energy generated by small slack first heuristic is no longer than that generated by large slack first heuristic, then

$$(x-a)C_{ef}\frac{1}{k^2}+(x-a)C_{ef}(\frac{x}{y+a})^2\frac{1}{k^2} \quad (x-a)C_{ef}\frac{1}{k^2}+(x-a)C_{ef}(\frac{y}{y+a})^2\frac{1}{k^2}$$

$$\Leftrightarrow (x-a)+(x-a)(\frac{x}{y+a})^2 \quad (x-a)+(x-a)(\frac{y}{y+a})^2$$

Since x>y>a>0 hold, then $(\frac{x}{y+a})^2 > (\frac{y}{y+a})^2$, the left part of above inequality is not true. Hence it leads to contradiction.

Lemma 2. Assuming that there exist two tasks T_i and T_j, the *WCET* and *ACET* of T_i are x and y-a respectively, the *WCET* and *ACET* of T_j are y and x-a respectively, and

$x>y$. The energy consumption generated by the execution order in which T_j is prior to T_i is higher than that generated by the reverse execution order.

Proof. We assume that the energy generated by small slack first heuristic is no longer than that generated by large slack first heuristic, then

$$(x-a)C_{ef}\frac{1}{k^2}+(y-a)C_{ef}(\frac{x}{y+a})^2\frac{1}{k^2} \quad (y-a)C_{ef}\frac{1}{k^2}+(x-a)C_{ef}(\frac{y}{x+a})^2\frac{1}{k^2}$$

$$\Leftrightarrow \quad (x-a)+(y-a)(\frac{x}{y+a})^2 \qquad (y-a)+(x-a)(\frac{y}{x+a})^2$$

$$\Leftrightarrow (x-a)[1-(\frac{y}{x+a})^2]\le(y-a)[1-(\frac{x}{y+a})^2]$$

Since x>y>a>0 hold, then $(x-a)>(y-a)$ and $(\frac{x}{y+a})^2>(\frac{y}{x+a})^2$, $[1-(\frac{y}{x+a})^2]>[1-(\frac{x}{y+a})^2]$, the left part of above inequality is not true. Hence it leads to contradiction

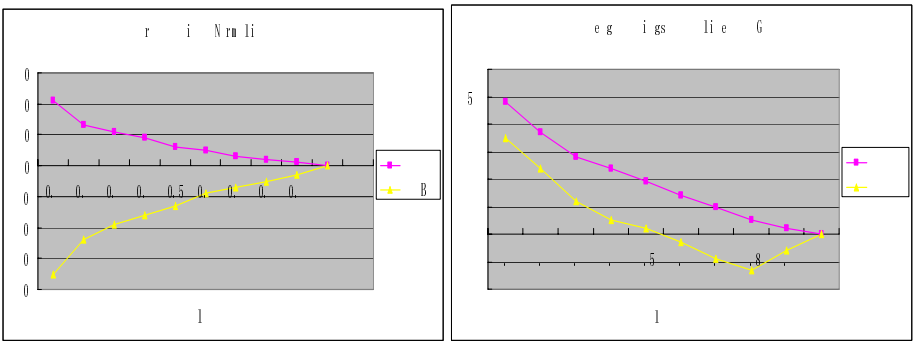
Thus, we can see that both longest task first and shortest slack first are not always efficient strategies. The key problem to save energy is the size of slack used by the following tasks. Therefore, we can draw a conclusion that the strategy based on maximal slack first. The longer slack of task generates, the higher priority of the task has. If the two slacks generated by two tasks are equal, shortest task first can save more energy. Moreover, longer tasks may generate more dynamic slack during execution. Thus, we use the longest task first heuristic to determine task's priority when the two slacks are 0, which may save more energy consumption. Thus, we propose the MSF algorithm, which combine the strategy of maximal slack first and the concept of dynamic adjusting of speed of tasks. The MSF algorithm comprises of two parts: static algorithm and dynamic algorithm. Static algorithm computes the static slack of each task offline and determines the priorities of tasks. Dynamic algorithm do not modify priorities of tasks, but adjust speed of tasks according to the actual execution time and dynamic slack, which is shown in following.

```
Dynamic Algorithm of MSF
{ slack=0; k=1;
   while Ready-Q<>• Do    //Tasks Γ = {T1,…,Ts}
     Tk=Dequeue(Ready-Q);
     Sk = Smax * WCETk / (WCETk + slack) ; Execute Tk at speed Sk;
     ETk=AETk/Sk; slack=slack+WCETk- ETk; k=k+1;
   endwhile }
```

4 Experiment Results

In this section, we present results of simulations performed to compare three algorithms MSF, GSSR and STFBA1. We define α_i as average/worst case ratio for T_i's

execution time, the actual execution time T_i will be generated as a normal distribution around $\alpha_i \cdot WCET_i$. Let U(i, j) be a uniformly distributed integer in the range of [i, j], the number of tasks is U(40, 400), the WCET of each task is U(20, 200). The value of α is 0.1-1.0. Regardless of α_i of all tasks are same and different, we can see that MSF consume less energy than GSSR and STFBA1 from Fig.1.



(a) same α_i (b) different α_i

Fig. 1. Comparison of energy savings

Acknowledgments. This work has been supported by the national natural science foundation of china under grant no. 60503048 and 60672059, the 863 program no. 2006AA01Z233.

References

1. Krishna C.M., Lee Y.H..:Voltage Clock Scaling Adaptive Scheduling Techniques or Low Power in Hard Real-Time Systems. Proc.6th IEEE Real-Time Technology and Applications Symp.(2000)
2. Mosse D., et al.. :Compiler-Assisted Dynamic Power-Aware Scheduling for Real-Time Applications. Proc.Workshop Compiler and OS for Low Power (2000)
3. Hakan Aydin, Rami Melhem, Daniel Mosse.: Power-Aware Scheduling for Periodic Real-Time Tasks, IEEE Transaction on Computer. 5(2004)584–600
4. Chen J.J, Kuo T.W. :Multiprocessor Energy-Efficient Scheduling for Real-Time Tasks with Different Power Characteristics. Proc. 2005 International Conference on parallel Processing (2005
5. Zhu D, Rami Melhem, Bruce Childers. :Scheduling with Dynamic Voltage/Speed Adjustment Using Slack Reclamation in Multiprocessor Real-Time Systems. IEEE Transaction on Parallel and Distributed Systems, 7 (2003)686–699.
6. Han J.J, Li Q.H,. :Dynamic Power-Aware Scheduling Algorithms for Real-Time Task Sets with Fault-Tolerance in Parallel and Distributed Computing Environment, Proceedings of the 19th IEEE International Parallel and Distributed Processing Symposium (2005)

An EDF Interrupt Handling Scheme for Real-Time Kernel: Design and Task Simulation

Peng Liu[1], Ming Cai[1], Tingting Fu[2], and Jinxiang Dong[1]

[1] Institute of Artificial Intelligence, Zhejiang University, Hangzhou, PRC, 310027
{perryliu,cm,djx}@zju.edu.cn
[2] Institute of Graphics and Image, Hangzhou Dianzi University, Hangzhou, PRC, 310018

Abstract. Traditional model of interrupt managemnent has been used for several decades. But it is often incapacity to incorporate reliability and temporal predictability demanded on real-time systems. Many solutions have been proposed to improve efficiency of interrupt handling such as In-line Interrupt Handling and Predictable Interrupt Management In this paper we propose a model that schedules interrupts in terms of their deadlines. Hard priorities of IRQs are still left to hardware. We only manager interrupts that can enter the kernel space so that hard real-time can be assured. Each interrupt will be scheduled only before its first execution according to their arrival time and deadlines so that the scheme is very simple and easy to be implemented. The scheme tries to make as many as possible ISRs finish their work within the time limit. Finally we do some experiments, which prove there is a big decrease of nested overtime interrupts, by means of task simulation on VxWorks.

Keywords: Real-time system, Interrupt scheduling, Task simulation, Similar Earliest-Deadline-First.

1 Introduction

Most embedded systems include a lot of external devices. The interrupt mechanism is a very important interface between the kernel and peripherals, which communicate the system with its external environment. Too many interrupt sources will cause the number of expired ISRs getting too high. It will also cause potential system instability or inconsistency. For example, as seen in fig. 1(left chart), I_1, I_2, I_3 are three interrupt routines which priorities meet $PI_1 < PI_2 < PI_3$. Their trigger time and endurance time can be seen in table 1. According to traditional scheme of interrupt management I_1 is preempted by I_2 even if it is about to finish its work. And I_2 is preempted by I_3 twice in succession. So I_1 completes its work at moment 15. Because its endurance time is 3 its work has been made no sense. I_2 has been overtime as well.

A number of research works propose alternatives to avoid the difficulties of the traditional interrupt model for real-time applications. Some have adopted radical solutions where most external interrupts are disabled and treat all peripherals by polling [1]. Other strategies have been proposed to obtain some degree of integration among the different types of asynchronous activities [2]. In [3], interrupts are treated as threads. In [4], a schedulability analysis integrating static scheduling techniques and

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 969–972, 2007.
© Springer-Verlag Berlin Heidelberg 2007

response time computation is proposed. In [5] an integrated scheme of tasks and interrupts to provide predictable execution times to real-time systems is presented. In this paper, we propose a novel strategy that manages interrupts using EDF scheduling. It is supposed to avoid many drawbacks of above schemes.

Table 1. Four ISRs (Tasks) invoked in the system

IRQ Number	Arrival T	Service T	Max. alive T	Priority
I_1 (K_1)	1	3	3	PI_1 (PK_1)
I_2 (K_2)	3	7	10	PI_2 (PK_2)
I_3 (K_3)	8	2	4	PI_3 (PK_3)
I_3 (K_3)	11	2	4	PI_3 (PK_3)

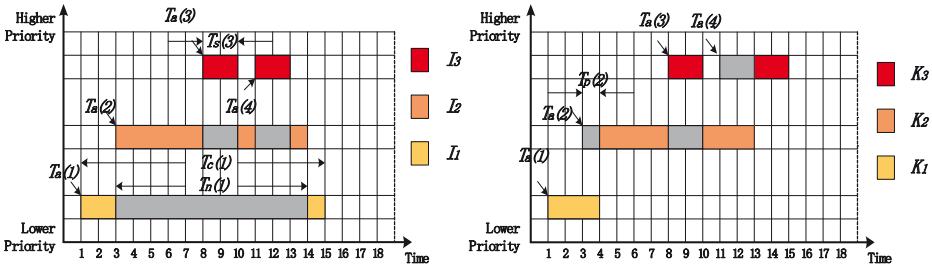


Fig. 1. Results of two interrupt management schemes

2 Model and Mechanism

Earliest-Deadline-First algorithm is a dynamic scheduling scheme. Priorities of tasks are dynamic according to their start-time and deadline. Those whose deadline is nearest to the current time have higher priorities. Priorities should be calculated again after the end of a task. The last step of scheduling is to choose a ready task that has highest priority.

Our scheme is a Similar EDF algorithm which sorts interrupt service routines according to three factors: hard priority, arrival time and deadline. Only when higher ISRs have spare time to allow lower ISRs to finish their work the scheduling will happen. Otherwise higher ISRs will preempt lower ones even though they may be overtime. Furthermore every ISR will be scheduled at most once for the sake of simplicity. Just like a stack, once an ISR is scheduled to run as if it is put into "stack". Others may preempt it as put into "stack" above it. Only when it is the top item of the "stack" can it be performed again. In our scheme following properties are considered: Processor availability A(i); Maximum alive time (deadline) $T_b(i)$; Arrival time $T_a(i)$; Service time $T_s(i)$; Execution time $T_e(i)$; Nested time $T_n(i)$; Hard Priority PI(i)

Maximum alive time means how long the ISR could stay in the system to finish his work before the deadline. Arrival time denotes when the interrupt is triggered. Service time means how long it would take for an ISR to complete a mission. Execution time indicates how long the ISR has taken to do its work. Nested time represent the period that the ISR is preempted by others. So we have:

$$U(t) = \text{Min}((U(t\text{-}1) - T_s(i)), (T_b(i) - T_s(i))) \quad (1)$$

$$T_e(i) <= T_s(i) \quad (2)$$

An interrupt j can be scheduled only and only if:

$$T_b(j) - T_s(j) >= T_s(i) - T_e(i) \quad (3)$$

We encapsulate the scheduling codes and user ISR into a code segment called realIntISR and use it to replace user ISR. It includes entering scheduling logic and exiting scheduling and is transparent to users.

3 Task Simulation and Performance

Here we have three tasks to simulate three interrupts I_1, I_2 and I_3 respectively on VxWorks. Their parameters are set as shown in table 1. We first create three tasks synchronously. Then put them into sleep until their pre-set trigger time is matched. Tasks just print message "intX is executing" onto the screen every second. We augment time scale to second to see the result clearly. After testing of the task simulation, the original system brings a result like what is shown in fig. 1(left chart). As expected, I_1 (simulated by K_1) is preempted by I_2, then by I_3 and I_4. In improved system, EOI will be sent after scheduling immediately for not to prevent the same level interrupt from entering. As shown in fig. 1(right chart), the total time consumed by three simulated task is not changed, but all three interrupts finished their work in time while two of them failed in original system.

The performance of the algorithm is tested also using task simulation. In each ISR, execution timer is recorded. If it is overtime a global count will be increased. We set five tasks which can be seen in table 2.

Table 2. Parameters of five tasks

IRQ Number	1	2	3	4	5
Service time	990ms	150ms	1.9ms	1.8ms	1ms
Lifecycle(Deadline)	4800ms	600ms	6ms	6ms	4.2ms
Interrupt Frequency	12/60s	60/60s	3000/60s	3000/60s	4200/60s

We keep those tasks running for a fixed period on the original and improved systems respectively. The result can be seen in fig. 2. The left chart shows new algorithm can do more interrupts than traditional algorithm in an equal period. The right chart shows new algorithm has less overtime interrupts than traditional algorithm. In conclusion, our scheme can improve the performance about 30% in situation that there are a good many interrupts and most of them have limited execution time while some of them could wait for execution for some time. Due to additional code introduced by the scheme, the performance of the system maybe decreases when a lot of interrupts have very short service time and restricted maximum alive time.

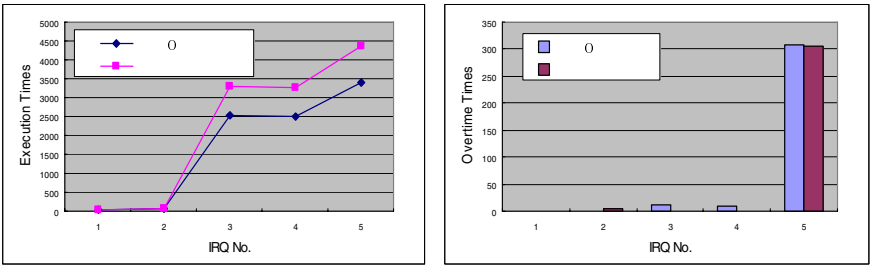


Fig. 2. Performance contrast

4 Conclusion

Most embedded systems have many interrupt sources and these interrupts will occur asynchronously. When there are too many nested interrupts those in low level are likely to run out of time, which leads to failure of their work. In this paper we presented a Similar Earliest-Deadline-First handling scheme to provide schedulability to interrupt management of real-time systems. Its algorithm and architecture were discussed. A simulation using tasks was presented. Result of performance test, which was carried out based on that simulation, was given. It was proved that using Similar EDF scheduling could greatly reduce interrupt failure caused by nested interrupts and enhance robustness of the real-time system.

References

1. Hermann, K., et-al.: Distributed Fault-Tolerant Real-Time Systems: the MARS Approach. IEEE Micro, Vol. 9, Issue 1. (1989) 25-40
2. Tullio, F., et-al.: Non-Preemptive Interrupt Scheduling for Safe Reuse of Legacy Drivers in Real-time systems. In: Proceedings of the 17th EuroMicro Conference on Real-Time Systems. (2005) 98-105
3. Steve, K., Joe, E.: Interrupts as Threads. ACM SIGOPS Operating Systems Review, Vol. 29, Issue 2, (1995) 21-26
4. K, Sanstrom., C, Erikssn., G, Fohler.: Handling Interrupts with Static Scheduling in an Automotive Vehicle Control Systems. In: Proceedings of 5 International Conference on Real-Time computing Systems and Applications. (1998) 158-165
5. Luis E, L-d-F., Pedro, M-A., Dionisio de, N.: Predictable Interrupt Management for Real Time Kernels over Conventional PC Hardware. In: Proceedings of IEEE Real-Time and Applications Symposium. (2006) 14-23
6. Aamer, J., Bruce, J.: In-Line Interrupt Handling and Lock-Up Free Translation Lookaside Buffers (TLBs). IEEE Transaction on Computers, Vol. 55. (2006) 559-574

Real-Time Controlled Multi-objective Scheduling Through ANNs and Fuzzy Inference Systems: The Case of DRC Manufacturing

Ozlem Uzun Araz

Dokuz Eylul University Industrial Engineering Department, 35100, Izmir, Turkey
ozlem.uzun@deu.edu.tr

Abstract. In this paper, we developed an integrated multi-objective real-time controlled scheduling methodology for Dual Resource Constrained systems. The proposed methodology integrates simulation, neural network and fuzzy inference system approaches to obtain a schedule considering both state of the system and objectives. By means of a case study, we have demonstrated that the proposed methodology can be an effective tool for dynamic scheduling of DRC systems.

Keywords: Dynamic Scheduling, Fuzzy Inference Systems, Artificial Neural Networks, Dual Resource Constrained Manufacturing System.

1 Introduction

Scheduling decisions are increasingly seen as strategic functions for manufacturing firms. Because of its increasing importance, many researchers have paid attention to develop numerous scheduling approaches for machine-constrained manufacturing systems. However, until recent years, the scheduling of Dual-Resource Constrained (DRC) manufacturing systems has not received much attention in the literature. Scheduling of these systems is more complicated than the others. Most of the DRC offline scheduling approaches developed in the literature use dispatching rules [1, 2, 3, 4, 5, etc.].

A drawback, however, of using dispatching rules is that their performances are dependent on the state of the system, but there is no rule that is the best performer for all the possible state of the system [6]. Therefore, for flexible and dynamic scheduling (DS) decisions, an effective tool is required to help the decision maker in selecting the best rule for each particular state of the system. Many researchers have proposed simulation and dispatching rule based dynamic scheduling approaches for machine constrained manufacturing systems [7, 8, 9, etc]. Due to the time consuming nature of simulation, some researchers have developed the artificial intelligence techniques for effective solutions of dynamic scheduling problems [10, 11, etc.].

To the best of our knowledge, the applicability and effectiveness of the artificial neural networks (ANN) based dynamic scheduling methods have not yet been fully explored on complex DRC manufacturing systems. In this paper, a novel real-time controlled multi-objective DRC scheduler is proposed to dynamically select dispatching rules. The proposed methodology, tested in a hypothetical manufacturing system, proves to be an effective method in a dynamic DRC manufacturing system.

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 973–976, 2007.
© Springer-Verlag Berlin Heidelberg 2007

2 The Multi-objective DRC Scheduler

The proposed scheduler (see Fig. 1) integrates several tools, namely; a simulation model, a backpropagation neural network (BPNN) and a fuzzy inference system (FIS), to dynamically select dispatching rules. A simulation experiment is conducted to collect predefined performance measures corresponding to the current dispatching rule set and the system states. BPNN is used to obtain the performance measures for each shop configuration. A shop configuration considers various decision variables such as dispatching rule sets, due date tightness, arrival rates, etc. Then the system is real time controlled and its status is analyzed in order to detect pre-defined symptoms. When a symptom becomes active, BPNN generates performance measures for all possible shop configurations. In order to compare the system performance between all shop configurations, an aggregated performance measure for each configuration is provided using a fuzzy inference system. Finally, the best configuration is determined and used until a new symptom becomes active.

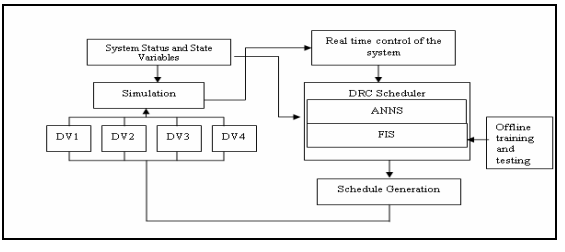


Fig. 1. The proposed real time scheduling mechanism. DV1: decision variables 1, etc

3 An Application of DS for DRC Manufacturing System

In order to investigate the potentials and effectiveness of the proposed methodology, the hypothetical DRC manufacturing system that consists of 24 departments is studied. There are 43 machines and 25 workers in the shop floor. Ten different parts can be processed by several flexible routing sequences. In this study, four decision variables are considered for scheduling of the system.

- Queue disciplines: First in first out (FIFO), shortest processing time, earliest due date, shortest remaining processing times, critical ratio, minimum slack time, critical ratio/ shortest processing time.
- When labor assignment rules: Centralized and decentralized rules are selected to determine when the worker should move to another station.
- Where labor assignment rules: The longest number in queue (LNQ), the longest waiting time in the queue and the work center with job which has shortest processing time and traveling time rules are selected to schedule of the worker.
- Alternative routing selection: Smallest number in queue (SNQ), shortest processing time at an operation, lowest average utilization first.

Four performance measures are considered to evaluate the alternative scheduling policies: mean tardiness, number of tardy jobs, mean flow time and throughput rate. Mean queue time, WIP, average and maximum utilization of machines and operators, arrival rate, flow allowance factor and number of part types in the system are selected as system state variables.

The simulation model was developed by using ARENA 3.0 simulation software. The simulation length is determined 50000 minutes and 10 independent replications. Then, BPNNs model for each performance measures are developed using 500 randomly generated training and testing samples. The inputs to the BPNNs are decision rule set and current system state variables, while the output of the system is the considered performance measures. After training and testing the ANN models, FIS model is generated by using MATLAB 7.0 fuzzy logic toolbox. The FIS system consists of four inputs which represent the performance measures of the DRC system, one output which is the aggregated fuzzy rating and 81 rules. The alternative which has highest rating score is selected as new schedule.

In order to investigate the effectiveness of the approach, it is compared with fixed scheduling approach (denoted as [1111]) in which "when rule" is centralized, "where rule" is LNQ, part selection by machine is FIFO, machine selection by part is SNQ. The manufacturing system is simulated for 15.000 minutes with both fixed scheduling and the proposed dynamic scheduling approaches. During the simulation period, system state variables were changed randomly in order to ensure the dynamic nature of the system. The fixed scheduling approach is regulated by fixed decision rules for decision variables at the start of the simulation and this decision rule set is not changed during the scheduling period.

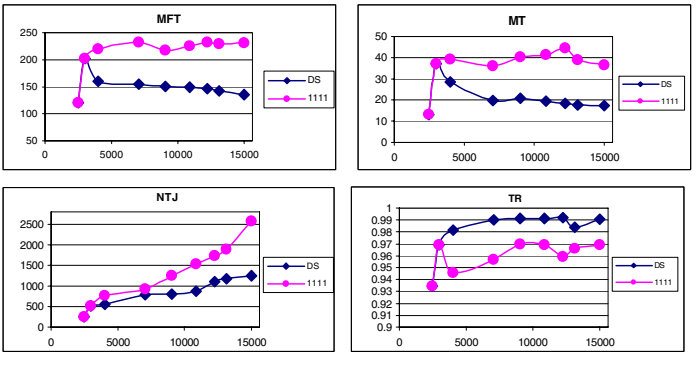


Fig. 2. Performance measures at each rescheduling point

During the application of dynamic scheduling approach, eight rescheduling points was detected. Performance measure values at each rescheduling point obtained from the fixed and dynamic scheduling approaches are illustrated in Fig. 2. As can be seen in figures, the proposed dynamic scheduling approach is more effective than the fixed scheduling approach in terms of all performance measures. It can easily be realized that when the value of a performance measure deteriorates, the proposed approach renews the schedule so that the overall performance of the system can be improved. It

should also be noted that the performance of the proposed approach is compared with all alternative fixed rule sets. These results also prove that the proposed approach is superior to all fixed scheduling rules in terms of all performance measures.

4 Conclusion

In this paper, we presented a new multi-objective Dual Resource Constrained scheduler for dynamic scheduling. The proposed methodology integrates three tools namely; simulation, neural network and fuzzy inference system. In spite of the complexity of DRC manufacturing, the result of this study indicates that the proposed methodology is an effective method to find appropriate dispatching rules.

References

1. Nelson, R.T.: Labor and machine limited production systems. Management Science. 13 (1967) 648-671.
2. Treleven, M.: Applications and implementation. The timing of labor transfers in a dual resource-constrained systems: “push” vs. “pull” rules. Decision Sciences (1987) 18 73-88.
3. Bobrowski, P.M., Park, P.S.: An evaluation of labor assignment rules when workers are not perfectly interchangeable. Journal of Operation Management. 11(1993) 257-268.
4. Malhotra, M.K., Kher, H.V.: An evaluation of worker assignment policies in dual resource-constrained job shops with heterogeneous resources and worker transfer delays. International Journal of Production Research. 32 (1994) 1087-1103.
5. Bokhorst, J.A.C., Slomp, J., Gaalman, G.J.C.: On the who-rule in Dual Resource Constrained manufacturing systems. Int. J. Prod. Res. 42 (23) (2004) 5049-5074.
6. Priore, P., Fuente, D., Puente, J., Parreno, J. : A comparison of machine-learning algorithms for dynamic scheduling of flexible manufacturing systems. Engineering Applications of Artificial Intelligence. 19(3) (2006) 247-255.
7. Kim, C.O., Kim, Y.D. Simulation based real-time scheduling in a flexible manufacturing system. Journal of Manufacturing System. 13 (1994) 85-93.
8. Sabuncuoğlu, I., Kızılışık, O.B.: Reactive scheduling in a dynamic and stochastic FMS environment. International Journal of Production Research. 41(17) (2003) 4211-4231.
9. Chan, F.T.S., Chan, H.K., Lau, H.C.W., Ip, R.W.L.: Analysis of dynamic dispatching rules for a flexible manufacturing system. Journal of Material Processing Technology. 138 (2003) 325-331.
10. Min, H.S., Yih, Y., Kim, C.O.: A competitive neural network approach to multi-objective FMS scheduling. International Journal of Production Research. 36(7) (1998) 1749-1765.
11. Min, H.S., Yih, Y.: Selection of dispatching rules on multiple dispatching decision points in real-time scheduling of a semiconductor wafer fabrication system. International Journal of Production Research. 41(16) (2003) 3921-3941.

Recursive Priority Inheritance Protocol for Solving Priority Inversion Problems in Real-Time Operating Systems*

Kwangsun Ko, Seong-Goo Kang, Gyehyeon Gyeong, and Young Ik Eom

School of Information and Communication Eng., Sungkyunkwan University,
300 Cheoncheon-dong, Jangan-gu, Suwon, Gyeonggi-do 440-746, Korea
{rilla91,lived,gyehyeon,yieom}@ece.skku.ac.kr

Abstract. In this paper, a protocol, called recursive priority inheritance (RPI) protocol, is proposed to solve complicated priority inversion problems as well as basic one. Our proposed protocol is implemented and tested in the Linux kernel. Additionally, the performance of our proposed protocol is evaluated and compared with previous protocols in the aspect of time and space complexity analyses.

1 Introduction

Nowadays, real-time operating systems are widely used in various fields where each task must finish execution by its deadline and should meet several requirements such as effective scheduling policies, reducing interrupt handling time, solving priority inversion problems, and so on. Among them, in this paper, the priority inversion problem [1], which means the situation that a high-priority process waits for acquiring the resource that is already locked by a low-priority process, is focused on. This situation occurs sometimes in real-time environments but certainly should be solved to remove the bad effects to the systems. The various solutions to this problem have been presented, among which two recommended protocols are basic priority inheritance (hereafter, BPI) protocol [2][3][4] and priority ceiling emulation (hereafter, PCE) protocol [5][6][7]. However, they cannot solve some complicated problems, such as the cases when a process locks several resources and when resources are locked and requested recursively. In this paper, a protocol, called recursive priority inheritance (hereafter, RPI) protocol, is proposed. The RPI protocol efficiently solves complicated priority inversion problems as well as basic one, just requiring low time/space complexity compared with pre-existing protocols.

* This research was supported by the MIC(Ministry of Information and Communication), Korea, under the ITRC(Information Technology Research Center) support program supervised by the IITA(Institute of Information Technology Advancement) (IITA-2006-C1090-0603-0027).

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 977–980, 2007.
© Springer-Verlag Berlin Heidelberg 2007

2 RPI Protocol

RPI protocol can solve the complicated priority inversion problems using recursive data structure, which is assigned to each individual process and keeps the entire information on both the resources locked by the process and the processes requesting these resources. (see Figure 1). (Let P_x denote processes, and R_x a resource; each process has different priority, and the priority of P_0 is the lowest.)

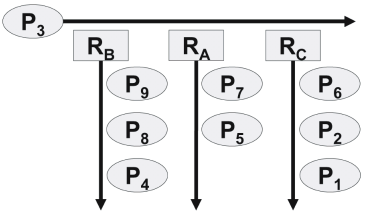


Fig. 1. Conceptual recursive data structure of a process P_3

As can be seen, P_3 locks three resources R_A, R_B, and R_C, and each resource has its own waiting queue for all processes that requested the resource. Processes in the waiting queue of each resource are sorted in the order of their priorities, and the resources locked by P_3 are also sorted in the order of priorities of the first processes in their waiting queues. Therefore, the first process in the waiting queue of the first resource has always the highest priority among the processes in the queues, and P_3 inherits the priority of the process; that is, P_3 inherits the highest priority.

Basically, RPI protocol uses several data structures: a recursive data structure and three new data structures. The former is implemented by adding a few fields to the *task_struct* structure, which is a structure that contains the entire information on the corresponding process in the Linux kernel. This data structure, a new *task_struct* structure, is initialized in *copy_process()* kernel function, which is invoked when a process is created. Additionally, our proposed protocol uses two operations, *P()*/*V()*, related a semaphore to implement the operations of 'locking' and 'unlocking' a resource. Operating systems, such as Unix and Linux, which support POSIX standards uses *semop()* system call for *P()*/*V()* operations of a semaphore. In the Linux kernel, *semop()* system call invokes the *sys_semtimedop()* kernel function, and we modified this function so that it can perform what we want when *P()*/*V()* operations are called.

3 Test and Evaluation

Our proposed RPI protocol can solve the problem when resources are locked and requested recursively; this problem cannot be solved by previously recommended protocols. (see Figure 2).

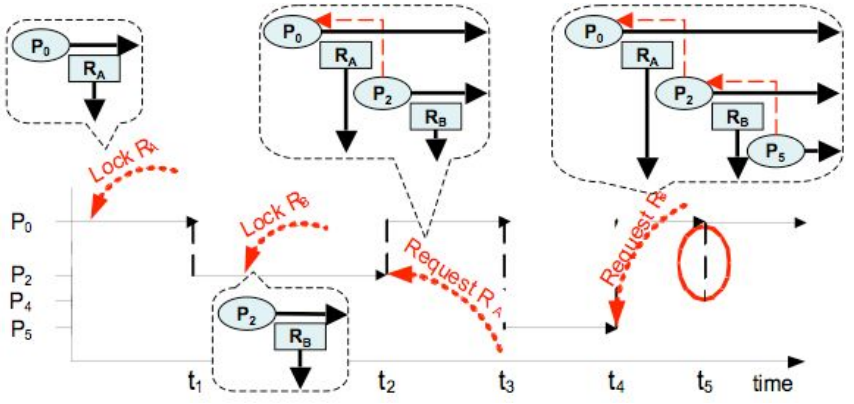


Fig. 2. A procedure when resources are locked and requested recursively

As can be seen, P_0 and P_2 lock R_A and R_B, respectively. At time t_2 when P_2 requests R_A, P_0 inherits the priority of P_2, and at time t_4 when P_5 requests R_B, P_2 inherits the priority of P_5. In the aspect of P_0, the priority of P_2 is changed, and so, that of P_0 is also changed. In result, at time t_5 when P_4 starts to run, P_0 is not preempted by P_4 because P_0 inherits the priority of P_5.

Now, we test and validate our proposed RPI protocol. The priorities of processes are changed using *setpriority()* system call which uses internally *set_user_nice()* kernel function, and *P()*/*V()* operations are performed using *semget()*/*semop()* system calls, respectively. We assume that the respective priorities of P_0, P_2, and P_5 are 10, 0, and -10 in all our tests; the priority of P_0 is the lowest while that of P_5 is the highest. In order to customize a general Linux kernel to a real-time operating system, uCLinux patch is applied to the Linux kernel of version 2.6.12 [8]. In this real-time environment, we define a resource as a semaphore and 'lock' (resource request)/'unlock' (resource release) as *P()*/*V()* operations of the semaphore, respectively. In addition, although we can directly modify the priorities of the processes in the system, we used the method that modifies the priorities of the processes using *set_user_nice()* kernel function, which modifies the *nice* value of a process; the priorities of processes may be modified by other factors, such as scheduling policy.

Comparison criteria are categorized into whether to solve the priority inversion problems or not and the time/space complexity for priority inheritance. (n and m are the number of processes and resources, respectively) (see Table 1)

As can be seen, our proposed RPI protocol can solve both the basic priority inversion problem and the complicated problems, however, incurs additional expense because the recursive data structure of each process is realigned whenever a process requests or releases a resource. Therefore, the time complexity for priority inheritance is *O(n+m)*. Additionally, each process maintains a recursive data structure for the processes or the resources involved, so the space complexity is also *O(n+m)*. In result, our proposed RPI protocol uses additional data

Table 1. Comparisons of the RPI protocol with BPI/PCE protocols

		BPI protocol	PCE protocol	Proposed protocol
Basic problem		O	O	O
Complicated problems	Case I [a]	X	O	O
	Case II [b]	X	X	O
Complexity	Time	*O(n)*	*O(nm)*	*O(n+m)*
	Space	*O(1)*	*O(m)*	*O(n+m)*

[a] Case I: when a process locks several resources
[b] Case II: when resources are locked and requested recursively

structure than both the previous two protocols and takes additional time than BPI protocol. However, our protocol can solve complicated priority inversion problems which previous protocols could not have solved as well as basic one.

4 Conclusion

Various priority inversion problems are addressed to design and implement real-time operating systems, where each task must finish execution by its deadline. In this paper, RPI protocol was proposed, which solves basic or complicated problems while pre-existing protocols cannot solve the complicated problems completely. The proposed protocol is implemented in the Linux kernel and validated with some test cases. Additionally, the performance of the proposed system is analyzed and compared with the previous recommended protocols.

References

1. TimeSys Inc., Priority Inversion: Why You Care and What to Do About It. A White Paper, 2004.
2. L. Sha, R. Rajkumar, and J. P. Lehoczky, "Priority Inheritance Protocols: An Approach to Real-Time Synchronization," IEEE Transactions on Computers, Vol. 39 No. 9, Sep. 1990.
3. B. Akgul, V. Mooney, H. Thane, and P. Kuacharoen, "Hardware Support for Priority Inheritance," Proc. of the IEEE Real-Time Systems Symp., Dec. 2003.
4. D. Zöbel, D. Polock, and A. Van Arkel, "Testing for the Conformance of Real-time Protocols Implemented by Operating Systems," Electronic Notes in Theoretical Computer Science, Vol. 133, May 2005.
5. J. B. Goodenough and L. Sha, "The Priority Ceiling Protocol," Special Report CMU/SEI-88-SR-4, Mar. 1998.
6. B. Dutertre, "Formal Analysis of the Priority Ceiling Protocol," Proc. of IEEE Real-Time Systems Symp., Nov. 2000.
7. V. Yodaiken, "Against Priority Inheritance, " FSMLABS Technical Paper, 2003.
8. Embedded Linux/Microcontroller Project, http://www.uclinux.org.

An Improved Simplex-Genetic Method to Solve Hard Linear Programming Problems

Juan Frausto-Solís and Alma Nieto-Yáñez

ITESM Campus Cuernavaca
Reforma 182-A, Col Lomas De Cuernavaca,
62589, Temixco Morelos, México
{juan.frausto,delia.nieto}@itesm.mx

Abstract. Linear programming (LP) is an important field of optimization. Even though, interior point methods are polynomial algorithms, many LP practical problems are solved more efficiently by the primal and dual revised simplex methods (RSM); however, RSM has a poor performance in hard LP problems (HLPP) as in the Klee-Minty Cubes problem. Among LP methods, the hybrid method known as Simplex-Genetic (SG) is very robust to solve HLPP. The objective of SG is to obtain the optimal solution of a HLPP, taking advantages from each one of the combined methods -a genetic algorithm (GA) and the classical primal RSM-. In this paper a new SG method named Improved Simplex Genetic Method (ISG) is presented. ISG combines a GA (with special genetic operators) with both primal and dual RSM. Numerical experimentation using some instances of the Klee-Minty cubes problem shows that ISG has a better performance than both RSM and SG.

Keywords: Genetic algorithm, Simplex method, Hybrid methods, Simplex-Genetic, Linear Programming.

1 Introduction

Linear programming problems (LPPs) have a wide range of real applications on diverse research areas [1], for example production planning, market analysis, routing, among others [2]. The development of the Simplex Method (SM) to solve LPPs in 1947 marks the start of the modern era in optimization; since then many algorithms, mainly interior point algorithms [3][4] have been developed. SM is a deterministic method and guarantees to find the optimal solution; however, for some LPPs, SM is not efficient and has poor worst-case behavior. Even though SM is an exponential algorithm, it is still the most popular and efficient method for many practical problems [5][6]. The most popular SM implementations are the primal and dual revised simplex methods (RSM). On other hand, genetic algorithms (GAs) are global search methods based on natural evolution and genetics. GAs have the property of maintaining a population of potential solutions using a selection process based on the fitness of each individual, and its evolution is made using genetic operators. GAs do not have optimality conditions, therefore they are not able to know when an optimal solution is found.

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 981–988, 2007.
© Springer-Verlag Berlin Heidelberg 2007

The hybrid approach Simplex-Genetic (SG) presented in [7], takes advantages of both methods and combines them with the goal of solving an LPP more efficiently. In [7] a GA is combined with only the primal version of RSM. SG was tested with hard linear programming problems (HLPP) known as Klee-Minty Cubes.

In this paper a new method named Improved Simplex Genetic (ISG) is presented. The principal characteristics of ISG are: i) the combination of a GA with both the primal and dual RSM, ii) the LPP representation is made for both bounded and unbounded variables and iii) new genetic operators are used in order to promote feasibility and optimality searching features. Numerical experimentation using some instances of Klee-Minty Cubes problems (the number of instances is limited by the IEEE infinitum definition [8]) shows that ISG has better performance than both RSM and SG.

This paper is organized as follows: In section 2, LPPs and RSM are briefly described, while the principal characteristics of a classical GA are presented in section 3; in section 4 SG is reviewed; in section 5 the new hybrid algorithm ISG is presented; finally in section 6, numerical experimentation with Klee-Minty Cubes is presented; the conclusion and future works are then given in section 7.

2 Linear Programming Problem and the Simplex Method

Linear Programming deals with linear mathematical models with the goal of optimizing a linear objective function subject to linear constraints. A LPP can be modeled as:

$$\begin{aligned} &\begin{bmatrix}\text{Max}\\ \text{Min}\end{bmatrix} && \sum_{j=1}^{n} c_{ij}x_j \\ &\text{subject to} && \sum_{j=1}^{n} a_{ij}x_j = b_i \qquad (i = 1,2,...,m) \\ & && l_j \le x_j \le u_j \qquad (j = 1,2,...,n) \end{aligned} \quad . \tag{1}$$

In (1), n is the number of variables, m is the number of constraints, x is the vector of variables to be solved, l_j and u_j represent the lower and the upper bound of the variable x_j, c is the cost vector, each equation $a_{ij}x_j=b$ is a constraint and the expression "cx" is the objective function.

Many methods have been proposed to solve LPPs. Even though SM was proposed by Dantzig since 1947, it is the most efficient for many practical problems [6]. As it is well known, the fundamental theorem of LP establishes that the optimal solution can only occur at a boundary point of the feasible region; therefore, SM is based on a search of solutions over the extreme points of a LPP. The SM first identifies an initial feasible solution and with this solution the initial tableau is built. The SM constructs a sequence of tableaux until the optimal solution is achieved. In every iteration, a basic variable is selected to leave the basis and one non-basic variable is selected to enter.

Revised Simplex Method (RSM) is the more popular SM implementation. All RSM variants [9] for solving model (1) are similar and the principal difference is the criterion of selecting the entering and leaving variables. Dual Simplex Method (DSM)

is another SM variant. DSM is based on the duality concept and was designed by Lemke in 1954 [10]. It also starts with a simplex tableau that defines an optimal but infeasible solution, and their objective is to move toward feasibility without loss of optimality [5].

SM behavior can be improved by changing: i) the pivoting form, that is, the way of selecting the entering and leaving variables [11][12]; and ii) the initial solution. A good initial solution is the key for efficient algorithms and so several alternatives are available: i) SM classical strategies, where the two-phase method and the big M method are the most popular and ii) Cosine Simplex Method (CSM) [13]; the latter uses a heuristic named Theta Heuristic [14] to determine the initial vertex based on the analysis of the angles between the gradient of the objective function and the gradients of the constraints. The initial solution can be classified in four cases [14]: Case I, the initial solution is feasible and optimal; case II, the initial solution is feasible but not optimal; case III, the solution is optimal but infeasible; and case IV, the solution is neither feasible nor optimal.

3 Genetic Algorithms

Genetic algorithms (GAs) are global search algorithms that are based on genetic principles and on the evolution [15]. The GAs were proposed by Holland in 1975. Since then, new applications have emerged in a wide variety of fields like scheduling, data fitting, and many combinatorial optimization problems [16]. A GA can be defined through the following components [15]: a genetic representation of the problem, a manner of generating an initial population, an evaluation function, genetic operators and parameters values (population size, probabilities of applying genetic operators, etc.). The performance of the GAs can be improved by an adequate selection of these components, parallel implementations and hybridization with other heuristic or exact algorithms.

4 Simplex-Genetic Method

In order to solve an LPP efficiently, hybrid methods can be designed; in [7] a Simplex-Genetic is presented; it combines a GA with the primal RSM and proves their efficiency using Klee-Minty Cubes; the hybridization is sequential. In the first phase (or genetic phase) a GA is used, while RSM is used in the second phase (or

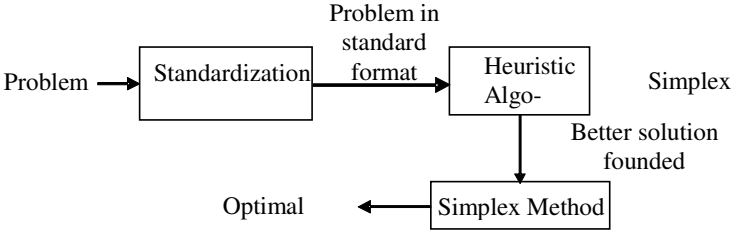


Fig. 1. Hybrid approach used to solve a LPP

simplex phase). RSM starts with the best feasible solution obtained in the first phase.Notice that in this hybrid algorithm the last solution of the genetic phase must be a feasible one, and then a primal problem should be solved by the simplex phase; so we rename this method as Primal Simplex-Genetic (PSG). The general scheme is shown in figure 1. Numerical results show that SG has better behavior than primal RSM [7].

The genetic phase of SG is characterized by [7]:

- Genetic representation of the solutions. Each string represents a basic solution for LPPs, that is, if an LPP has n variables, m constraints and h slack variables, the length of the string is $n+h$ and the string must have m basic variables. Any basic variable in a string is represented by a one and all the $(n+h-m)$ non-basic variables are represented by a zero.
- A tabu list, the purpose of which is to measure the feasibility of the variables during the process is used (see figure 2). The tabu list is built by the average of infeasibility of each variable in the population; in this case one variable is infeasible if the non-negative constraint is violated. Each position in the list represents one variable. If the variable was feasible in the previous population then the list has a zero, if not, the list has a positive number that measures how far from the feasibility region is the variable. In each generation this list is updated.

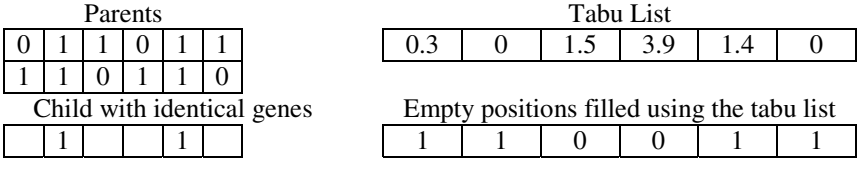


Fig. 2. Tabu aproach for crossover operator

- Two genetic operators: To maintain a population of basic solutions, special crossover and mutation operators are used. The crossover operator works with two parent solutions producing a child solution as is common in GA, but in this case two tasks are done: 1) identical bits in the parents are copied in the child and the others are empty sites; 2) the empty sites are filled using the tabu list, zeroes are set in the sites where the tabu list has a high value and the child must be a basic solution. In the figure 2 an example of this operator is shown. The mutation operator interchanges two different values selected at random and avoids non-basic solutions (i.e. only corners are permitted).

5 Improved Simplex-Genetic Method

In [17] different strategies to combine heuristics are presented. In this case we proposed a similar scheme used in [7] (and presented in figure 1). In this scheme the hybridization is sequential: GA in the first phase and RSM in the second one. Now in the genetic phase of ISG, a hybrid elitist GA is used; then, in the second phase, the

primal or dual RSM are used. Additionally, the genetic phase can be stopped with a feasible solution or with a non-feasible but optimal solution therefore the genetic phase stops until the improvement between one generation and the next is not significant (tends to zero).

The elitism included in the genetic phase consists only in keeping the best solution (feasible or non-feasible but optimal). Other improvements proposed in the genetic phase are as follows:

- Genetic representation of the solutions. It is important that the solution can represent basic solutions to LPPs. The majority of real LPPs have explicit bounds. Therefore an alternative representation is proposed. The length string is $(n+h)$ where n is the number of variables and h is the number of slack variables. The number of basic variables is equal to m, it is the number of constraints. In the string a 1 represents a basic variable; a 0 represents a non-basic variable in its lower bound while a non-basic variable in its upper bound is represented by -1.
- Generation of the initial population. This generation is random but the basic variables are added one by one verifying that they do not have dependent columns in the basis (matrix B) using LU factorization. Another approach used to generate the initial population is using the cosines calculated with the equations presented in [13][14], with the purpose of minimizing the number of individuals with dependent columns. Similar cosines form a cluster, and then basic variables are randomly selected, taken care that two basic variables do not belong to the same cluster. If there are not enough clusters to select the m variables required, new variables are randomly generated.
- Evaluation Function. The evaluation function is the objective function of the LPPs and the infeasibility is penalized adding this to the original objective function.
- Genetic Operators. New operators were proposed with the objective to promote feasibility and optimality.
  - Mutation promoting feasibility (Mut Feasible). Using a tabu list of infeasibility, selecting the basic variable more infeasible and interchange it by one non-basic variable randomly selected.
  - Mutation using the cosines (Mut Cosine). The cosines of the angles between the gradient of constraints and the gradient of objective function are calculated with the equations presented in [13][14] and according to the KKT in the optimal solution the cosine tends to be infinite, therefore the mutation pretends that the basic variables are those associated with the greater cosines. The mutation operator works as follows: first it selects a basic variable (randomly) and then interchanges it with the non-basic variable associated with the greater cosine.
  - These two operators are jointly used (Mut Fea-Cos), additionally the mutation operator presented in [7] is added. The results are shown in the next section.

6 Numerical Experimentation

In order to prove the performance of ISG the Klee-Minty Cubes problem [18] is used; as is generally known, RSM has an exponential behavior and the number of iterations required to solve this problem is 2^n-1 [5]. The model of this problem is:

$$\begin{aligned} &\text{Max} \quad \sum_{j=1}^{n} 10^{n-j} x_j \\ &\text{subject to} \quad \left(2 \sum_{j=1}^{i=1} 10^{i-j} \right) + x_i \leq 100^{i-1} \qquad i = 1,2,...,n \\ &\qquad\qquad x_j \geq 0 \qquad\qquad j = 1,2,...,n \end{aligned} \quad . \tag{2}$$

The instances used in this paper have a dimension equal to (n), the value of which is less than nineteen because for bigger values, the right hand side of some constraints is bigger than the IEEE definition of infinite [8]. The population size used here was determined by experimentation and its resulting value was two times the chromosome size; in this paper we use this value in both SG and ISG. The ISG is compared with SG and with RSM. SG combines the operators described in section 4 (with 80% of mutations and 20% of crossovers), the GA phase stops when the improvement between one generation and the next one is not significant and the best solution is feasible. The mutation operators (described in section 5 for ISG) are used first separately and then combined in a second experiment (50% each). The experimentation was made in a computer with an Intel® Pentium® M processor 1.60GHz 590MHz and 503MB of RAM. In figure 3, the comparison between the SG versus ISG with the proposed genetic operators is presented (using 30 executions in average); we notice that the operators using cosines have better performance and this improvement increases as the problem size increases. In the figure 4 the comparison between the cosines operator and RSM starting with the slacks as basic variables is shown.

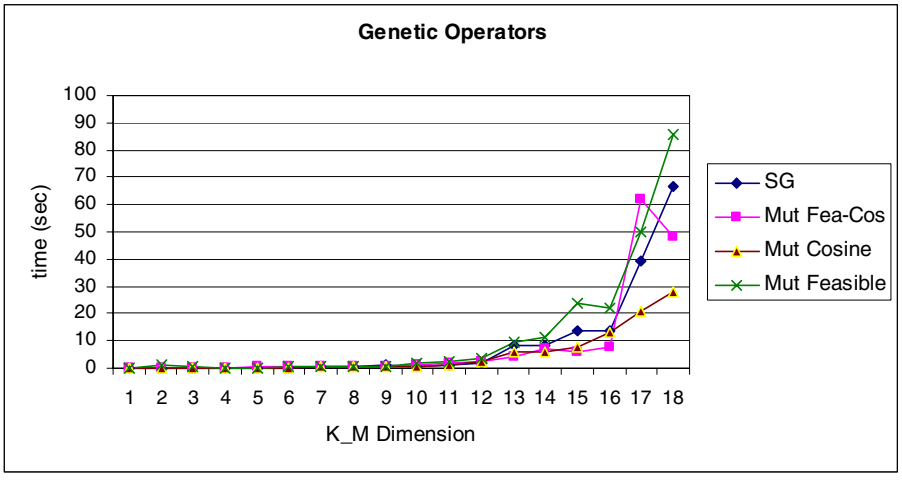


Fig. 3. Comparison of SG and the new operators in ISG

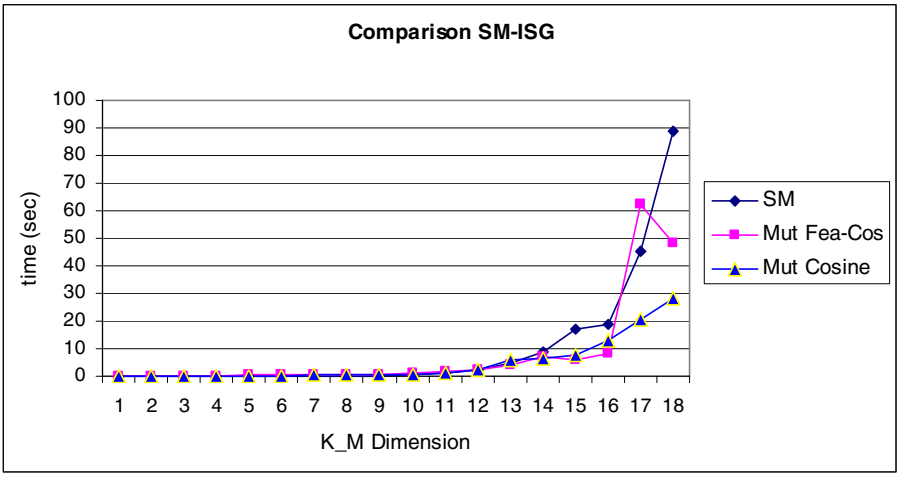


Fig. 4. Comparison of SM ant the new operators in ISG

7 Conclusions and Future Works

In this paper an Improved Simplex Genetic algorithm (ISG) is presented, which includes new genetic operators. These genetic operators use the cosine of the angles between the constraints and the objective function and use a list of infeasibility of each variable updated from previous solutions with the objective to promote optimality (using cosine values) and to avoid infeasibility (using a tabu list of infeasibility). The representation proposed here allows solving LPPs with bounded variables. The best solution obtained in the genetic phase of ISG is a basic solution for both, RSM or DSM. ISG, SG and RSM always finish with the optimal solution (if there is any); therefore, the comparison is made using only execution time. Experimentation presented in this paper for Klee-Minty Cubes problems showed that ISG has better performance than both RSM and the classical Simplex Genetic algorithms.

References

1. McMillan, C. Jr.: Mathematical Programming. John Wiley and Sons, Inc., New York (1970)
2. ILOG Optimization: http://www.ilog.com/products/optimization/industry/index.cfm. Last access: (2007)
3. Adler, I., Karmarkar, N., Resende, M., Veiga, G.: An Implementation of Karmarkar's Algorithm for Linear Programming. Mathematical Programming Vol. 44 (1989) 297-335
4. Mitra, G., Tamiz, M.: Experimental Investigation of an Interior Search Method within a Simplex Framework. Communications of the ACM, Vol. 31, Issue 12 (1988) 1474-1482
5. Chvátal, V.: Linear Programming. W. H. Freeman and Company, New York (1983)
6. Garey, M., Johnson, D.: Computers and intractability. A guide to the theory of NP-Completeness. Nineteenth printing, W.H. Freeman and Company, New York (1997)

7. Frausto-Solís, J., Rivera, R., Ramos, F.: A Simplex-Genetic method for solving the Klee-Minty cube. WSEAS Transactions on Systems Vol. 2, No. 1 (2002) 232-237
8. IEEE: Standard 754 for Binary Floating-Point Arithmetic. Lecture Notes on the Status of the IEEE 754 (1997)
9. Morgan, S.: A Comparison of Simplex Method Algorithms. Master Thesis, U. Florida (1997)
10. Lemke, C.: The dual method of solving the lineal programming problem. Naval Research Logistic Quarterly (1954)
11. Terlaky, T., Zhang, S.: A survey on pivot rules for linear programming. Delft University of Technology, Report No. 91-99, ISSN 0922-5641 (1991)
12. Bland, R.: New finite pivoting rules for the simplex method. Mathematics on Operations Research Vol. 2 (1977) 103-107
13. Trigos, F., Frausto-Solís, J., Rivera, R.: A Simplex Cosine Method for Solving the Klee-Minty Cube, Advances in Simulation, System Theory and Systems Engineering, WSEAS Press, ISBN 960852 70X (2002) 27-32
14. Trigos, F., Frausto-Solís, J.: Experimental Evaluation of the Theta Heuristic for Starting the Cosine Simplex Method. International Conference on Computational Science and its Applications, Singapore, ISBN 981-05-3498-1 (2005)
15. Michalewicz, Z.: Genetic Algorithms + Data Structures= Evolution programs. Third Edition, Springer-Verlag, Berlin Heidelberg New York (1996)
16. Illinois Genetic Algorithms Laboratory: http://www-illigal.ge.uiuc.edu/index.php3, Director David Goldberg, last access: April (2006)
17. Puchinger, J., Günther R.: Combining Metaheuristics and Exact Algorithms in Combinatorial Optimization: A Survey and Classification. In Proceedings of the First International Work-Conference on the Interplay Between Natural and Artificial Computation, Part II, Vol. 3562 of LNCS, Springer (2005) 41-53
18. Klee, V., Minty, G.: How good is the Simplex Algorithm? In Inequalities III, Shisha, O., Ed., Academic Press, New York (1972) 159-179

Real-Observation Quantum-Inspired Evolutionary Algorithm for a Class of Numerical Optimization Problems*

Gexiang Zhang and Haina Rong

School of Electrical Engineering, Southwest Jiaotong University,
Chengdu 610031 Sichuan, China
gxzhang@ieee.org

Abstract. This paper proposes a real-observation quantum-inspired evolutionary algorithm (RQEA) to solve a class of globally numerical optimization problems with continuous variables. By introducing a real observation and an evolutionary strategy, suitable for real optimization problems, based on the concept of Q-bit phase, RQEA uses a Q-gate to drive the individuals toward better solutions and eventually toward a single state corresponding to a real number varying between 0 and 1. Experimental results show that RQEA is able to find optimal or close-to-optimal solutions, and is more powerful than conventional real-coded genetic algorithm in terms of fitness, convergence and robustness.

Keywords: Evolutionary computation, quantum-inspired evolutionary algorithm, real observation, numerical optimization.

1 Introduction

Quantum-inspired evolutionary algorithm (QEA) is an unconventional algorithm of evolutionary computation. QEA inherits the structure and probabilistic search way of conventional genetic algorithm (CGA), and some concepts and operations of quantum computing, such as quantum-inspired bit (Q-bit), quantum-inspired gate (Q-gate) and quantum operators including superposition, entanglement, interference and measurement [1,2]. Up to now, as a better optimization method than CGA, QEA has been used in several applications of knapsack problem [2,3], digital filter design [4], feature selection [1]. Extensively experimental results manifest its advantages of good global search capability, rapid convergence and speediness [1-4].

In the existing QEA, only binary strings can be obtained by observing the probability amplitudes of Q-bits. Accordingly, the evolutionary strategy (update strategy of the rotation angles of Q-gates) was derived from a class of combinatorial optimization problems and was represented with binary code. The QEA in the existing literature is called binary-observation QEA (BQEA). Like binary-coded CGA, BQEA suffers from several disadvantages when it involves real

* This work was supported by the Scientific and Technological Development Foundation of Southwest Jiaotong University under the grant No.2006A09.

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 989–996, 2007.
© Springer-Verlag Berlin Heidelberg 2007

number optimization problems [5-7]. First of all, there is Hamming Cliff when real variables are encoded as binary strings. The Hamming distances exist between the binary codes of adjacent integers. For example, the integers 31 and 32 are represented respectively with binary codes 011111 and 100000, which have a Hamming distance of 6. To improve the code of 31 to that of 32, BQEA must alter all bits simultaneously, which is a difficult problem. Then, in encoding real number as binary strings, discretization error will inevitably be introduced in BQEA that cannot operate on a continuous space but on an evenly discretized space. Discretization error comes from the discrepancy between the binary representation space and a real space. Two points close to each other in a real space might be very far in the binary representation space. Finally, the encoding and decoding operations make BQEA more computationally expensive because the binary chromosome will have a huge string length when binary substring representing each real parameter with the desired precision are concatenated to form a chromosome. What is more, Han [2] also made it clear that the representation of real number may be more suitable for numerical optimization than that of binary string.

To overcome the drawbacks of BQEA, this paper proposes a real-observation QGA (RQEA) which is more suitable than BQEA for a wide range of real-world numerical optimization problems. Experiments on several functions are carried out to verify the effectiveness. Experimental results show that RQEA is able to find optimal or close-to-optimal solutions, and is more powerful than conventional real-coded genetic algorithm (CRGA) in terms of fitness, convergence and robustness.

2 RQEA

Quantum mechanical system is a probabilistic system. Like a classical probabilistic system, the probabilities of each state need be specified to describe the behavior of quantum mechanical system [8]. A quantum state vector $|\Psi\rangle$ can best be used to describe the location of a quantum particle and a weighted sum which in the case of two possible locations A and B equals $\alpha|A\rangle + \beta|B\rangle$, where α and β are complex number weighting factors of the particle being in locations A and B, respectively, and where $\alpha|A\rangle$ and $\beta|B\rangle$ are themselves state vectors [9]. Each two state quantum system is referred to as a Q-bit, which is also the smallest information unit in a two-state quantum computer [8]. The quantum state $|\Psi\rangle$ may be in the A state, in the B state, or in any superposition of the two. The quantum state $|\Psi\rangle$ can be represented as

$$|\Psi\rangle = \alpha|A\rangle + \beta|B\rangle \ . \tag{1}$$

where α and β satisfy the normalization equality

$$|\alpha|^2 + |\beta|^2 = 1 \ . \tag{2}$$

where $|\alpha|^2$ and $|\beta|^2$ are the probabilities that the Q-bit will be observed in A state, in the B state, respectively, in the act of observing the quantum state.

In this paper, states A and B are not considered as only the states 1 and 0, respectively, but an arbitrary pair of states between 1 and 0, which satisfy (1) and (2).

For a quantum system with n Q-bits, there will be n quantum states and 2^n information states. Of course, n different pairs of complex number weighting factors are needed to describe the quantum system and each describes the weighted probability of the particle being at that location [9]. A basic operation in quantum computing is that of a *fair coin flip* performed on a single Q-bit whose states are A and B. In an 2^n information state quantum system, this operation is performed on each Q-bit independently and can change the state of each Q-bit [8]. Through this operation, a Q-bit in the state A or B can be transformed into a superposition of the two states. In RQEA, this operation is implemented by using a Q-gate. If there are n locations as given by n state vectors, the particle is said to be all n locations at the same time.

Quantum mechanical systems have a deeper structure, and consequently, besides having a certain probability of being in each state, they also have a phase associated with each state [8]. In quantum computing, four quantum operators, superposition, entanglement, interference and measurement, are mainly used [10]. Superposition operator is applied for junction of possible solution spaces in a single unified solution space. Entanglement operator is employed to search the optimal solution as an unknown marked state. Interference and measurement operators are applied to extract the marked state with highest probability.

Instead of numeric, binary or symbol representation, Q-bit representation is used to represent the individuals of population in RQEA [1-4]. The probability amplitude of a Q-bit is defined firstly.

Definition 1. *The probability amplitude of a Q-bit is defined by a pair of numbers* (α, β) *as*

$$[\alpha \ \ \beta]^T \ . \tag{3}$$

where α *and* β *satisfy normalization equality (2).* $|\alpha|^2$ *and* $|\beta|^2$ *denote the probabilities that the qubit will be found in A state and in B state in the act of observing the quantum state, respectively. Note that in general, the probability amplitudes can be complex quantities. However, in this paper or in RQEA, we only need real amplitudes with either positive or negative signs.*

For quantum systems, in addition to having a certain probability of being in each state, they also have a phase associated with each state. The definition of Q-bit phase is given in the following.

Definition 2. *The phase of a Q-bit is defined with an angle* ξ *as*

$$\xi = \arctan(\beta/\alpha) \ . \tag{4}$$

where $\xi \in [-\pi/2, \pi/2]$. *The sign of Q-bit phase* ξ *indicates which quadrant the Q-bit lies in. If* ξ *is positive, the Q-bit is regarded as being in the first or third quadrant, otherwise, the Q-bit lies in the second or fourth quadrant.*

According to Def.1, the probability amplitudes of n Q-bits are represented as

$$\begin{bmatrix} \alpha_1|\alpha_2|\cdots|\alpha_n \\ \beta_1|\beta_2|\cdots|\beta_n \end{bmatrix} \quad . \tag{5}$$

where $|\alpha_i|^2 + |\beta_i|^2 = 1, i = 1, 2, \cdots, n$. The phase of the i^{th} Q-bit is

$$\xi_i = \arctan(\beta_i/\alpha_i) \; . \tag{6}$$

Q-bit representation can represent a linear superposition of states probabilistically. As shown in (5), n Q-bits are able to represent a linear superposition of 2^n states. Different from CGA, Q-bit representation makes it possible that four main quantum operators including superposition, entanglement, interference and measurement are implemented. So Q-bit representation is greatly superior to other representations in population diversity. This is what distinguishes RQEA from CGA.

According to the above Q-bit representation, the structure of RQEA is described as Algorithm 1, in which each step is explained briefly as follows.

Algorithm 1. Algorithm of RQEA

```
   Begin
(1)  Set initial values of parameters; % Evolutionary generation g=0;
(2)  Initialize P(g); %
     While (not termination condition) do
       g=g+1;
(3)    Generate R(g) by observing P(g-1); %
(4)    Evaluate R(g); %
(5)    Store the best solution among R(g) and B(g-1) into B(g);
(6)    Update P(g) using Q-gates; %
(7)    Migration operation;
       If (catastrophe condition)
(8)     Catastrophe operation;
       End if
     End
   End
```

(1) Population size n_p, the number n_v of variables and the initial evolutionary generation g need be set.

(2) In this step, population $\boldsymbol{P}(g)$=$\{\boldsymbol{p}_1^g, \boldsymbol{p}_2^g, \cdots, \boldsymbol{p}_{n_p}^g\}$, where $\boldsymbol{p}_i^g (i = 1, 2, \cdots, n_p)$ is an arbitrary individual in population $\boldsymbol{P}(g)$ and $\boldsymbol{p}_i^g$ is represented as

$$\boldsymbol{p}_i^g = \begin{bmatrix} \alpha_{i1}^g|\alpha_{i2}^g|\cdots|\alpha_{in_v}^g \\ \beta_{i1}^g|\beta_{i2}^g|\cdots|\beta_{in_v}^g \end{bmatrix} \quad . \tag{7}$$

where $\alpha_{ij}^g = \beta_{ij}^g = 1/\sqrt{2}$ $(j = 1, 2, \cdots, n_v)$, which means that all states are superposed with the same probability.

Table 1. Look-up table of function $f(\alpha, \beta)$ (Sign is a symbolic function)

$\xi_1 > 0$	$\xi_2 > 0$	$f(\alpha, \beta)$	
		$\xi_1 \geq \xi_2$	$\xi_1 < \xi_2$
True	True	+1	-1
True	False	$\text{sign}(\alpha_1 \cdot \alpha_2)$	
False	True	$\text{-sign}(\alpha_1 \cdot \alpha_2)$	
False	False	$\text{sign}(\alpha_1 \cdot \alpha_2)$	$\text{-sign}(\alpha_1 \cdot \alpha_2)$
$\xi_1, \xi_2 = 0$ or $\pi/2$		± 1	

(3) According to probability amplitudes of all individuals in $\boldsymbol{P}(g-1)$, observed states $\boldsymbol{R}(g)$ is constructed by observing $\boldsymbol{P}(g-1)$. Here $\boldsymbol{R}(g)$=$\{\boldsymbol{a}_1^g, \boldsymbol{a}_2^g, \cdots, \boldsymbol{a}_{n_p}^g\}$, where $\boldsymbol{a}_i^g (i = 1, 2, \cdots, n_p)$ is an observed state of an individual $\boldsymbol{p}_i^{g-1}$ $(i = 1, 2, \cdots, n_p)$. $\boldsymbol{a}_i^g$ is a real number with the dimension n_v, that is $\boldsymbol{a}_i^g = b_1 b_2 \cdots b_{n_v}$, where b_j $(j = 1, 2, \cdots, n_v)$ is a real number between 0 and 1. Observed states $\boldsymbol{R}(g)$ is generated in probabilistic way. For the probability amplitude $[\alpha\ \beta]^T$ of a Q-bit, a random number r in the range [0, 1] is generated. If $r < |\alpha|^2$, the corresponding observed value is set to $|\alpha|^2$, otherwise, the value is set to $|\beta|^2$.

(4) The fitness are calculated by using the obtained real parameter values.

(5) The best solution are stored into $B(g)$.

(6) In this step, the probability amplitudes of all Q-bits in population $\boldsymbol{P}(g)$ are updated by using Q-gates given in (8).

$$\boldsymbol{G} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \quad . \tag{8}$$

where θ is the rotation angle of Q-gate and θ is defined as $\theta = k \cdot f(\alpha, \beta)$, where k is chosen as

$$k = 0.5\pi e^{-\frac{mod(g,100)}{10}} \quad . \tag{9}$$

and $f(\alpha, \beta)$ is obtained by using the look-up table shown in Table 1, in which $\xi_1 = \arctan(\beta_1/\alpha_1)$ and $\xi_2 = \arctan(\beta_2/\alpha_2)$, where α_1, β_1 are the probability amplitudes of the best solution stored in $B(g)$ and α_2, β_2 are the probability amplitudes of the current solution.

(7) Within an individual, the probability amplitudes of one Q-bit are migrated to those of another Q-bit, i.e.

$$\begin{bmatrix} \alpha_{11}| \leftrightarrow \alpha_{12}| \leftrightarrow \cdots | \leftrightarrow |\alpha_{1i}| \leftrightarrow \cdots | \leftrightarrow \alpha_{in_v} \\ \beta_{11}| \leftrightarrow \beta_{12}| \leftrightarrow \cdots | \leftrightarrow |\alpha_{1i}| \leftrightarrow \cdots | \leftrightarrow \beta_{in_v} \end{bmatrix}$$

(8) The catastrophe condition is a prescribed generation C_g, such as 10 or 20.

3 Experiments

To test the effectiveness of RQEA, 13 functions $f_1 - f_{13}$ are used to bring into comparison with CGA [5-7]. The evolutionary strategies of CRGA include elitism

selection, uniform crossover and uniform mutation. The crossover and mutation probabilities are set to 0.8 and 0.1, respectively. RQEA and CRGA use the same population size 20 and the same maximal generation 500. The parameter C_g is set to 20 in RQEA. RQEA and CRGA are performed 50 independent runs for each test function, respectively. The mean best values and the standard deviations are recorded for each test function. Experimental results are listed in Table 2, in which m, σ, g and p represent the mean best, the standard deviation, the maximal number of generations and the population size, respectively. The results are averaged over 50 runs. From Table 2, it can be seen that RQEA obtains far better results than CRGA in terms of both the mean best solutions and the standard deviations for all of the test functions.

(I) Sphere function

$$f_1(\mathbf{x}) = \sum_{i=1}^{N} x_i^2, \ \ -100.0 \le x_i \le 100.0, \ N = 30 \, . \tag{10}$$

(II) Ackley function

$$f_2(\mathbf{x}) = 20 + e - 20 \exp\left(-0.2\sqrt{\frac{1}{N}\sum_{i=1}^{N} x_i^2}\right) - \exp\left(\frac{1}{N}\sum_{i=1}^{N} \cos(2\pi x_i)\right) \, .$$
$$-32.0 \le x_i \le 32.0, \quad N = 30 \tag{11}$$

(III) Griewank function

$$f_3(\mathbf{x}) = \frac{1}{4000}\sum_{i=1}^{N} x_i^2 - \prod_{i=1}^{N}\left(\frac{x_i}{\sqrt{i}}\right) + 1, \ \ -600.0 \le x_i \le 600.0, \ N = 30 \, . \tag{12}$$

(IV) Rastrigin function

$$f_4(\mathbf{x}) = 10N + \sum_{i=1}^{N} (x_i^2 - 10\cos(2\pi x_i)), \ \ -5.12 \le x_i \le 5.12, \ N = 30 \, . \tag{13}$$

(V) Schwefel function

$$f_5(\mathbf{x}) = 418.9829N - \sum_{i=1}^{N} \left(x_i \sin\left(\sqrt{|x_i|}\right)\right), \ \ -500.0 \le x_i \le 500.0, \ N = 30 \, . \tag{14}$$

(VI) Schwefel's problem 2.22

$$f_6(\mathbf{x}) = \sum_{i=1}^{N} |x_i| + \prod_{i=1}^{N} |x_i|, \ \ -10 \le x_i \le 10, \ N = 30 \, . \tag{15}$$

(VII) Schwefel's problem 1.2

$$f_7(\mathbf{x}) = \sum_{i=1}^{N} \left(\sum_{j=1}^{i} x_j\right)^2, \ \ -100 \le x_j \le 100, \ N = 30 \, . \tag{16}$$

Table 2. Comparisons of RQEA and CRGA

			RQEA		CRGA		Global
	g	p	m	σ	m	σ	minimum
f_1	500	20	1.11×10^{-7}	2.19×10^{-7}	1.55×10^4	2.20×10^3	0
f_2	500	20	2.62×10^{-4}	1.29×10^{-3}	15.67567	0.478948	0
f_3	500	20	1.50×10^{-6}	5.63×10^{-6}	141.9507	15.16764	0
f_4	500	20	1.44×10^{-7}	3.21×10^{-7}	239.6483	12.61521	0
f_5	500	20	0.194603	0.684961	9.80×10^3	3.21×10^2	0
f_6	500	20	1.78×10^{-4}	2.82×10^{-4}	1.29×10^9	2.32×10^9	0
f_7	500	20	3.07×10^{-6}	8.15×10^{-6}	6.29×10^4	1.74×10^4	0
f_8	500	20	6.02×10^{-5}	7.34×10^{-5}	39.63123	2.649442	0
f_9	500	20	0	0	1.60×10^4	1.51×10^4	0
f_{10}	500	20	2.06×10^{-3}	1.92×10^{-3}	12.15373	3.226462	0
f_{11}	500	20	0.998004	1.82×10^{-10}	1.578839	0.658432	≈ 1
f_{12}	500	20	-1.031628	4.42×10^{-9}	-0.966542	0.069581	-1.031628
f_{13}	500	20	0.397904	8.63×10^{-5}	0.525740	0.169221	0.397877

(VIII) Schwefel's problem 2.21

$$f_8(\mathbf{x}) = \max_{i=1}\{|x_i|, 1 \le i \le 30\}, \ \ -100 \le x_i \le 100\ . \tag{17}$$

(IX) Step function

$$f_9(\mathbf{x}) = \sum_{i=1}^{N} (\lfloor x_i + 0.5 \rfloor)^2, \ \ -100 \le x_i \le 100\ . \tag{18}$$

(X) Quartic function, i.e. noise

$$f_{10}(\mathbf{x}) = \sum_{i=1}^{N} i x_i^4 + random[0, 1), \ \ -1.28 \le x_i \le 1.28\ . \tag{19}$$

(XI) Shekel's Foxholes function

$$f_{11}(\mathbf{x}) = \left[\frac{1}{500} + \sum_{j=1}^{25} \frac{1}{j + \sum_{i=1}^{2}(x_i - a_{ij})^6}\right]^{-1}, \ \ -10 \le x_j \le 10\ . \tag{20}$$

where $(a_{ij}) = \begin{pmatrix} -32 & -16 & 0 & 16 & 32 & -32 & \cdots & 0 & 16 & 32 \\ -32 & -32 & -32 & -32 & -32 & -16 & \cdots & 32 & 32 & 32 \end{pmatrix}$.

(XII) Six-hump camel-back function

$$f_{12}(\mathbf{x}) = 4x_1^2 - 2.1x_1^4 + \frac{1}{3}x_1^6 + x_1x_2 - 4x_2^2 + 4x_2^4, \ \ -5 \le x_i \le 5\ . \tag{21}$$

(XIII) Branin function

$$f_{13}(\mathbf{x}) = \left(x_2 - \frac{5.1}{4\pi^2}x_1^2 + \frac{5}{\pi}x_1 - 6\right)^2 + 10\left(1 - \frac{1}{8\pi}\right)\cos(x_1) + 10 \quad . \qquad (22)$$
$$-5 \le x_1 \le 10, \quad 0 \le x_2 \le 15$$

4 Concluding Remarks

Extending two states 1 and 0 to an arbitrary pair of states between 1 and 0 in quantum system, this paper proposes RQEA to solve numerical optimization problems. RQEA can be considered as the extensive version of BQEA to real number solution space. Extensive experiments show that RQEA is a competitive algorithm for numerical optimization problems. Our future work will be concentrated on the applications of RQEA.

References

1. Zhang, G.X., Hu, L.Z., Jin, W.D.: Quantum Computing Based Machine Learning Method and Its Application in Radar Emitter Signal Recognition. In: Torra, V., Narukawa, Y., (eds.): Lecture Notes in Artificial Intelligence, Vol.3131. Springer-Verlag, Berlin Heidelberg New York (2004) 92-103
2. Han, K.H., Kim, J.H.: Quantum-Inspired Evolutionary Algorithms with a New Termination Criterion, H_ε Gate, and Two-Phase Scheme. IEEE Transactions on Evolutionary Computation **8** (2004) 156-169
3. Han, K.H., Kim, J.H.: Quantum-Inspired Evolutionary Algorithms for a Class of Combinatorial Optimization. IEEE Transactions on Evolutionary Computation **6** (2002) 580-593
4. Zhang, G.X., Jin, W.D., Li, N.: An Improved Quantum Genetic Algorithm and Its Application. In: Wang, G., et al., (eds.): Lecture Notes in Artificial Intelligence, Vol.2639. Springer-Verlag, Berlin Heidelberg New York (2003) 449-452
5. Oyama, A., Obayashi, S., Nakahashi, K.: Real-Coded Adaptive Range Genetic Algorithm and Its Application to Aerodynamic Design. International Journal of Japan Society of Mechanical Engineers, Series A **43** (2000) 124-129
6. Qing, A.Y., Lee, C.K., Jen, L.: Electromagnetic Inverse Scattering of Two-Dimensional Perfectly Conducting Objects by Real-Coded Genetic Algorithm. IEEE Transactions on Geoscience and Remote Sensing **39** (2001) 665-676
7. Wang, J.L., Tan, Y.J.: 2-D MT Inversion Using Genetic Algorithm. Journal of Physics: Conference Series **12** (2005) 165-170
8. Grover, L.K.: Quantum Computation. In: Proceedings of the 12th Int. Conf. on VLSI Design (1999) 548-553
9. Narayanan, A.: Quantum Computing for Beginners. In: Proc. of the 1999 Congress Evolutionary Computation (1999) 2231-2238
10. Ulyanov, S.V.: Quantum Soft Computing in Control Process Design: Quantum Genetic Algorithm and Quantum Neural Network Approaches. In: Proc. of World Automation Congress vol.17 (2004) 99-104

A Steep Thermodynamical Selection Rule for Evolutionary Algorithms

Weiqin Ying[1], Yuanxiang Li[1,2], Shujuan Peng[2], and Weiwu Wang[1]

[1] State Key Lab. of Software Engineering, Wuhan University, Wuhan 430072, China
[2] School of Computer Science, Wuhan University, Wuhan 430079, China
{weiqinying,yxli62}@yahoo.com.cn

Abstract. The genetic algorithm (GA) often suffers from the premature convergence because of the loss of population diversity at an early stage of searching. This paper proposes a steep thermodynamical evolutionary algorithm (STEA), which utilizes a steep thermodynamical selection (STS) rule. STEA simulates the competitive mechanism between energy and entropy in annealing to systematically resolve the conflicts between selective pressure and population diversity in GA. This paper also proves that the rule STS has the approximate steepest descent ability of the free energy. Experimental results show that STEA is both far more efficient and much stabler than the thermodynamical genetic algorithm (TDGA).

Keywords: Evolutionary algorithms, thermodynamics, selection rule, population diversity, free energy.

1 Introduction

The genetic algorithm (GA) is an optimization technique based on the mechanism of evolution by natural selection [1]. However, it has some disadvantages yet for solving large-scale combinatorial optimization problems because the astronomical size of search spaces with local optima often lead GA to extremely slow search and "premature convergence" [2]. The premature convergence also causes the low stability of GA. Whitley [3] argues that population diversity and selective pressure are the two primary factors in genetic search. Increasing selective pressure speeds up the search, while it also results in a faster loss of population diversity. Maintaining population diversity can help the search to escape local optima, while it offsets the effect of increasing selective pressure. These two factors are inversely related.

Some techniques on controlling population diversity have been proposed, such as scaling the fitness [1], sharing the fitness [1], and driving all individuals moving [4] etc. But they are not yet sufficiently systematical and effective for large-scale combinatorial problems. Kirkpatrick *et al.* [5] have proposed another general optimization algorithm called the simulated annealing (SA). SA controls search systematically by the cooling temperature and the Metropolis rule. Mori *et al.* [6,7] have proposed a method of combining SA and GA, called the thermodynamical genetic algorithm (TDGA). They introduce a greedy thermodynamical

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 997–1004, 2007.
© Springer-Verlag Berlin Heidelberg 2007

selection rule in TDGA, getting a hint from the principle of minimal free energy. TDGA attempts to utilize the concepts of temperature and entropy in annealing to control population diversity systematically. TDGA is effective to some extent, while its performance is still inadequately stable and its computational cost is extremely high.

This paper proposes a steep thermodynamical evolutionary algorithm (STEA) to improve the stability and the computational efficiency of TDGA. STEA simulates the competitive mechanism between energy and entropy in annealing to systematically resolve the conflicts between selective pressure and population diversity in GA. The measurement of population diversity and the minimization process of free energy at each temperature are mended and a steep thermodynamical selection (STS) rule is proposed in STEA. The paper is organized as follows. We briefly review the thermodynamic background on STEA in Section 2, describe the outline of STEA in Section 3, and prove the approximate steepest descend ability of the rule STS in Section 4. Finally, the experimental results are presented in Section 5 and the conclusion is given in Section 6.

2 Brief Thermodynamic Background on STEA

In thermodynamics and statistical mechanics, annealing can be viewed as an adaptation process to optimize the stability of the final crystalline solid. In an annealing process, a metal, initially at high temperature and disordered, is slowly cooled so that the system at any temperature approximately reaches thermodynamic equilibrium [5]. As cooling proceeds, the system becomes more ordered and approaches a "frozen" ground state at the temperature T=0. There are a few observations about annealing which are helpful for STEA:

1. If the initial temperature is too low or cooling is done insufficiently slowly, the system may become quenched forming defects or trapped in a local minimum energy state.
2. Any change from non-equilibrium to equilibrium of the system at each temperature follows the principle of minimum free energy. In other words, the system will change spontaneously to achieve a lower total free energy and the system reaches equilibrium when its free energy seeks a minimum. In thermodynamics, the free energy F is defined as $F = E - HT$, where E is the energy of the system and H its entropy.
3. In thermodynamics, the entropy can quantificationally measure the energy dispersal of the particles in the system.
4. Any change of the system can be viewed as a result of the competition between its energy and its entropy. The temperature T determines their relative weights in the competition.

3 Steep Thermodynamical Evolutionary Algorithm

There are some deep and useful similarities between annealing in solids and convergence in GA. The population and the individuals in GA can be regarded

as the system and the particles. Then the mean of negative fitness, the population diversity and a controllable weight parameter can play the roles of the energy, the entropy and the temperature respectively. Every population state exactly full of global optima can be interpreted as the ground state. This analogy provides an approach for STEA to simulate the competitive mechanism between energy and entropy in annealing to systematically resolve the conflicts between selective pressure and population diversity in GA.

3.1 Measurement of Population Diversity

It is a critical part how to measure population diversity when introducing the competitive mechanism into GA. TDGA uses the sum of the information entropy at each locus, called a gene-based entropy in this paper. However, it has two disadvantages. Firstly, its repetitive calculations at all loci cause high computational costs. Secondly, the thermodynamic entropy can measure the energy dispersal of particles that is equivalent to the fitness dispersal of individuals in the population, while the gene-based entropy in TDGA can't measure the fitness dispersal. In this section, we propose a level-based entropy by grading the fitness.

Definition 1. *Let S be the search space, $f : S \to \mathbb{R}$ be the objective function, and $X_r \in S$ be one individual. Then the individual energy $e(X_r) = f(X_r)$ for minimum problems and $e(X_r) = -f(X_r)$ for maximum problems. Assume e_u and e_l be respectively an upper bound and a lower bound of the individual energy. Then $\pi = \{g_i | 0 \le i \le K-1\}$ is called a level partition on $[e_l, e_u]$ if*

$$g_i = \left(\frac{2^{i-1}-1}{2^{K-1}-1}(e_u - e_l) + e_l, \frac{2^i - 1}{2^{K-1}-1}(e_u - e_l) + e_l\right) \cap [e_l, e_u]. \tag{1}$$

We shall say that X_r is at level g_i if $e(X_r) \in g_i$.

Definition 2. *Let $P = (X_1, X_2, \ldots, X_N) \in S^N$ be one population of size N, $\pi = \{g_i | 0 \le i \le K-1\}$ be a level partition, and n_i be the number of individuals in g_i of population P. Then $H(\pi, P)$ is called the level-based entropy of P for π where*

$$H(\pi, P) = -\sum_{i=0}^{K-1} \frac{n_i}{N} \log_K \frac{n_i}{N}, \ 0 \le i \le K-1. \tag{2}$$

The value of the level-based entropy varies from 0 to 1 depending on the level distribution of P. It measures the fitness dispersal with very low computational costs.

3.2 Minimization of Free Energy at Each Temperature

Definition 3. *For $P \in S^N$, $E(P)$ is called the population energy of P where*

$$E(P) = \frac{1}{N} \sum_{r=1}^{N} e(X_r), \ X_r \in P. \tag{3}$$

$F(\pi, T, P)$ is called the free energy of P at temperature T for partition π where

$$F(\pi, T, P) = E(P) - H(\pi, P)T. \tag{4}$$

The free energy is the driving force toward equilibrium in annealing. Similarly, we should force the population to minimize its free energy at each temperature T_k during evolution. However, there is only one generation of competition at each temperature T_k in TDGA. This insufficient competition only lowers the free energy very slightly, and the population can't reach the minimum free energy (or equilibrium) at T_k. In order to approach equilibrium, STEA holds L_k generations of competitions at T_k. These competitions at T_k form a $Markov$ chain of length L_k, where L_k should grow with the hardness of problems.

3.3 Steep Thermodynamical Selection Rule

In order to minimize the free energy rapidly at each temperature T_k, we should design a thermodynamical selection rule to descend the free energy of the next generation most steeply. Its mission is to select N individuals from N parent individuals and M offspring individuals as the next generation with the minimum free energy. However, It is infeasible to exactly minimize the free energy for each generation because of the extremely high complexity $O((N+K)C_{N+M}^{N})$. Hence, TDGA uses a greedy thermodynamical selection (GTS) rule with the complexity $O(N^2K)$. But its reliability can't be guaranteed. In this section, we proposes a steep thermodynamical selection (STS) rule by assigning the free energy of the population to its individuals.

Definition 4. *Let $P = (X_1, X_2, \ldots, X_N) \in S^N$ be one population of size N and $\pi = \{g_i | 0 \leq i \leq K-1\}$ be a level partition. For an individual $X_r \in P$ at level $g_d \in \pi$, its free energy component in P at temperature T for π is defined as:*

$$F_c(\pi, T, P, X_r) = e(X_r) + T\log_K(\frac{n_d}{N}), \tag{5}$$

where n_d is the number of individuals at g_d of P.

The steep selection rule $\boldsymbol{STS}(\pi, T, P_t, O_t)$ is described as follows:

1. Produce an interim population P'_t of size $N+M$ by appending M individuals in the offspring population O_t to the parent population P_t.
2. Calculate the free energy component $F_c(\pi, T, P'_t, X_r)$ for each individual $X_r \in P'_t$.
3. Pick the M individuals with the largest free energy components from P'_t.
4. Form the next generation P_{t+1} by removing these M individuals from P'_t.

STS has the lower complexity $O((N+M)M)$. It's proved in Section 4 that STS has the approximate steepest descent ability of the free energy.

3.4 Outline of STEA

Figure 1 provides the general outline of the whole STEA described above.

```
01 Create N individuals randomly as an initial population P_0 and evaluate them;
02 Determine the energy bounds e_u and e_l for a level partition π;
03 Configure the length L_k of the Markov chain at each temperature T_k;
04 T_0 = 10(e_u − e_l);
05 t = 0; k = 0;
06 while(Termination_test(P_t)==False)
07 {
08    for(i = 0; i < L_k; i++)
09    {
10       Generate M offspring by uniform selection, crossover and mutation;
11       Organize these M offspring as an offspring population O_t and evaluate them;
12       P_{t+1} = STS(π, T_k, P_t, O_t);
13       t = t + 1;
14    }
15    k = k + 1;
16    T_k = T_0/(1 + k);
17 }
```

Fig. 1. The outline of STEA

4 Approximate Steepest Descent Ability of STS

Lemma 1. *Assume P_{t+1}^G is the next generation population with the exact minimum free energy and P_{t+1} is the next generation generated by STS. Then*

$$\frac{1}{N}\sum_{X_r\in P_{t+1}} F_c(\pi,T_k,P_t',X_r)-\frac{1}{N}\sum_{X_r\in P_{t+1}^G} F_c(\pi,T_k,P_t',X_r)\leq 0, \tag{6}$$

$$0\leq F(\pi,T_k,P_{t+1})-\frac{1}{N}\sum_{X_r\in P_{t+1}} F_c(\pi,T_k,P_t',X_r)\leq T_k\log_K((N+M)/N), \tag{7}$$

$$0\leq F(\pi,T_k,P_{t+1}^G)-\frac{1}{N}\sum_{X_r\in P_{t+1}^G} F_c(\pi,T_k,P_t',X_r)\leq T_k\log_K((N+M)/N). \tag{8}$$

Here we omit the proof of Lemma 1 for paper length limitation.

Theorem 1. *Let $y=\frac{M}{N+M}$ and $D(\pi,T_k,P_t,O_t)=F(\pi,T_k,P_{t+1})-F(\pi,T_k,P_{t+1}^G)$. Then*

$$\lim_{y\to 0} D(\pi,T_k,P_t,O_t)=0. \tag{9}$$

Proof. Subtracting (8) from (7), we obtain an inequation:

$$D(\pi,T_k,P_t,O_t)\leq T_k\log_K((N+M)/N)+ \\ \left(\frac{1}{N}\sum_{X_r\in P_{t+1}} F_c(\pi,T_k,P_t',X_r)-\frac{1}{N}\sum_{X_r\in P_{t+1}^G} F_c(\pi,T_k,P_t',X_r)\right). \tag{10}$$

Then we substitute (6) into (10) to get:

$$D(\pi, T_k, P_t, O_t) \leq T_k \log_K ((N+M)/N). \tag{11}$$

Since P_{t+1}^G has the exact minimum free energy, we also get:

$$D(\pi, T_k, P_t, O_t) \geq 0. \tag{12}$$

Furthermore, there exists such a limit:

$$\lim_{y \to 0} (T_k \log_K ((N+M)/N)) = 0. \tag{13}$$

Applying the *Squeeze Law* of limits, we can obtain (9) from (11), (12) and (13).

Theorem 1 asserts that the selection rule STS has the approximate steepest descent ability of the free energy if $M \ll N$. According to Theorem 1, M and N have been set to satisfy $y < 0.1$ in all our experiments.

5 Experiments and Results

In this section, we present the experimental results on one instance of 0-1 knapsack problems, called KP2. KP2 is generated by the Pisinger's algorithm [8] of constructing test instances with these parameters: *data rang* R=1000, *instance size* n=100, *problem type* t=*weakly correlated*, *instances sum* S=1000 and *random seed* i=750. Note that here the profits p_j, the weights w_j and the capacity c are positive integers. We apply the simple genetic algorithm (SGA) with elitism, the steady state genetic algorithm (SSGA), TDGA and STEA to this instance. They all utilize the uniform crossover with its probability $P_c = 0.8$, the uniform mutation with its probability $P_m = 0.05$, and the same population size $N = 80$. The termination condition is satisfied when 3.2×10^6 individuals are searched. The offspring population for SSGA, TDGA and STEA has the size $M = 8$. All experiments are performed on a Pentium-4 3.0G computer.

Assume that the items are ordered according to their efficiency such that $p_i/w_i \geq p_j/w_j$ when $i < j$. Martello [9] has proved that the greedy value $f_u = \sum_{j=1}^{b-1} p_j + \lfloor x_b p_b \rfloor$ is an upper bound of the objective function f where

$$\sum_{j=1}^{b-1} p_j \leq c, \ \sum_{j=1}^{b} p_j > c, \text{and } x_b = (c - \sum_{j=1}^{b-1} w_j)/w_b. \tag{14}$$

It's also obvious that $f_l = 0$ is an infimum of f. Therefor we can get an upper bound $e_u = -f_l = 0$ and a lower bound $e_l = -f_u = -38249$ of energy for the level partition π about KP2. The other parameters of STEA have the following values: level number $K = \frac{1}{5}N = 16$ and chain length $L_k = 10n = 1000$.

There were respectively 40 trials on KP2 for each algorithm. Table 1 provides the statistic results of 40 trials for each algorithm, including the rate of hitting

Table 1. Statistic results for four algorithms

Algorithm	Hitting rate	Worst	Mean	Best	Time	First hitting time
SGA	1/40	38145	38215.700	38245†	152.2	>151.8
SSGA	8/40	38240	38241.000	38245†	303.2	>254.0
TDGA	23/40	38240	38242.875	38245†	56930.4	>26275.7
STEA	40/40	38245†	38245.000†	38245†	191.8	=30.8

† optimal solution

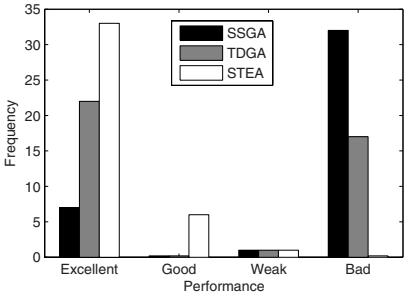


Fig. 2. Performance distributions

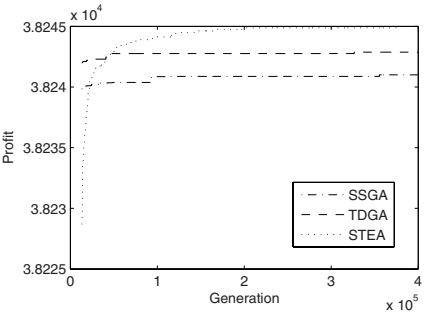


Fig. 3. Convergence curves

optimum successfully, the worst result, the average result, the best result, average running time (seconds), and average time (seconds) of first hitting optimum. We score the performance during each trial for each algorithm according to the number s of searched individuals when first hitting optimum as follows: (1) *excellent* if $s \leq 8\times10^5$; (2) *good* if $8\times10^5 < s \leq 1.6\times10^6$; (3) *weak* if $1.6\times10^6 < s \leq 3.2\times10^6$; (4) *bad* if $s > 3.2\times10^6$. Figure 2 shows the frequencies of four scores in 40 trials for SSGA, TDGA and STEA. Figure 3 shows their average convergence curves in 40 trials of the best individual of each generation. There are a few interesting observations which can be made on the basis of the experiments:

1. The results in Table 1 demonstrate clearly the stability of STEA. Its hitting rate is much higher than that of the other three algorithms. Note that for STEA the optimum was found in all 40 trials. Moreover, the quality of its solutions is averagely superior to that of the others due to its stability.
2. The above results also illustrate the high computational efficiency of STEA. STEA nearly spends the same running time as SGA. It contrasts clearly with the extremely high computational cost of TDGA, which is about 300 times as much as the others.
3. The first hitting time is a very significant performance goal for GA. It's very exciting that STEA has the far more rapid first hitting time than the others because of its high stability and low computational cost.
4. TDGA as well as SSGA presents the serious polarization phenomenon in Figure 2. It indicates that TDGA often get trapped in local optima.

However, STEA avoids this phenomenon due to keeping a systematical balance between selective pressure and population diversity successfully.

5. Figure 3 shows that the convergence speeds of SSGA and TDGA are faster than STEA at the early stage. However, they are in a nearly stagnant state after that stage and then the speed of STEA exceeds theirs. STEA gains the more rapid global convergence at the very little cost of the early stage.

6 Conclusions

This paper proposes a steep thermodynamical evolutionary algorithm (STEA), which utilizes a steep thermodynamical selection rule. STEA simulates the competitive mechanism between energy and entropy in annealing to systematically resolve the conflicts between selective pressure and population diversity in GA. The experimental results show that STEA not only speeds up the global convergence of TDGA remarkably at the very little cost of the early stage, but also improves the stability and the computational efficiency of TDGA greatly.

Further research will concentrate on the analysis of the convergence traits about STEA from the viewpoint of statistical mechanism.

Acknowledgement. This research is supported by the National Natural Science Foundation of China under Grant No.60473014.

References

1. Goldberg, D.E.: Genetic algorithms in search, optimization, and machine learning. Addison-Wesley. (1989)
2. Su, X.H., Yang, B., Wang, Y.D.: A genetic algorithm based on evolutionary stable strategy. Journal of Software. **14**(11) (2003) 1863-1868 (in Chinese)
3. Whitley, D.: The GENITOR algorithm and selection pressure: Why rank-based allocation of reproductive trials is best. Proc. of Int. Conf. on Genetic Algorithms. (1989) 116-123
4. Li, Y.X., Zou, X.F., Kang, L.S., Michalewicz, Z.: A new dynamical evolutionary algorithm based on statistical mechanics. J. Computer Science and Technology. **18** (2003) 361-368
5. Kirkpatrick, S., Gelatt, C.D., Vecchi, M.P.: Optimization by simulated annealing. Science. **220** (1983) 671-680
6. Mori, N., Yoshida, J., Kita, H., Nishikawa, Y.: A thermodynamical selection rule for the genetic algorithm. IEEE Conf. on Evolutionary Computation. (1995) 188-192
7. Kita, H., Mori, N., Nishikawa, Y.: Maintenance of Diversity by means of Thermodynamical Selection Rules for Genetic Problem Solving. Proc. of Int. Symposium on Artificial Life and Robotics. (1999) 330-333
8. Pisinger, D.: Core problems in knapsack algorithms. Operations Research. **47** (1999) 570-575
9. Martello, S., Toth, P.: Knapsack problems: algorithms and computer implementations. John Wiley & Sons. (1990)

A New Hybrid Optimization Algorithm Framework to Solve Constrained Optimization Problem

Huang Zhangcan[1] and Cheng Hao[2]

[1] School of Science, Wuhan University of Technology, Wuhan 430070, China
[2] School of Mechanical and Electrical Engineering, Wuhan University of Technology, Wuhan 430070, China
{huangzc,h.cheng}@whut.edu.cn

Abstract. Evolutionary Computation made great success from the theory of natural selection devised by Charles Darwin. It was a process of randomly searching but not emphasizing each individuals respective functions. This paper proposed a hybrid optimization algorithm framework trying to incorporate natural selection and survival of the fittest and birds of a feather flock together. Aiming at balancing search results and search speed, we adopted the search strategy to classify the individuals by their fitness. Individuals classification differentiated respective function in search process, thats the excellent individuals mine the local optimal solution and others explore the search domain to find new local optimal solution. Experimental findings support the theoretical basis of the proposed framework.

1 Introduction

Solving a constrained optimization problem with inequality, upper bound, and lower bound constraints usually includes mathematical model building and algorithm designing. Many methods have been posed from Newton Method, Linear Search to Evolutionary Computation [1], which has respective characteristics. Evolutionary computation from the theory of Natural Selection and Survival of the Fittest has long been accepted as a powerful search tool in both academia and industry, with numerous applications in varies science and engineering problem. The Darwinian-type evolutionary computation uses mutation (an asexual reproduction with variation), crossover (a recombination or sexual reproduction) and selection (survival of the fitness). These operators are simple to execute and domain-independent [2]. Michalski employs machine learning to generate new populations, called Learnable Evolutionary Model, briefly LEM. LEM integrated machine learning and evolutionary computation, which focused on why certain individuals are superior to others [3]. SUN employs similartaxis and dissimilation instead of crossover and mutation operators in Mind Evolutionary Computation, briefly MEC [4]. For low efficiency of conventional methods to dealing with highly complex problems, more methods like LEM, MEC trying to

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 1005–1012, 2007.
© Springer-Verlag Berlin Heidelberg 2007

emphasis individuals characteristics. The novel idea proposed in this research is to incorporate "birds of a feather flock together", a universal rule of human activity and "natural selection and survival of the fittest", Darwinian-type EC together to arrange each individual respective functions.

One of the most critical of these features is perhaps that how to build effective algorithm framework to solve more and more complex mathematical model. Jim Gray believes a good long-range goal should have five key properties: Understandable, Challenging, Useful, Testable, Incremental [5]. According to these rules, we try to build a hybrid optimization algorithm framework which focus on two goals. One is to unify the optimization problems into a framework and then use the framework to guide the mathematical models building, the other is to incorporate different algorithms search kernel.

Subsequent sections of this paper deal with this issue of usage of the framework to design algorithm to solve constrained optimization problem. The organization of the paper is as follows: In Section 2 defines the general constrained optimization problems. Section 3 presents the framework and its major functional aspects. Some experimental results and discussions are presented in Section 4, while conclusions are drawn in Section 5.

2 Statement of the Problem

The general constrained optimization problem can be defined as follows:

$$\begin{aligned} & & minimize \quad & f(\boldsymbol{x}) \\ subject \quad & to \quad & g_j(\boldsymbol{x}) \leq 0, j = 1, ..., J \\ & & h_m(\boldsymbol{x}) = 0, m = 1, ..., M \\ & & x_i^L \leq x_i \leq x_i^U, i = 1, ..., n \end{aligned} \tag{1}$$

where $f(\boldsymbol{x})$ is an objective function, $\boldsymbol{x}$ is the vector of decision variables $\boldsymbol{x} = [x_1, x_2, ..., x_n]$, J is the number of inequality constraints, and M is the number of equality constraints (objective function, constraints could be linear or nonlinear). x_i^L and x_i^U are the upper bound and the lower bound of x_i, respectively. Denote with F to the feasible region and with S to search space. The feasible solutions exist in $F \subseteq S$.

3 Description of Hybrid Optimization Algorithm Framework

In our previous work, multi-population adaptive-gathering evolutionary algorithm [6] using domains of attraction to separate and gather the initial population automatically to deal with function optimization, which divide the entire search space of objective function into several domains. In another paper, neighborhood exploring evolutionary strategy [7] consists of assigning neighborhood of different size to different solution to deal with multi-objective optimization problem. The population classification and neighborhood assignment behind these ideas are also the foundation of the Hybrid Optimization Algorithm Framework as our forward work.

3.1 Population Classification

The first critical step in the process of the framework is to calculate the individuals' fitness at every generation and to sort the population by their fitness. The individuals are selected randomly and they have their performance, better or worse. Most Evolutionary Computation algorithms work with population of fixed size keeping at least the best individual always in the current population [1]. No signs indicate part of the worse individuals remain in next population will hinder optimum seeking. If we ignore some worse individuals we lost the diversity of the population especially in multi-modal objective function space. Based on this idea, we intend to classify the population into several levels in the current population. We assign individuals in each level with different tasks to generate new offspring. The individuals in the next population will be selected from each level. Hence, we arrange each individual respective function and at the same time keep the diversity of the population.

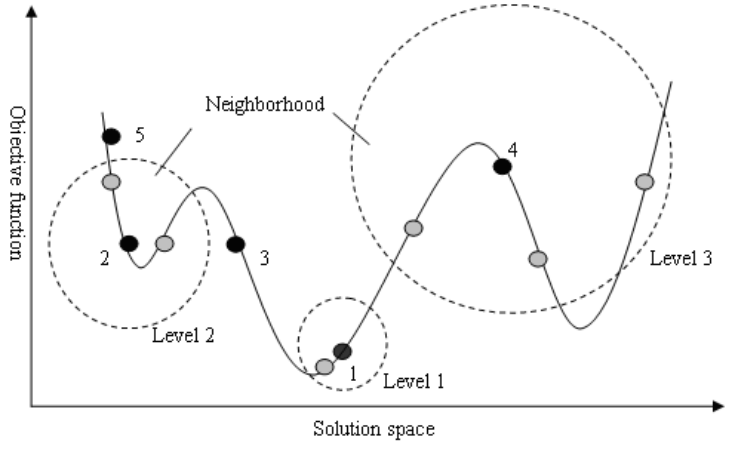


Fig. 1. Illustration of generating new offspring using (1+λ)-ES. The black dots denote father individuals and the grey dots denote offspring.

Figure 1 illustrates an example of such classification. There are five solutions in the solution space that are classified into three group, 1,2,3,4,5according to the individual's fitness. We assign each solution a level, for instance, in Figure 1, solution 1 in level 1, solution 2, 3 in level 2, solution 4, 5 in level 3.

3.2 Neighborhood Assignment

After population classification, the next step is to decide the neighborhood for every solution in the population. The neighborhood is designed in the decision variable space and the actual shape of the neighborhood can be various.

In this way, we can assign neighborhood to each level. The solutions with lower levels have small sized neighborhood and vice verse. The solutions in the same level have same sized neighborhood. For convenience, we use a circle denote a neighborhood and several different neighborhoods are showed in Figure 1.

3.3 (1+λ)-ES (Evolutionary Strategy) Selection

(1+ λ)-ES is an efficient selection method in evolutionary strategy, which generates λ offspring using the information 1 individual in the population [2]. The (1+λ)-ES can be used to generate new offspring. In this case, the individual in the level 1 uses (1+1)-ES, level 2 uses (1+2)-ES and level 3 uses (1+3)-ES.

3.4 Hybrid Optimization Algorithm Framework

Evolutionary algorithms deserve a special mention as powerful global optimizers, but which not emphasis on each individuals' respective function. It is a main difficult to balance search speed and result.

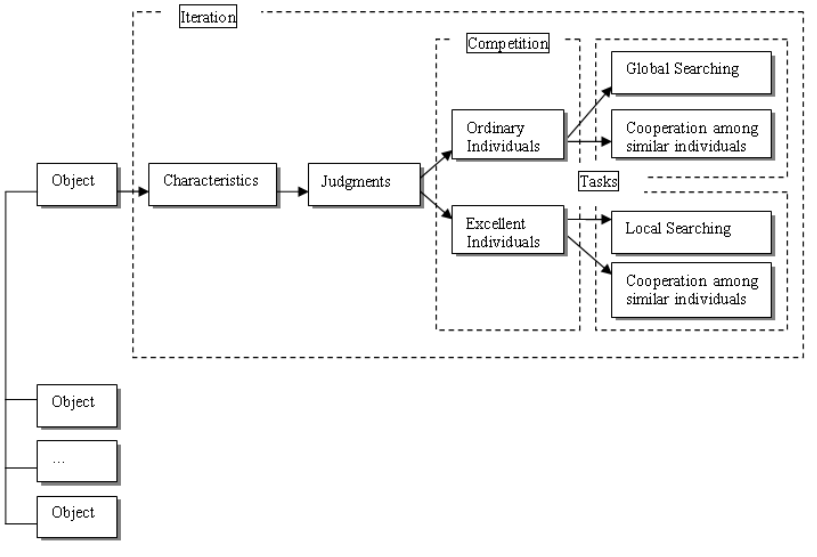


Fig. 2. The proposed optimization algorithm framework

We try to arrange the task to each individual from the theory of Birds of a Feather Flock Together. No matter how the optimization objects and objective changed, the characters of objectives on optimization objects have two respects. Firstly, capability similarity, that's ordinary individuals gather to excellent individuals. Secondly, structure similarity, that's those individuals with similar capability have similar structure in some kinds.

The Hybrid Optimization Algorithm Framework is based from randomly population search. To explain how the framework works, we consider an optimization problem as the object and the object functions as their characteristics. By judging each characteristic, we classify the population into different level. By competition among the called ordinary individuals and excellent individuals, we arrange their respective tasks. That is by cooperation among ordinary individuals to explore hopefully domain with global searching and excellent individuals to mine the local minima with local searching (see Figure 2).

3.5 Using the Framework to Design Algorithm

Under the guidance of the framework, we can design algorithms to solve constrained optimization problems. The process can be described as following:

```
Randomly initialize objects S(0)
Evaluate S(0)
t=0
while the termination condition is not satisfied
  Classify objects into k levels by the evaluated fitness
  Generate new individuals S' in each level using father
   individuals' level
  Add S' to S
  t=t+1
end while
```

How to classify objects into k levels according to the problem's complex and the human personal experience. We consider individuals' similarity in their characteristic.

$$f(x^*_{min}) + \frac{p}{k}L < f(x_i) \le f(x^*_{min}) + \frac{p+1}{k}L \quad x_i \in p \quad level, p = 1, 2, ..., k-1 \tag{2}$$

Where $x_i \in S$, $f(x_i)$ is the fitness of each individual (see Equation (3),(4)) and L is the distance between two individuals (see Equation (5)).

$$f(x^*_{max}) = maxf(x_i) \tag{3}$$

$$f(x^*_{min}) = minf(x_i) \tag{4}$$

$$L = maxf(x_i) - minf(x_i) \tag{5}$$

4 Statement of the Problem

Examples mentioned in this chapter are representatives of different kinds of functions. Example 1 is cited from Ref.[8], which has been known that when $n = 2$, the function has 720 local optimal solutions, 18 of which are global optimal solutions. Example 2 is cited from Ref.[9], which is nonlinear and its

global modal is unknown. We adopted real code and the experiment parameters are: population size $N = 20$, initialized radius $r_0 = c2^{p-1}$ and level $k = 4$. Here, p is the individual's level, c is a constant, which is set by domain radius.

Example 1: n-demension Shubert function

$$\begin{aligned} min \quad & f(x_1, x_2, ..., x_n) = \prod_{i=1}^{n} \sum_{j=1}^{5} j \cos((j+1)x_i + j) \\ & x_1 \in [-10, 10], i = 1, 2, ..., n \end{aligned} \tag{6}$$

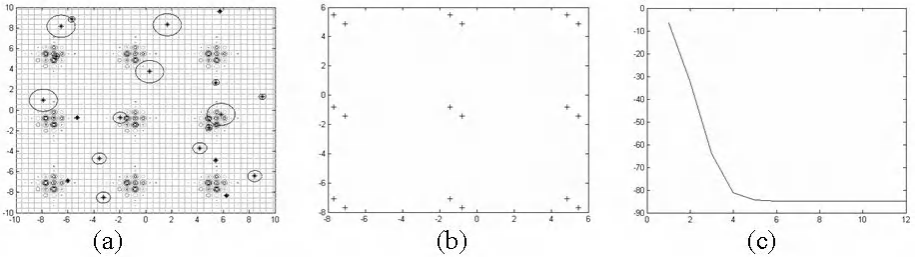


Fig. 3. (a)Distribution of Initial population and its levels. (b)distribution of the 12th generation. (c)Curve of convergence.

Table 1. Results observed under different contractive ratio c

c	Average modals found	Running times	Times approaching modals	Average generations running
1	3	10	5	730
0.9	6	10	9	88
0.8	5	10	10	46
0.7	5	10	10	30
0.6	4	10	10	23
0.5	3	10	10	17
0.4	2	10	10	14
0.3	1	10	10	11

Example 1 is a high-dimension function, which have 18 global optimal solutions (see Figure 3). At 12th generation the algorithm obtained global solutions.

To test the effects of the parameters on the algorithm, we varied the ratio c on Shubert function. The result can be viewed in Table 1. When $c = 1$, the searching process is global searching, not decreasing the convergence speed but found few modals. When $c = 0.9, 0.8, 0.7$, more modals were found and speed up the convergence speed. When $c = 0.6, 0.5, 0.4, 0.3$, modals can be found in each run quickly, but more modals were failed to find. From the above result, we can see for ordinary optimization problems, the c should be in 0.6 to 1. From the point of searching, the smaller searching domain is the easier to converge to

local minima. When holding a larger c and in a larger searching domain, it is more easy to find global minima. When decrease c, the algorithm has a quick convergence speed but probably the result is local minima.

Example 2: Bump Function

$$f(X) = -|\frac{\sum_{i=1}^{n} \cos^4(x_i) - 2\prod_{i=1}^{n} cos^2(x_i)}{\sqrt{\sum_{i=1}^{n} ix_i^2}}| \tag{7}$$
$$\prod_{i=1}^{n} x_i \geq 0.75, \sum_{i=1}^{n} x_i \leq 0.75n, 0 \leq x_i \leq 10, i = 1, 2, ...n$$

Table 2. Results of bump function in 20, 50, 100 dimension

Dimensions	20	50	100
Framework	-0.803615	- 0.835196	- 0.842786
Ref.[8],[9]	-0.8035	- 0. 8352620	-0.8410

Example 2 is a high-dimension bump function, which is very difficult to solve. Generally, we get satisfied results in 20, 50, 100 dimensions (see Table 2).

5 Conclusion and Future Works

In this study, natural selection and survival of the fittest, birds of a feather flock together, are incorporated into a hybrid optimization algorithm framework by arranging each individual's task.

With regards to algorithm from the framework, further work is necessary to improve the method of individuals' classification and local speed search. Also, the way of deciding the proper contractive ratio c should be devised. The used method$c$ is set to 0.7, but there may be a more appropriate value to enhance the quality of convergence. It may be desirable to devise a guide to choose the proper value of c. To test the high dimensional bump function is a remarkable performance, we just adopted simply acceleration and many excellent methods, like greedy algorithm, are value to test in the framework, how to make them cooperate perfectly is our future work.

References

1. Christian Blum, Andrea Roli: Metaheuristics in Combinatorial Optimization: Overview and Conceptual Comparison. ACM Computing Surveys, **35** (3). ACM, Inc., New York (2003) 268-308
2. Zbigniew Michalewicz, David B.Fogel: How to Solve it: Modern Heuristics. Springer-Verlag Berlin Heidelberg.(2000)

3. Ryszard S.Michalski: Learnable Evolutionary Model: Evolutionary Processes Guided by Machine Learning. Machine Learning. **38** (1-2). Springer, Netherlands (2000) 9-40
4. Sun Chengyi, Zhou Xiuling, Wang Wanzhen: Mind Evolutionary Computation and Applications. Journal of Communication and Computer. **1** (1). USA-China Business Review (Journal), Inc., USA (2004) 13-21
5. Jim Gray: What Next? A Dozen Information-Technology Research Goals. Journal of the ACM, **50** (1). ACM, Inc., New York (2003) 41-57
6. Chen Siduo, Huang Zhangcan: Multi-population adaptive-gathering evolutionary algorithm in function optimization. Proceedings of the 2000 Evolutionary Computation Congress, **1**. IEEE (2000) 817-821
7. Hu Xiaolin, Carlos A. Coello Coello, Huang Zhangcan: A New Multi-objective Evolutionary Algorithm: Neighborhood Exploring Evolution Strategy. Engineering Optimization, **37**. Taylor and Francis Ltd, UK(2005) 351-379
8. Huang Yuzhen, Kang Lishan, Zhou Aimin: Two-Phase Genetic Algorithm Applied in the Optimization of Multi-Modal Function. Wuhan University Journal of Natural Sciences, **8**. Wuhan University Journals Press, Wuhan(2003) 259-264
9. Kang Zhuo, Li Yan, Liu Pu, et al: Two Asynchronous Parallel Algorithms for Function Optimization. Wuhan University Journal of Natural Sciences, **48** (1). Wuhan University Journals Press, Wuhan (2002)33-36

In Search of Proper Pareto-optimal Solutions Using Multi-objective Evolutionary Algorithms

Pradyumn Kumar Shukla

Institute of Numerical Mathematics
Department of Mathematics
Technische Universität Dresden
Dresden PIN 01069, Germany
pradyumn.shukla@mailbox.tu-dresden.de

Abstract. There are multiple solution concepts in multi-objective optimization among which a decision maker has to select some good solutions usually which satisfy some trade-off criteria's. The need for potentially good solutions has always been one of the primary aims in multi-objective optimization. A complete representation of all these solutions is only possible with population based approaches like multi-objective evolutionary algorithms since then trade-off's can be calculated at each generation from the population members. Thus this paper proposes the use of multi-objective evolutionary algorithms for obtaining a complete representation of these good solutions. Theoretical results show how one can integrate search procedure for obtaining these solutions in population based evolutionary algorithms and some convergence results. Finally simulation results are presented on a number of test problems.

Keywords: Multi-objective optimization, Trade-off, Evolutionary algorithms.

1 Introduction

Multi-objective optimization is one of the most rapidly growing areas of modern optimization theory, see for example Deb [2], Miettinen [7] and the references therein. Since there are multiple solution concepts in multi-objective optimization it often becomes a challenging issue both in theory and practice. The set of all efficient points as is well known lies in the boundary of the objective space and is thus referred to as the efficient frontier. However all points on the frontier need not have equally nice properties which a decision maker may desire and thus one needs to filter out the bad Pareto points and keep the good ones. Such nice Pareto points are referred to in the literature proper Pareto solutions. Thus the need for potentially good solutions has always been one of the primary aims in multi-objective optimization. Good solutions can be thought of as "knee-points" on the efficient frontier or that are good in trade-off with respect to other solutions. However in most of practical and large scale problems, the user may usually not get the exact efficient front and thus he has to

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 1013–1020, 2007.
© Springer-Verlag Berlin Heidelberg 2007

be content with approximate solutions. This usually happens if one uses population based approach like *Multi-objective Evolutionary Algorithm (MOEA)* or any other algorithm. In these algorithm the obtained solutions may be thought of as approximate representation of the efficient front. We need to filter out the bad ones and keep the so called ε-proper Pareto solutions. A complete representation of ε-proper Pareto solutions is only possible using population based algorithms since trade-off's need to be calculated for them. This paper presents a way to obtain all these ε-proper Pareto solutions using MOEA's. Theoretical results are also presented so show convergence of such an algorithm.

The paper has been organized in three sections of which this is the first. Definitions of ε-proper efficiency theoretical results are presented in Section 2 while simulation results and conclusions are presented in Section 3.

2 Theoretical Results

Consider the following general multi-objective optimization problem (MP):

$$\min_{x \in X} f(x) = (f_1(x), f_2(x), \ldots, f_m(x))$$

where each $f_i : \mathbb{R}^n \to \mathbb{R}$ and $X \subseteq \mathbb{R}^n$. In what follows we will consider $\varepsilon \in \mathbb{R}^m_+$, i.e. $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_m)$, $\varepsilon_i \geq 0$ for all i. In some cases we will set $\varepsilon_i = \varepsilon'$, for all i and then $\varepsilon = e\varepsilon'$ where $e = (1, \ldots, 1) \in \mathbb{R}^m_+$.

Definition 1 ε-Pareto optimal. *Let $\varepsilon \in \mathbb{R}^m_+$ be given then a point $x^* \in X$ is said to be an ε-Pareto optimal of (MP) if there exists no $x \in X$ such that,*

$$f_i(x) \leq f_i(x^*) - \varepsilon_i, \quad \forall i \in \{1, 2, \ldots, m\}. \tag{1}$$

and with strict inequality holding for at least one index.

Observe that if $\varepsilon = 0$, the above definition reduces to that of a Pareto optimal. Let us denote the set of ε-Pareto points as $X_{\varepsilon - par}$ and the set of Pareto points as X_{par}. The notion of trade-off is inherent in multi-objective optimization which led to more robust solution concept defined next.

Definition 2 Proper Pareto optimal. *$x_0 \in X$ is called proper Pareto optimal if x_0 is Pareto optimal and if there exists a number $M > 0$ such that for all i and $x \in X$ satisfying $f_i(x) < f_i(x_0)$, there exists an index j such that $f_j(x_0) < f_j(x)$ and moreover $(f_i(x_0) - f_i(x))/(f_j(x) - f_j(x_0)) \leq M$.*

There are many other notions of proper Pareto optimal solutions and the above one based on trade-off's was introduced by Geoffrion [5]. Let us denote the set of all Geoffrion properly Pareto optimal solution as X_G.

Lemma 1. *A point $x_0 \in X_G$ if and only if there exists $M > 0$ such that the following system is inconsistent (for all $i = 1, 2, \ldots, m$ and for all $x \in X$).*

$$\begin{aligned} &-f_i(x_0) + f_i(x) < 0 \\ &-f_i(x_0) + f_i(x) < M(f_j(x_0) - f_j(x)) \quad \forall j \neq i. \end{aligned}$$

Proof: If $x_0 \in X_G$ then it is clear from the definition that above system is inconsistent.

Suppose the system is inconsistent for some $M > 0$. We claim that $x_0 \in X_{par}$. If not then on the contrary then there exists $x \in X$ such that $f_l(x) < f_l(x_0)$ for some index l, and $f_k(x) \leq f_k(x_0)$, for all $k \neq l$. Thus one easily sees that system 1 has a solution for index $i = l$. Hence $x_0 \in X_{par}$. If $x_0 \notin X_G$ then for all $\overline{M} > 0$ there is an index i, and some $x \in X$ satisfying $-f_i(x_0) + f_i(x) < 0$, $-f_i(x_0)+f_i(x) < \overline{M}(f_j(x_0)-f_j(x))$ for all j such that $-f_j(x_0)+f_j(x) > 0$ (such a j exists since $x_0 \in X_{par}$). For j such that $-f_j(x_0)+f_j(x) \leq 0$, $-f_i(x_0)+f_i(x) < \overline{M}(f_j(x_0) - f_j(x))$ is trivially true. Thus the system is consistent for all $\overline{M} > 0$, and hence a contradiction. □

Note that in Geoffrion's definition $x \in X$. However as shown in next lemma, when $Y = f(X)$ is $\mathbb{R}^m_+$ compact (i.e. the sections $(y - \mathbb{R}^m_+) \cap Y$ are compact for all $y \in Y$) then this can be replaced by $x \in X_{par}$.

Lemma 2. *Suppose that $Y = f(X)$ is $\mathbb{R}^m_+$ compact, then $x_0 \in X_G$ if x_0 is Pareto optimal and if there exists a number $M > 0$ such that for all i and $x \in X_{par}$ satisfying $f_i(x) < f_i(x_0)$, there exists an index j such that $f_j(x_0) < f_j(x)$ and moreover $(f_i(x_0) - f_i(x))/(f_j(x) - f_j(x_0)) \leq M$.*

Suppose that x_0 satisfies the conditions of the lemma. Then using Lemma 1, we obtain that for all $\hat{x} \in X_{par}$ the following system which we mark as (System 1), has no solutions,

$$\begin{aligned} &-f_i(x_0) + f_i(\hat{x}) < 0 \\ &-f_i(x_0) + f_i(\hat{x}) < M(f_j(x_0) - f_j(\hat{x})) \qquad \forall j \neq i. \end{aligned}$$

Take any $x \in X, x \notin X_{par}$. Now since $Y = f(X)$ is $\mathbb{R}^m_+$ compact so there exists $\hat{x} \in X_{par}$ such that

$$\begin{aligned} f_i(\hat{x}) - f_i(x) \leq 0 \qquad &\forall i = 1, 2, \ldots, m \\ f_k(\hat{x}) - f_k(x) < 0 \qquad &\text{for some} \quad k. \end{aligned}$$

Since the System 1 has no solutions, thus we obtain that the following system also has no solutions

$$\begin{aligned} &-f_i(x_0) + f_i(\hat{x}) < f_i(\hat{x}) - f_i(x) \\ &-f_i(x_0) + f_i(\hat{x}) < M(f_j(x_0) - f_j(\hat{x})) + M(f_j(\hat{x}) - f_j(x)) + f_i(\hat{x}) - f_i(x) \;\; \forall j \neq i. \end{aligned}$$

which is equivalent to saying that the following system is inconsistent,

$$\begin{aligned} &-f_i(x_0) + f_i(x) < 0 \\ &-f_i(x_0) + f_i(x) < M(f_j(x_0) - f_j(x)) \qquad \forall j \neq i. \end{aligned}$$

Thus System 1 has no solutions for *any* $x \in X$. Thus $x \in X_G$ □

Definition 3 ε-properly Pareto optimal (Liu [6]). *A point, $x^* \in X$ is called ε-proper Pareto optimal, if x^* is ε-Pareto optimal and there exists a number $M > 0$ such that for all i and $x \in X$ satisfying $f_i(x) < f_i(x^*) - \varepsilon_i$, there exists an index j such that $f_j(x^*) - \varepsilon_j < f_j(x)$ and moreover $(f_i(x^*) - f_i(x) - \varepsilon_i)/(f_j(x) - f_j(x^*) + \varepsilon_j) \leq M$.*

Observe that if $\varepsilon = 0$, the above definition reduces to that of a Geoffrion proper Pareto optimal. Let us denote the set of all Liu properly Pareto optimal solution as $X_L(\varepsilon)$. Let us however observe in the above definition and definition 2.2, M is arbitrary. On the other side M provides a bound on the trade off between the components of the objective vector. It is more natural to expect in practice the decision maker will provide a bound on such trade offs. Thus we are motivated to define the following.

Definition 4 Geoffrion M properly Pareto optimal. *Given a positive number $M > 0$, $x_0 \in X$ is called Geoffrion M proper Pareto optimal if x_0 is Pareto optimal and if for all i and $x \in X$ satisfying $f_i(x) < f_i(x_0)$, there exists an index j such that $f_j(x_0) < f_j(x)$ and moreover $(f_i(x_0) - f_i(x))/(f_j(x) - f_j(x_0)) \leq M$.*

Let us denote the set of all Geoffrion M properly Pareto optimal solution as X_M. It is to be noted that a similar modified definition is also possible for Liu ε-proper Pareto optimal solutions. Let us denote the set of all M ε-proper Pareto optimal as $X_M(\varepsilon)$.

Theorem 1. *Let $\varepsilon = \varepsilon' e$ where $\varepsilon' \in \mathbb{R}$, $\varepsilon' > 0$ and $e = (1, 1, \ldots, 1)$, then for any fixed M,*

$$X_M = \cap_{\varepsilon' > 0} X_M(\varepsilon) \tag{2}$$

Proof: Let $x_0 \in \cap_{\varepsilon' > 0} X_M(\varepsilon)$. Hence for any $\varepsilon' > 0$, and for all i, the following system

$$\begin{aligned} -f_i(x_0) + f_i(x) + \varepsilon' &< 0 \\ Mf_j(x) + f_i(x) - Mf_j(x_0) - f_j(x_0) + M\varepsilon' + \varepsilon' &< 0 \end{aligned}$$

has no solutions in $x \in X$ Let $W = \mathbb{R}^m \backslash (-int\mathbb{R}^m_+)$ and consider the vectors $F^i(\varepsilon)$ (for all $i = 1, \ldots, m$) whose first component is given by $-f_i(x_0) + f_i(x) + \varepsilon'$ and whose j^{th} component is equal to $Mf_j(x) + f_i(x) - Mf_j(x_0) - f_j(x_0) + M\varepsilon' + \varepsilon'$, for all $j = 2, \ldots, m$ then $F^i(\varepsilon) \in W$ for all $x \in X$. Now since W is a closed cone for each i

$$\lim_{\varepsilon \to 0} F^i(\varepsilon) \in W$$

This shows that the following system

$$\begin{aligned} -f_i(x_0) + f_i(x) &< 0 \\ Mf_j(x) + f_i(x) - Mf_j(x_0) - f_j(x_0) &< 0 \end{aligned}$$

is inconsistent for all $x \in X$. Thus by Lemma 1 x_0 is M-properly Pareto optimal, or $x_0 \in X_M$. This shows that $\cap_{\varepsilon'>0} X_M(\varepsilon) \subset X_M$ Conversely, let $x_0 \in X_M$, thus for all $i = 1, \ldots, m$ following system

$$\begin{aligned} -f_i(x_0) + f_i(x) &< 0 \\ Mf_j(x) + f_i(x) - Mf_j(x_0) - f_j(x_0) &< 0 \end{aligned}$$

is inconsistent for all $x \in X$, thus the following system

$$\begin{aligned} -f_i(x_0) + f_i(x) &< -\varepsilon' \\ Mf_j(x) + f_i(x) - Mf_j(x_0) - f_j(x_0) &< -M\varepsilon' - \varepsilon' \end{aligned}$$

is also inconsistent for all $x \in X$.Thus x_0 is M ε-properly Pareto for all $\varepsilon' > 0$ Hence $x_0 \in \cap_{\varepsilon'>0} X_M(\varepsilon)$. □

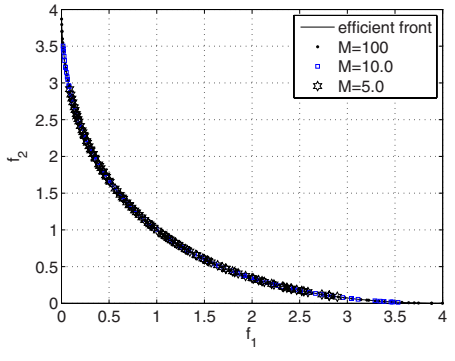


Fig. 1. Proper Pareto optimal solutions obtained using M=100, 10 and 5 on SCH

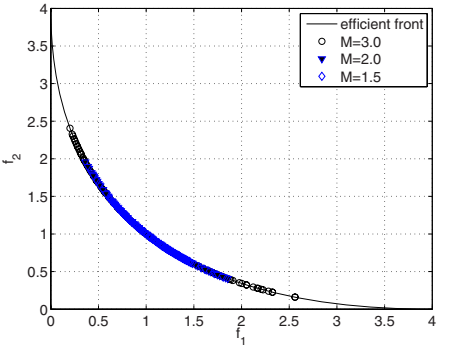


Fig. 2. Proper Pareto optimal solutions obtained using M=3, 2 and 1.5 on SCH

Note that for checking an obtained solution for proper Pareto optimality one needs to checks boundedness of trade-off's with *all* feasible points. Using Lemma 2 however we need to check this for only solutions that belong to the non-dominated front. Thus the above theoretical results can be effectively applied to population based multi-objective evolutionary algorithms since in these algorithms one gets a approximate non-dominated front at each iteration. Theorem 1 says that if an algorithm is computing approximate proper Pareto optimal solutions then in the end we obtain exact proper Pareto optimal solutions. In order to compute approximate proper Pareto optimal solutions we take the elitist non-dominated sorting GA or NSGA-II [3] and introduce constraints as follows. Given any solution $f(x_0)$ at any generation, we calculate the following constraint violation

$$c(x_0) = \sum_x \min\{0, \{M - (f_i(x_0) - f_i(x))/(f_j(x) - f_j(x_0))\}\}$$

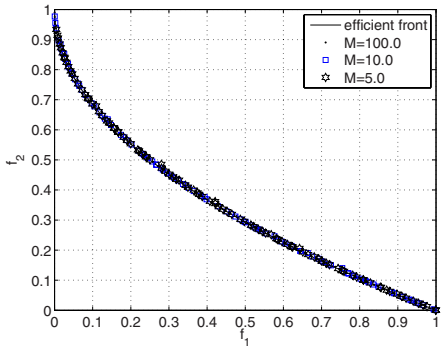


Fig. 3. Proper Pareto optimal solutions obtained using M=100, 10 and 5 on ZDT1

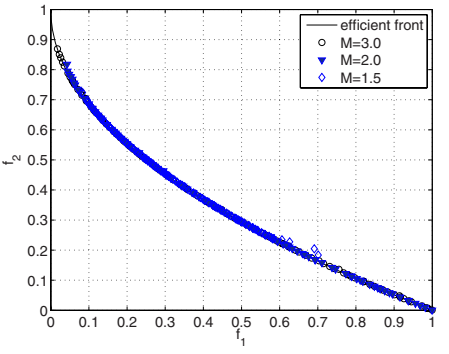


Fig. 4. Proper Pareto optimal solutions obtained using M=3, 3 and 1.5 on ZDT1

for all solutions x and all indices (i, j) that satisfy $f_i(x) < f_i(x_0), f_j(x) > f_j(x_0)$. Thus for any solution that is non Pareto optimal $c(x_0)$ provides a measure of how far is the solution from proper Pareto optimal. For proper Pareto optimal solutions the constraint violation is zero. Thus this provides a measure to penalty solutions that are non Pareto optimal.

3 Simulation Results and Conclusions

In this section, we apply the above constraint approach to obtain proper Pareto optimal solutions for different values of M. For the NSGA-II, we use a standard real-parameter SBX and polynomial mutation operator with $\eta_c = 10$ and $\eta_m = 10$, respectively [2] (unless otherwise stated). For all problems solved, we use a population of size 100. We set the number of function evaluations as 20000 for each problems.

First we consider a one variable bi-objective Schaffler's test problem (SCH). This problem is unconstrained and has a convex efficient frontier. The Pareto optimal solutions correspond to $x = [0, 2]$. Figure 1 shows the obtained efficient front corresponding to M=100, 10 and 5. The complete efficient front is also shown. The unrestricted efficient front can be seen as solutions corresponding to proper Proper optimal solutions with $M = \infty$. It can be seen that all the efficient front is proper Pareto optimal with respect to M=100. The efficient front reduces when smaller values of M are used with f_1 =0.36 to 1.89 for M=2 (Figure 2).

The box-constrained ZDT1 problem has a convex Pareto optimal front for which solutions correspond to $0 \leq x_1^* \leq 1$ and $x_i^* = 0$ for $i = 2, 3, \ldots, 30$. Here Figure 3 and 4 show the obtained efficient front corresponding to different values of M. It can be seen that (as opposed to SCH) here only the part corresponding to minimization of f_1 is chopped out as M value is reduced. This is because M proper Pareto optimal solutions are based on trade-off's and are thus

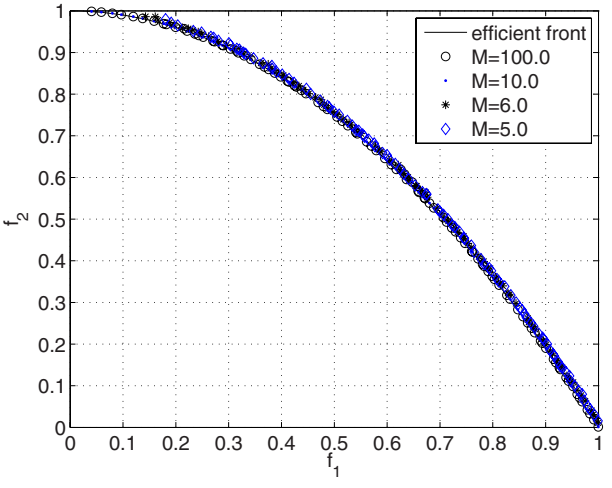


Fig. 5. Proper Pareto optimal solutions obtained using M=100, 10, 6 and 5 on ZDT2

related to slope of efficient front. In the ZDT1 problem slope is more near the f_1 minimum.

The box-constrained ZDT2 problem has a non-convex Pareto optimal front for which solutions correspond to $0 \leq x_1^* \leq 1$ and $x_i^* = 0$ for $i = 2, 3, \ldots, 30$. This problem is non-convex so in this case guided domination approach [1] cannot be used for finding intermediate regions in this problem. However even in this case concept of trade-off can be applied to obtain M proper Pareto optimal solutions. Figure 5 shows the obtained efficient front corresponding to different values of M. As opposed to ZDT1 it is observed that for M values smaller than 5.0 only a very small part of efficient front is obtained while no feasible solution is M proper Pareto optimal values for M=2. Thus in this case these values give range of realistic trade-off values to a decision maker.

Finally we consider a constrained test problem (CTP7). This problem has disconnected set of continuous regions. As suggested in [4] we use five decision variables. Here the efficient front consists of six disconnected convex regions. Figure 6 and 7 show the obtained efficient front corresponding to different values of M. It can be seen that all six parts of efficient front are obtained with M proper Pareto optimal solutions corresponding to M=100, 5 and 2. With M=1.5 however only four parts remain and for M=1.2 only one part of efficient front remains.

The decision maker could also use different M values for trade-off's in different objectives. It is a simple exercise to show the theoretical results presented in Section 2 remain valid in such a case. Finally it is to be noted that the convergence result of Theorem 1 could be applied to any other population based multi-objective evolutionary algorithm other than NSGA-II. In such cases non M proper Pareto optimal solutions could be penalized and then some constraint handling approach can also be used.

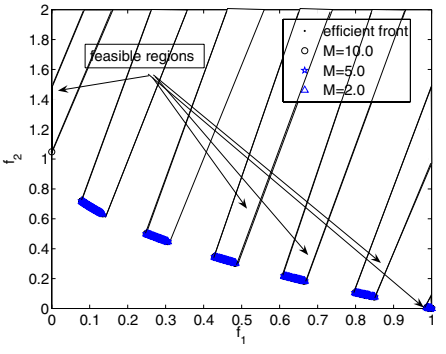


Fig. 6. Proper Pareto optimal solutions obtained using M=10, 5 and 2 on CTP7

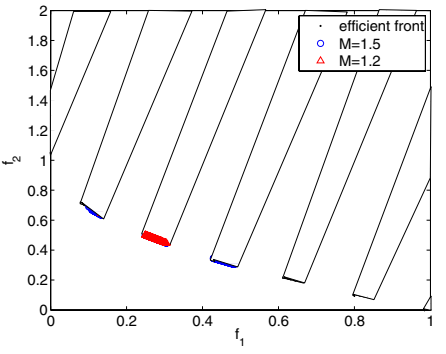


Fig. 7. Proper Pareto optimal solutions obtained using M=1.5, and 1.2 on CTP7

Acknowledgements

The author acknowledges the partial financial support by the Gottlieb-Daimler- and Karl Benz-Foundation under Project No. 02-13/05.

References

1. J. Branke, T. Kauβler, and H. Schmeck. Guidance in evolutionary multi-objective optimization. *Advances in Engineering Software*, 32:499–507, 2001.
2. K. Deb. *Multi-objective optimization using evolutionary algorithms*. Chichester, UK: Wiley, 2001.
3. K. Deb, S. Agrawal, A. Pratap, and T. Meyarivan. A fast and elitist multi-objective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, 6(2):182–197, 2002.
4. K. Deb, A. Pratap, and T. Meyarivan. Constrained test problems for multi-objective evolutionary optimization. In *Proceedings of the First International Conference on Evolutionary Multi-Criterion Optimization (EMO-01)*, pages 284–298, 2001.
5. Arthur M. Geoffrion. Proper efficiency and the theory of vector maximization. *J. Math. Anal. Appl.*, 22:618–630, 1968.
6. Jen-Chwan Liu. ϵ-properly efficient solutions to nondifferentiable multiobjective programming problems. *Appl. Math. Lett.*, 12(6):109–113, 1999.
7. K. Miettinen. *Nonlinear Multiobjective Optimization*. Kluwer, Boston, 1999.

Cultural Particle Swarm Algorithms for Constrained Multi-objective Optimization

Fang Gao[1], Qiang Zhao[2], Hongwei Liu[1], and Gang Cui[1]

[1] School of Computer Science and Technology, Harbin Institute of Technology, 150001 Harbin, China
[2] School of traffic, Northeast forestry university, 150040 Harbin, China
gaofang@hit.edu.cn, qyangzhao@163.com,
{lhw,cg}@ftcl.hit.edu.cn

Abstract. In this paper, we propose to integrate particle swarm optimization algorithm into cultural algorithms frame to develop a more efficient cultural particle swarm algorithms (CPSA) for constrained multi-objective optimization problem. In our CPSA, the population space of cultural algorithms consists of n+1 subswarms which are used to search for the n single-objective optimums and an additional multiobjective optimum. The belief space accepts 20% elite particles form each subswarm and further takes crossover to create Pareto optimums. Niche Pareto tournament selection is further executed to ensure Pareto set to distribute uniformly along Pareto frontier. Additional memory of Pareto optimums spool is allocated and updated in each iteration to keep resultant Pareto solutions. Besides, a direct comparison method is employed to handle constraints without needing penalty functions. Two examples are presented to demonstrate the effectiveness of the proposed algorithm.

Keywords: Multi-objective optimization, Cultural algorithms, Particle swarm optimization, Pareto optimums, Crossover.

1 Introduction

Multi-objective optimization (MOP) has assumed greater importance in many real engineering applications. A large number of algorithms based on evolutionary and swarm intelligence have been proposed for the solution of MOP during the last decades such as NSGA, VEGA and other GA-based algorithms [1,2,3]. During the past decade, the particle swarm optimization (PSO) algorithm, proposed by James Kennedy and Russell Eberhart in 1995 [4], has gained more attraction for its simplicity and effectiveness. Recently, many researchers began to use PSO to solve multi-objective optimization problems. However, the performance of simple PSO greatly depends on its parameters, and it often suffers from the problem of being trapped in local optima that is to cause premature convergence. For example, the original PSO had difficulties in controlling the balance between exploration and exploitation because it tends to favor the intensification search around the 'better' solutions previously found [5,6,7]. Therefore, many approaches have proposed various kinds of approaches to improve the standard PSO algorithm. Meantime Cultural Algorithms

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 1021 – 1028, 2007.
© Springer-Verlag Berlin Heidelberg 2007

(CA) proposed by Reynolds [8], is a novel evolutionary computational frame based on concept of culture of human society, which shows higher intelligence in treating all kinds of complicated problems [9,10]. In this paper we propose to integrate PSO into the frame of CA and put forward a cultural particle swarm algorithm (CPSA) for constrained multi-objective optimization problems. It synthesizes both the advantages of PSO algorithm and CA and overcomes their drawbacks

2 Introduction to PSO Algorithm

The particle swarm optimization algorithm first randomly initializes a swarm of particles. Each particle is represented as $\boldsymbol{X}_i = (x_{i,1}, x_{i,2}, \cdots, x_{i,n}) \quad i = 1,2,\cdots,N,$ where N is the swarm size, and n is the total dimension number of each particle. Each particle adjusts its trajectory towards its own previous best position ***pbest*** and the previous global best position ***gbest*** attained by the whole swarm. In the k th iteration, the i th particle with respect to the j th dimension is updated by

$$v_{i,j}^{(k+1)} = v_{i,j}^{(k)} + c_1 r_1 (pbest_{i,j}^{(k)} - x_{i,j}^{(k)}) + c_2 r_2 (gbest_j^{(k)} - x_{i,j}^{(k)}), \tag{1}$$

$$x_{i,j}^{(k+1)} = x_{i,j}^{(k)} + v_{i,j}^{(k+1)}, \tag{2}$$

where $x_{i,j}^{(k)}$ and $v_{i,j}^{(k)}$ are the current position and velocity, respectively. c_1 and c_2 are acceleration constants, r_1 and r_2 are random numbers within the interval of [0,1].

The procedure of the PSO algorithm for optimization can be described as follows:

(1) Initializing: Generate N particles with random positions $X_1, X_2, \cdots, X_N$, and velocities $V_1, V_2, \cdots, V_N$, $i = 1,2,\cdots,N$.
(2) Evaluate the fitness of each particle based on the problem's objective function.
(3) Individual and global best positions updating: If $f(\boldsymbol{pbest}_i) > f(\boldsymbol{X}_i)$, then let $\boldsymbol{pbest}_i = \boldsymbol{X}_i$, and search for the minimum value $f_{\min}$ among $f(\boldsymbol{pbest}_i)$, If $f(\boldsymbol{gbest}) > f_{\min}$, let $\boldsymbol{gbest} = \boldsymbol{X}_{\min}$, $\boldsymbol{X}_{\min}$ is the particle associated with $f_{\min}$.
(4) Velocity and position updating with the Eq. (1) and (2).
(5) Repeat Steps of (2) to (5) until a given maximum iterations number are achieved.

3 Frame of Cultural Algorithms

Cultural algorithms provide for interaction and mutual cooperation of two distinct levels of evolution: a population space and a belief space. It models cultural evolution by allowing a collection of information at a macro evolutionary level. Such information can be shared among single individuals whose evolutionary dynamics constitute the micro level. The two levels influence each other through the communication protocol. The presence of the belief space provides for a global knowledge repository and can guide the search towards better solutions, by using the knowledge to prune large portions of the state space. A framework for a CA is shown in Fig. 1.

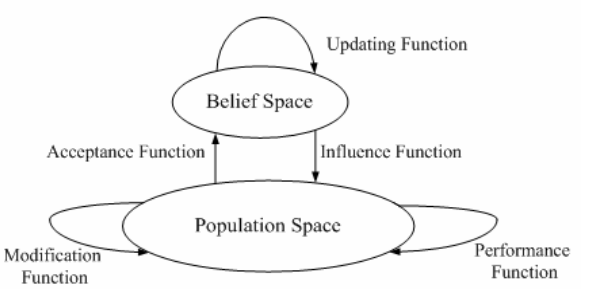


Fig. 1. Framework of cultural algorithms

As shown in Fig.1, the two spaces are connected together by an explicit communication protocol composed of two functions: an acceptance function and an influence function. The contents of the belief space can be altered via an updating function. In the population space, individuals are first evaluated by a performance function, and then new individuals are created by a modification function. In Reynolds's frame of cultural algorithm [8], an evolutionary planning (EP) model is employed in the population space. In this paper, multiple subswarms PSO algorithm will replace EP model for the constrained MOP problems.

4 Cultural PSO Algorithm (CPSA)

4.1 Frame of CPSA

In our dual evolution approach of CPSA, PSO algorithm is integrated into the frame of cultural algorithms for the solution of MOP as shown in Fig. 2.

The swarm intelligence of PSO is used for the evolution of population space. Since there are totally q objective functions, the population space include q subswarms, with the ith swarm evolves with $f_i(x)$ ($i=1,2,\cdots,q$) as its single optimization objective respectively. Besides, another additional subswarm is also added into the population space, and this subswarm randomly selects one function from $f_1(x), f_2(x),\cdots, f_q(x)$ as its single optimization objective in every cycle of k_0 iterations. After every cycle of k_0 iterations, each subswarm outputs 20% elite particles within its whole subswarm to the belief space. All the elite particles coming from different subswarms constitute the belief space and it further takes crossover operations to create Pareto optimums.

In this crossover, two particles are randomly selected as parents and performed as

$$X_i' = \alpha \cdot X_i + (1-\alpha) \cdot X_j, \tag{3}$$

$$X_j' = (1-\alpha) \cdot X_i + \alpha \cdot X_j, \tag{4}$$

where α is a random number extracted from region (0,1).

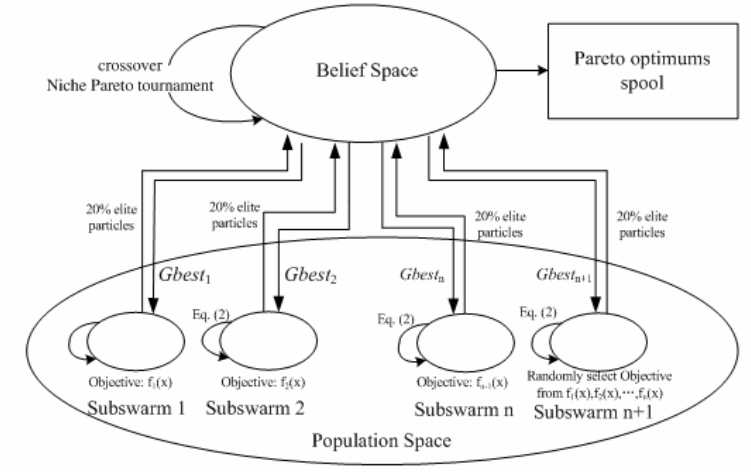


Fig. 2. Frame of cultural particle swarm optimization for MOP

Both the offsprings X_i' and X_j' are compared with their parent X_i and X_j. If X_i' and X_j' are nondominated, they replace their parent. Otherwise X_i' and X_j' will be discarded and their parent will be retained.

After crossover operation, a selection mechanism based on Niche Pareto tournament approach is executed to pick better particle for reproduction. The detailed selection procedure is as follows: Randomly take out two candidate particles, as well as a comparative set, which include a certain number of particles used for comparison, compare each candidate particle with the comparative set respectively, and two possible results maybe occur as:

(1) If one candidate particle is dominated by the comparative set and another is not, then the nondomiated one will be selected for reproduction.
(2) If both the candidate particle are dominated by or dominate the comparative set, then the one with less niche count will be selected for reproduction.

Niche count in procedure (2) can be obtained by computing the sum of sharing function of the whole swarm as follow [1]:

$$m_i = \sum_{i=1}^{swarm_size} sh(d_{i,j}), \tag{5}$$

where $sh(\cdot)$ is denoted as sharing function, which takes a power law function as

$$sh(d_{i,j}) = \begin{cases} 1 - \left(\dfrac{d_{i,j}}{\sigma_{share}} \right)^a & if\ d_{i,j} < \sigma_{share} \\ 0 & otherwise \end{cases}, \tag{6}$$

where σ_{share} is denoted as niche radius, a is constant and $d_{i,j}$ is distance between particle i and j measured in decision space.

The above Niche Pareto tournament approach ensures uniformly convergence to Pareto frontier, other than undesired premature.

As mentioned above, the belief space accepted the elite particles so as to be shared with the entire population. Whereas after the updating operation of crossover and selection, the belief space will supervise the evolution of the population space by adding a correction item to its velocity by updating formula in Eq. (1) as

$$v_{i,j}^{(k+1)} = v_{i,j}^{(k)} + c_1 r_1 (pbest_{i,j}^{(k)} - x_{i,j}^{(k)}) + c_2 r_2 (gbest_j^{(k)} - x_{i,j}^{(k)}) + c_3 r_3 (Gbest_j^{(k)} - x_{i,j}^{(k)}) \quad (7)$$

where c_3 is acceleration constant. r_3 is a random number within the interval of [0,0.5]. $Gbest_j$ is the position attained by the elite particle in belief space. Roulette wheel selection approach is employed to select one as $Gbest_j$ from the belief space.

To keep a set of Pareto optimums in the evolution, an independent external memory called population pool is allocated to store Pareto set as seen in Fig. 2. It is updated by deleting all dominated solutions and accepting new Pareto solutions from the belief space in each generation.

4.2 Constraints Handling

For general constrained optimization problem, some infeasible individuals maybe exist near the global optimum and holds high fitness values. Although they are infeasible in current iteration, further operations maybe make them to create new feasible offspring with higher fitness value. Thus it is helpful for the optimization to keep one small part of infeasible but good individuals. Inspired by [11,12,13], a feasibility-based constraint handling rule, called direct comparison-method is employed in this paper to handle constraints, it presents the comparison rules among individuals, as well as the adaptation strategy to keep certain proportion of infeasible individual in the population. To describe the magnitude that all inequality constraints are violated, a measuring function is defined as follows:

$$viol(x) = \sum_{j=1}^{m} f_j(x) \quad (8)$$

where $f_j(x)$, j =1,2,…,m, is a series of penalty functions to measure what extent each constraint is violated to, it is defined as:

$$f_j(x) = \begin{cases} \max\{0, g_j(x)\} & if \quad 1 \le j \le q \\ |h_j(x)| & if \quad q+1 \le j \le m \end{cases} \quad (9)$$

For a predefined constant ε ($\varepsilon > 0$), two individuals are compared and treated according to the follows handling rules:

(1) When two individuals are both feasible, select the one with higher fitness.
(2) When two individuals are both infeasible, select neither.
(3) When an individual x is feasible and another individual x' is infeasible, and if $viol(x) \le \varepsilon$, compare their fitness and select the one with higher fitness.

To maintain a rational proportion of infeasible individuals in the whole population, an adaptive updating strategy for ε can be done when every cycle of k_0 generations of evolution is completed, the update is as

$$\varepsilon_{new} = \begin{cases} 1.2\varepsilon_{old} & if \ \ a/pop_size \le p \\ 0.8\varepsilon_{old} & if \ \ a/pop_size > p \\ \varepsilon_{old} & else \end{cases} \tag{10}$$

where a is number of infeasible individuals, p is a predefined constant to determine the proportion of infeasible individuals.

For the two offsprings created by Eqs. (3) and (4) in the belief space, the above constraints handling rule can be directly used. For the particle updating formula in Eq. (1), the above handling rule can be executed as follows: suppose that $pbest^{(k)}$ represents *pbest* of the *i*th particle at generation *k,* and $x_i^{(k+1)}$ represents the newly generated position of the *i*th particle at generation *k*+1. $pbest^{(k)}$ will be replaced by $x_i^{(k+1)}$ at any of the following occasions:

(1) $pbest^{(k)}$ is infeasible, but $x_i^{(k+1)}$ is feasible.

(2) Both $pbest^{(k)}$ and $x_i^{(k+1)}$ are feasible, but $f(x_i^{(k+1)}) < f(pbest^{(k)})$.

(3) Both $pbest^{(k)}$ and $x_i^{(k+1)}$ are infeasible, but viol($x_i^{(k+1)}$) < viol($pbest^{(k)}$). Similarly, *gbest* is updated based on the above rule at every generation.

4.3 Procedure of CPSA

Our CPSA actually adopts a dual evolution mechanism. Each subswarm in the population space and the elite swarm in the belief space can evolve synchronously with a parallel multi-thread mode in one computer or multi-computer c/s network respectively to increase computing speed. However single-thread basic procedure of the CPSA can be described as follows:

```
Begin
t=0;
Initialize Population Space POP(t) and Belief Space BLF(t);
repeat
    repeat  //adjust Population Space
       For i=1 to n+1
           Evaluate subswarm i;
           Individual and global best positions updating
                for subswarm i;
           Velocity and position updating for subswarm i;
       End
     Until expected number of iterations achieved.
     //Adjust (BLF(t),Accept(POP(t)));
     Select 20% elite particles to belief space
     Crossover operation in belief space
     // Adjust(BLF(t));
     Update Pareto solutions set
     Niche Pareto tournament selection
```

```
      Update population pool
      Output Gbest_i to each subswarm  //Influence(POP(t));
Until termination condition achieved
End
```

5 Numerical Example

Two simple and typical single-variable and double-variable double-objective problem is first presented to demonstrate the proposed algorithm. It is described as follows:

(1) Example 1 : $\min \{ f_1 = x^2 ; \ f_2 = (x-2)^2 \quad \text{s. t. } x \in R^1 \}$

(2) Example 2: $f_1(x) = -25(x_1-2)^2 - (x_2-2)^2 - (x_3-1)^2 - (x_4-1)^2 - (x_5-1)^2$

$$f_2(x) = (x_1-1)^2 + (x_2-1)^2 + (x_3-1)^2 + (x_4-1)^2 - (x_5-1)^2$$

s. t. $\{ g_1(x) = x_1 + x_2 - 2 \geq 0 ; \ g_2(x) = 6 - x_1 - x_2 \geq 0 ; \ g_3(x) = 2 + x_1 - x_2 \geq 0 ;$

$g_4(x) = 2 - x_1 + 3x_2 \geq 0 ; \ g_5(x) = 4 - (x_3-3)^2 - x_4 \geq 0 ;$

$g_6(x) = (x_5-3)^2 + x_6 - 4 \geq 0 ; \ 10 \geq x_i \geq 0 , \ i = 1,2,3,4,5,6 \}$

Fig.3 presents the curves of objective functions f1(x) and f2(x) with decision variable x for example 1. Using the proposed algorithm, its optimums in criterion space are shown in Fig.4 and Fig.5 after 100 and 400 generations of iterations. The Pareto optimums are shown in Fig. 6 for example 2.

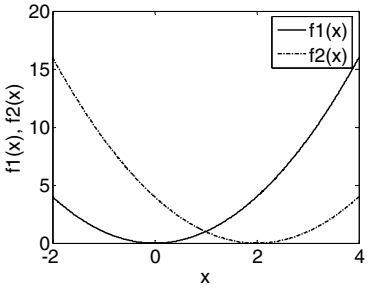


Fig. 3. $f_1(x)$ and $f_2(x)$ with x

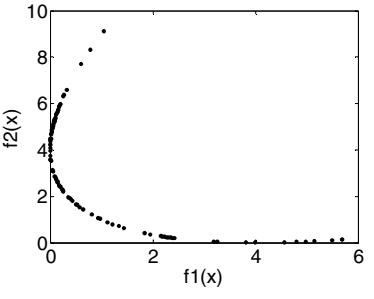


Fig. 4. $f_1(x)$ and $f_2(x)$ after 100 generations

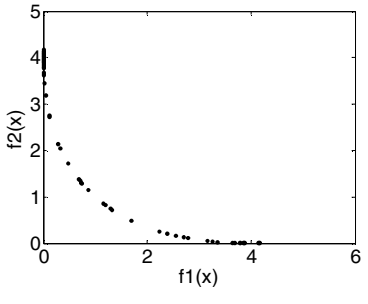


Fig. 5. $f_1(x)$ and $f_2(x)$ after 400 generations

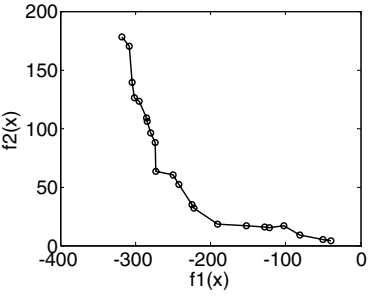


Fig.6. $f_1(x)$ and $f_2(x)$ of example 2

6 Conclusion

This paper has proposed a novel dual-evolutionary cultural particle swarm algorithm for constrained multi-objective optimization problem. Multiple particle subswarms make up of population space of cultural algorithms frame. These subswarms ensure to find each single-objective optimum, as well as a multi-objective solution. The belief space accepts a certain number of elite particles and takes crossover operation for Pareto optimums. Additional Nicho Pareto tournament selection can make Pareto set to be uniformly distributed in Pareto frontier. A direct comparison method for constraint handling overcomes the disadvantages of penalty function methods.

Simulation results showed that our proposed CPSA is of good performances, and more typical function test will be discussed in future.

References

1. Gen, M., Cheng R.: Genetic Algorithms and Engineering Optimization. John Wiley & Sons, Inc. (2000)
2. Schaffer, J.D.: Multiple objective optimization with vector evaluated genetic algorithms, in: Proceedings of an International Conference on Genetic Algorithms and Their Applications, Pittsburgh, PA, (1985)93–100
3. Srinivas, N., Deb, K.: Multiobjective optimization using nondominated sorting in genetic algorithms, Evolutionary Computation 2 (3) (1994) 221–248
4. Kennedy, J. E., Eberhart, R.: Particle Swarm Optimization. In: Proceedings of the IEEE International Conference on Neural Networks,Vol. 4. Perth. Australia. (1995) 1942-1948
5. Silva, A., Neves, A., and Costa, E.: An Empirical Comparison of Particle Swarm and Predator Prey Optimization. Lecture Notes Comput. Sci., 2464(2002) 103–110
6. Schutte, J. F., Groenword, A. A.: A Study of Global Optimization Using Particle Swarms. J. Global Optimiz., 31(2005) 93–108
7. Ho, S. L., Yang, S. Y., Ni, G. Z., and Wong, H. C.: A Particle Swarm Optimization Method with Enhanced Global Search Ability for Design Optimizations of Electromagnetic Devices. IEEE Transactions on Magnetics, 42(2006) 1107-1110
8. Reynolds, R. G.: An Introduction to Cultural Algorithms. Proceedings of the 3rd Annual Conference on Evolutionary Programming, World Scientific Publishing, (1994) 108–121
9. Jin, X.D., Reynolds, R. G.: Using Knowledge-Based Evolutionary Computation to Solve Nonlinear Constraint Optimization Problems: a Cultural Algorithm Approach, IEEE, (1999)1672-1678
10. Jin, X.D., Reynolds, R. G.: Mining Knowledge in Large Scale Databases Using Cultural Algorithms with Constraint Handling Mechanisms. Proceeding of the 2000 Congress on Evolutionary Computation., IEEE, (2000) 1498-1506
11. Li, M. Q., Kou, J. Z., Lin, D., Li, S. Q.: Based Theory and Application of Genetic algorithm, 3rd edition. Science Press, (2004)
12. Deb, K.: An Efficient Constraint Handling Method for Genetic Algorithms. Comput. Meth. Appl. Mech. Eng, 186 (2000) 311–338
13. He, Q., Wang, L.: A Hybrid Particle Swarm Optimization with a Feasibility-based Rule for Constrained Optimization, Applied Mathematics and Computation, (2006), doi:10.1016/j.amc.(2006) 134

A Novel Multi-objective Evolutionary Algorithm

Bojin Zheng[1] and Ting Hu[2]

[1] College of Computer Science,South-Central University For Nationalities, Wuhan,430074,China
[2] Department of Computer Science, Memorial University of Newfoundland, St. John's, NL, A1B 3X5, Canada
zhengbojin@gmail.com, tingh@cs.mun.ca

Abstract. Evolutionary Algorithms are recognized to be efficient to deal with Multi-objective Optimization Problems(MOPs) which are difficult to be solved with traditional methods. Here a new Multi-objective Optimization Evolutionary Algorithm named DGPS which is compound with Geometrical Pareto Selection Method (GPS), Weighted Sum Method (WSM) and Dynamical Evolutionary Algorithm (DEA) is proposed. Some famous benchmark functions are carried out to test this algorithm's performance and the numerical experiments show that this algorithm runs much faster than SPEA2, NSGAII, HPMOEA and can obtain finer approximate Pareto fronts which include thousands of well-distributed points.

Keywords: Multi-Objective Optimization, Evolutionary Algorithm, Geometrical Pareto Selection, Weighted Sum Method.

1 Introduction

Multi-objective optimization problems(MOPs)are very common in economics, and engineering etc., but they are a class of very difficult problems. In 1984, David Schaffer proposed Vector Evaluated Genetic Algorithm (VEGA) [1,2] to deal with MOPs. Since then many Multi-objective Optimization Evolutionary Algorithms (MOEAs) are proposed, most of them base on Pareto techniques. Since the works of E. Zitzler et al.[3,4,5] in 1999, the importance of elitism strategy in multi-objective search was recognized and supported experimentally.

In general, , the current popular MOEAs employ explicit or implicit "archive" [6,7] which is designed to store the non-dominated solutions to implement the elitism. So the schema of current MOEAs could be depicted as follow[8]:

$$\text{MOEA} = \text{Archive} + \text{Generator}$$

The equation means that if we want to design an efficient MOEA, we should pay attention to two aspects: firstly, design a more efficient algorithm to deal with archive, i.e., elitist space, secondly, design a more efficient algorithm in the production of new individuals.

As to the generator, hundreds of Evolutionary Algorithms would be meaningful to us. Among them, Dynamical Evolutionary Algorithm (DEA) [9] would be worth of researching. Zou [10] proposed High Performance MOEA (HPMOEA)

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 1029–1036, 2007.
© Springer-Verlag Berlin Heidelberg 2007

[11] which combines DEA with Ranking method and Niching method to deal with MOPs and reported good results.

As to the archiving algorithm, except the schema of "Ranking-alike operator + Niching-alike operator", there exists another schema, which is called "the sampling schema". Geometrical Pareto Selection (GPS) would be such a typical algorithm[12,13].

In this paper, we propose a novel MOEA which combines Geometrical Pareto Selection Method (GPS), Weighted Sum Method (WSM) and Dynamical Evolutionary Algorithm (DEA) together. The experimental results show that this algorithm runs much faster than SPEA2, NSGAII[14], HPMOEA and can obtain finer approximate Pareto fronts which include thousands of well-distributed points.

2 Background

In this section, we present three major components of DGPS to help understand our work. For the detail information please refer to [10],[13].

2.1 Geometrical Pareto Selection

For convenience of discussion, we only discuss how to deal with the two-objective optimization problems.

The method can be depicted as follows:

1. Use some techniques to estimate Pareto front,$F_1 \subset (L_1, U_1)$,$F_2 \subset (L_2, U_2)$. Here F is the objective functions.
2. Choose a constant integer κ, here κ is an angle constant which splits the Pareto front equally. Create an array: Array. And the array,i.e.,archive, is used to store the current best solutions (but not non-dominated solutions).
3. Choose a point A(x, y) which is far away from Pareto front.
4. When a new individual N(x_1,y_1) will be inserted into archive, use such rules:
   - Compute its slope α with point A(x,y).

     $$\alpha = \frac{y_1 - y}{x_1 - x}$$

   - Compute its distance dis with point A(x,y).

     $$dis = \sqrt{(y_1 - y)^2 + (x_1 - x)^2}$$

   - Compute the location pos in archive.

     $$pos = \left\lceil \frac{\alpha - \gamma}{\kappa} \right\rceil, \text{ here } \gamma = \frac{L_2 - y}{L_1 - x}$$

   - Modify the archive. If $dis > Array[pos].dis$, then $Array[pos] \coloneqq N(x_1, y_1)$.

For the above method, it's easy to generalize this method from 2-objective problems to N-objective problems. Namely, we can get the angles respectively by using the projective method. We should get N-1 angles totally and then find the very point to compare with the current point.

2.2 Dynamical Evolutionary Algorithm

DEA firstly is designed as a single objective optimizer. In DEA the individuals is considered as particles and the population as a dynamical system or particle system. So every particle is assigned a momentum and an activity, these two quantities are incorporated to control selections and to drive particles moving and searching all the time and everywhere. In DEA, the iterative step t in DEA like the generation in the traditional GA is called time t.

The momentum P of particle x_i at time t is defined as

$$P(t, x_i) = f(t, x_i) - f(t-1, x_i)$$

Here f is the value of objective function.

The activity of $a(t, x_i)$ particle x_i is defined as the count that x_i is selected as the parent individual at time t.

We can generally take a weight co-efficient $\lambda \in (0, 1)$ to indicate that which one of the two terms is more significant than the other in selection. Namely

$$slct(t, x_i) = \lambda \sum_{k=0}^{t} |p(k, x_i)| + (1-\lambda)a(k, x_i)$$

Based on $slct(t, x_i)$, not fitness, the individuals would be selected to evolve.

2.3 Weighted Sum Method

When searching in the solution space of MOPs, the individuals have to evolving toward multiple directions in common MOEAs. Actually, Evolutionary Algorithms are more efficient while evolving toward only one direction. So the Weighted Sum Method is employed to seek the extreme points, that is, the whole population is divided into several subpopulations, every subpopulation would evolving by a weight setting. Considering that trade-off solutions also are very important, an additional subpopulation is suggested to optimize the mean of all objectives.

While the objective functions are multi-modal, the method should cooperate with other operators to make the subpopulations evolve toward the right extreme points.

3 Introduction to New Algorithm

In this algorithm, GPS is an archiving algorithm, DEA is used as single-objective optimizer, WSM would guide the evolution of population. As to trade-off solutions, they will be generated by the optimizer during the optimization.

The algorithm can be depicted Figure 1:

Program DGPS;
1 Initialize Populations and archive
2 While not satisfy the stopping criteria do
3 For i=1 to SubPopNumber
4 Use DEA to optimize the corresponding objective
5 Try to insert every offspring into archive by GPS method
6 If successful in insertion or the offspring is better than its parent
7 then replace its parent
8 End For
9 End While
10 Using Cutoff strategy to clear all dominated solutions
11 Output all non-dominated solutions

Fig. 1. Pseudocode of DGPS

Note:
1) The definition of better

The better function is used to compare two individuals, we note the prototype as bool Better(Indi1, Indi2).When solving the single objective problems, DEA uses a simple strategy to compare two individuals. The individual whose fitness is smaller is the better(for minimization). But when solving the multi-objective problems, we modify the strategy as follows:

If the Indi1's corresponding objective value is smaller
then return true
Else if the Indi1's corresponding objective value equals to Indi2's
then if the average of all objectives of Indi1 is smaller
then return true
else return false

2) Cutoff Strategy

In GPS, some dominated solutions are not eliminated until the iteration stops. So after the iteration, these dominated solutions should be removed by this strategy.

4 Experimental Results and Analysis

We test this algorithm with some famous test-bed functions, the results are very satisfactory. Because HPMOEA employs DEA, but with different archiving algorithm and search strategy, we compare our results with it.

For a fair comparison, we set the same generation number and size of population (the same evaluation times) to both the algorithms.

Problem 1: KUR

$$\begin{aligned} &\min f_1(X) = \sum_{i=1}^{n-1} \left(-10 \exp(-0.2\sqrt{x_i^2 + x_{i+1}^2})\right) \\ &\min f_2(X) = \sum_{i=1}^{n} \left(|x_i|^{0.8} + 5 \sin x_i^3\right) \\ &x_i \in [-5\ ,\ 5], i = 1, \cdots, 3 \end{aligned} \tag{1}$$

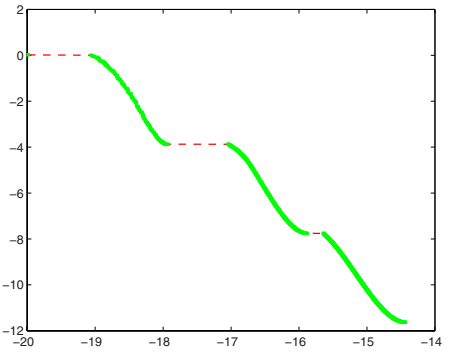


Fig. 2. APF of KUR by DGPS

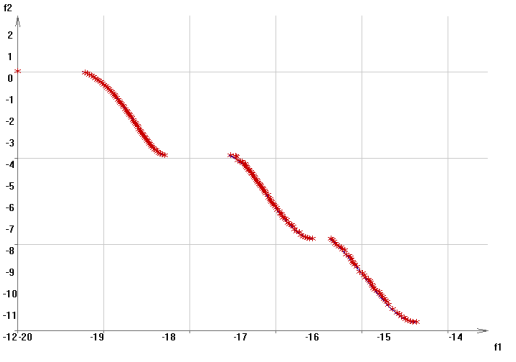


Fig. 3. APF of KUR by HPMOEA

The Pareto front of KUR is non-convex and disconnected. Because of the discrete nature of the Pareto-optimal regions, optimization algorithms may have difficulties in finding Pareto-optimal solutions in all regions. But both the proposed MOEA and HPMOEA obtain very good results, as shown in Figure 2 and 3.

Problem 2: VNT

$$\begin{aligned} &Min\ f_1(X) = 0.5(x_1^2 + x_2^2) + \sin(x_1^2 + x_2^2) \\ &Min\ f_2(X) = \frac{(3x_1 - 2x_2 + 4)^2}{8} + \frac{(x_1 - x_2 + 1)^2}{27} + 15 \\ &Min\ f_3(X) = \frac{1}{(x_1^2 + x_2^2 + 1)} - 1.1\exp(-x_1^2 - x_2^2) \\ &s.t.: \quad -3 \le x_1, x_2 \le 3 \end{aligned} \tag{2}$$

This is a difficult problem with high objective dimensions. But this algorithm generated a very fine pictures with obvious features, see Figure 4. From Figure 5, HPMOEA can not obtain the whole Pareto front, and its Pareto front does not include enough points.

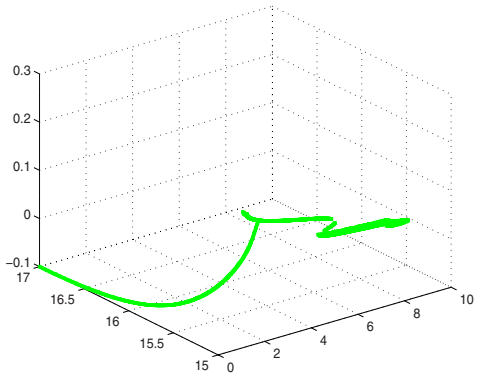


Fig. 4. POF of VNT by DGPS

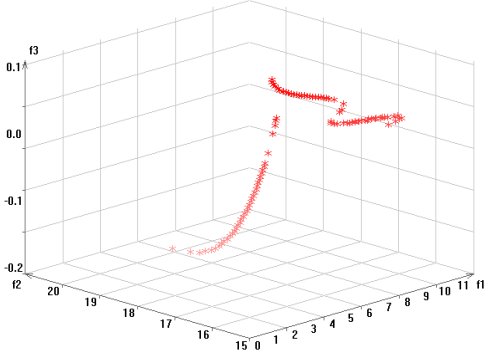


Fig. 5. POF of VNT by HPMOEA

From the pictures, we can see that this algorithm can obtain very fine curves in a single run. Obviously, this will contribute great convenience to decision-maker for the final decision and decease greatly the decision risk of lack of sampling points. It is remarkable that the sampling points distribute very equally.

In this algorithm, multi-subpopulation strategy make the population evolve towards multiple directions, so it is easier to obtain the extreme points. Though this algorithm get more sampling points, the consumed time still is less than some other MOEAs. We compare this algorithm with HPMOEA and other two famous MOEAs - SPEA2 and NSGA-II which base on PISA[15].

All the algorithms are carried out on a machine with one Intel PIV 2.4G CPU. In the experiments, every algorithm runs 30 times, the sizes of populations are set to 100, the numbers of generation of DGPS and HPMOEA are set to 2000 and the numberd of SPEA2 and NSGA-II are set to 200. The results are listed in Table 1.

Table 1. Comparison of Consumed Time

Problem	Algorithm	Avg. Num. of points	Average time
ZDT3	DGPS	1420.0	0.837s
	HPMOEA	100	21.840s
	SPEA2	100	73.006s
	NSGAII	100	72.646s
SRN	DGPS	2026.1	0.282s
	HPMOEA	100	29.815s
KUR	DGPS	1381.4	1.440s
	HPMOEA	100	19.761s

From Table 1 we can see that this algorithm runs much faster than HPMOEA and can get more sampling points, so we can conclude that this algorithm is better than HPMOEA. Literature[10] points out that HPMOEA have a better convergent rate than SPEA2, NSGAII, so we can conclude that this algorithm can run faster than SPEA2,NSGAII and get finer approximate Pareto Front.

5 Conclusions and Future Work

In this paper, we proposed a new MOEA. The experimental results show that this algorithm can obtain finer approximate Pareto front with less time. In contrast with the some previous MOEAs, which may spend hours to perform one run for only hundreds of non-dominated solutions, this algorithm can obtain tens of thousands of solutions in a single run under a bearable time limitation.

In this paper, The theoretical analyses such as convergence property, convergence ratio are not introduced here, we will leave it as the future work.

Acknowledgement. The authors gratefully acknowledge the financial support of the National Natural Science Foundation of China under Grant No.60473014 and No.60603008.

References

1. Schaffer, J.: Some Experiments in Machine Learning Using Vector Evaluated Genetic Algorithms. PhD thesis, Vanderbilt University (1984)
2. Schaffer, J.: Multiple objective optimization with vector evaluated genetic algorithms. In: Proceedings of the First International Conference on Genetic Algorithms. (1985) 93–100
3. Zitzler, E., Thiele, L.: Multiobjective evolutionary algorithms: a comparative case study and the strength pareto approach. Evolutionary Computation, IEEE Transactions on **3**(4) (1999) 257–271
4. Zitzler, E., Thiele, L., Laumanns, M., Fonseca, C.M., da Fonseca, V.G.: Performance assessment of multiobjective optimizers: an analysis and review. Evolutionary Computation, IEEE Transactions on **7**(2) (2003) 117–132

5. Laumanns, M., Zitzler, E., Thiele, L.: On the effects of archiving, elitism, and density based selection in evolutionary multi-objective optimization. Evolutionary Multi-Criterion Optimization, Proceedings **1993** (2001) 181–196
6. Knowles, J., Corne, D.: Properties of an adaptive archiving algorithm for storing nondominated vectors. Evolutionary Computation, IEEE Transactions on **7**(2) (2003) 100–116
7. Knowles, J.D., Corne, D.W., Fleischer, M.: Bounded archiving using the lebesgue measure. In: Evolutionary Computation, 2003. CEC '03. The 2003 Congress on. Volume 4. (2003) 2490–2497
8. Corne, D., Knowles, J.: Some multiobjective optimizers are better than others. In: Evolutionary Computation, 2003. CEC '03. The 2003 Congress on. Volume 4. (2003) 2506–2512
9. Li, Y., Zou, X., Kang, L., Michalewicz, Z.: A new dynamical evolutionary algorithm based on statistical mechanics. Journal of Computer Science and Technology **18**(3) (2003) 361–368
10. Zou, X.: Research On Theory of Dynamical Evolutionary Algorithm and their Applications. PhD thesis, Wuhan University (2003)
11. Zou, X.F., Kang, L.S.: Fast annealing genetic algorithm for multi-objective optimization problems. International Journal of Computer Mathematics **82**(8) (2005) 931–940
12. Zheng, B., Li, Y., Peng, S.: GPS: A geometric comparison-based pareto selection method. In Kang, L., Cai, Z., Yan, X., eds.: Progress in Intelligence Computation and Applications, International Symposium on Intelligent Computation and its Application, ISICA 2005. Volume 1., Wuhan,China (2005) 558– 562
13. Zheng, B.: Researches on Evolutionary Optimization. PhD thesis, Wuhan University (2006)
14. Deb, K., Pratap, A., Agarwal, S., Meyarivan, T.: A fast and elitist multiobjective genetic algorithm: Nsga-ii. Evolutionary Computation, IEEE Transactions on **6**(2) (2002) 182–197
15. Bleuler, S., Laumanns, M., Thiele, L., Zitzler, E.: Pisa - a platform and programming language independent interface for search algorithms. In Fonseca, C., ed.: Evolutionary Multi-Criterion Optimization(EMO 2003), LNCS 2632. (2003) 494–508

New Model for Multi-objective Evolutionary Algorithms

Bojin Zheng[1] and Yuanxiang Li[2]

[1] College of Computer Science, South-central University for Nationalities, Wuhan, 430074, China
[2] State Key Lab. of Software Engineering, Wuhan University, Wuhan, 430072, China
zhengbojin@gmail.com, yxli@whu.edu.cn

Abstract. Multi-Objective Evolutionary Algorithms (MOEAs) have been proved efficient to deal with Multi-objective Optimization Problems (MOPs). Until now tens of MOEAs have been proposed. The unified mode would provide a more systematic approach to build new MOEAs. Here a new model is proposed which includes two sub-models based on two classes of different schemas of MOEAs. According to the new model, some representatives algorithms are decomposed and some interesting issues are discussed.

Keywords: Multi-objective Optimization, Framework, Evolutionary Algorithm, Unified Model.

1 Introduction

Evolutionary Algorithms are an randomized searching approach based on Darwin's evolutionary theory. They play an important role in many fields such as optimization, control, game strategies, machine learning, and engineering design etc.

In 1984, David Schaffer introduced Vector Evaluated Genetic Algorithm (VEGA)[1,2] to solve Multi-objective Optimization Problems(MOPs). Henceforth, the research on Multi-Objective Evolutionary Algorithms(MOEAs) attracted more and more researchers. Up to now tens of MOEAs have been proposed.

To guide the efforts on MOEAs, some researchers tried to build unified models for popular MOEAs. For examples, Macro Laumanns et al.[3] proposed a unified model for the Pareto-based and elitist MOEAs in 2000. This model can describe most popular famous MOEAs, such as Non-dominated Sorting Genetic Algorithm II(NSGA-II[4], Strength Pareto Evolutionary Algorithm and its improvement(SPEA/SPEA2)[5], Pareto Archived Evolution Strategy(PAES)[6] and so on. [7] expressed the schema of MOEAs which employ archive with such a formula as follows:

$$\text{MOEA} = \text{Archive} + \text{Generator}$$

But this formula is quite simple. Recently, more and more MOEAs can not be accurately described by these models, such as Adaptive Grid Algorithm (AGA)[8],Rank-Density based Genetic Algorithm (RDGA)[9], Geometrical Pareto Selection (GPS)[10,11] and GUIDED [12] etc.

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 1037–1044, 2007.
© Springer-Verlag Berlin Heidelberg 2007

In this paper, we propose a new model to describe the advanced MOEAs. In section 2,the model is introduced . And then SPEA[5] ,AGA[8] and GPS [10,11] are decomposed in section 3 according to this model. Some interesting issues are discussed in section 4. In section 5, some conclusions are made.

2 Introduction to the New Model

The first MOEA – VEGA – is a non-Pareto algorithm. Subsequently, Goldberg D. E. [13] proposed to use Pareto dominance to compute the fitness of the individuals based on 'Ranking' method. Subsequent experiments prove that Pareto dominance based MOEAs are more efficient than non-Pareto MOEAs. Since the work of Zitzler et al.[5], the 'elitism' of MOEAs has been recognized: Elitism of MOEAs is especially beneficial in deal with MOPs and the use of elitism can speed up the convergence to the Pareto front. the Pareto based MOEAs with elitism is more efficient than the MOEAs without elitism. To implement the elitism, many MOEAs use a secondary 'elitist' population ,i.e., the archive, to store the elite individuals. According to MOEAs' formula, the pseudocode of common elitist MOEAs with archive can be depicted as Figure 1:

```
1 initialize the population and archive
2 evaluate the population
3 while the termination criterions have not been reached do
4     generate a solution by the generator
5     evaluate the new solution
6     try to update the archive
7     according to the feedback of archive, try to update the population
8 end while
```

Fig. 1. Pseudocode for Generic MOEAs with Archive

The pseudocode seemly does not mention the generation gap methods. Actually, the generation gap methods can be decomposed into this model, if we see the generation number and the replaced parent individuals as the additional parameters. Moreover, though Single-Objective Evolutionary Algorithms(SOEAs) and MOEAs are very similar, there still are three major differences:

1. Different to single-objective optimization, the generator of MOEAs may crossover some individuals in population with the individuals in population or archive
2. the fitness assignment is more complicated, because it is relative to two operators: fitness evaluation for the archive and fitness evaluation for the population
3. As to the elitism, the SOEAs just keeps only one fittest individual. But in MOEAs, the elitism, commonly, the strategy to update the archive is quiet complicated.

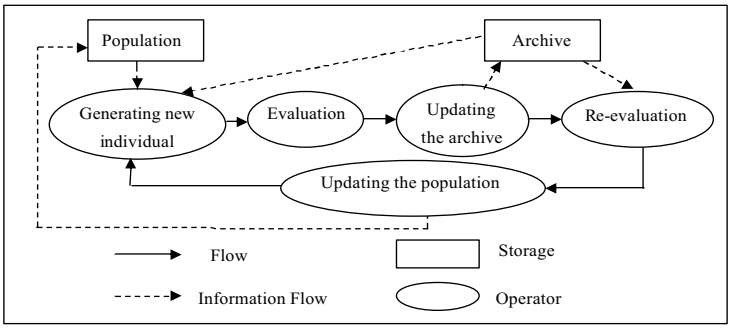


Fig. 2. The Model of Elitism MOEAs with Archive

In general, the new model can be depicted as Figure 2:

In Figure 2, updating the archive would retrieve information from the archive, so the link is not drawn in this framework. Moreover, the generator is similar to the generating process of SOEA. Actually, except the selection operators, they both are same. It can be depicted as Figure 3.

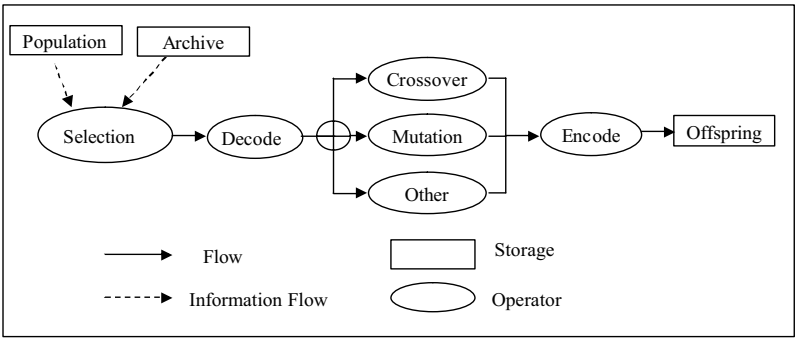


Fig. 3. The Framework for the Generator

Secondly, the strategy of updating the archive is different to SOEA and very complicated.

Very many MOEAs employ the ranking-alike operators and the niching-alike operators. In such a schema, ranking-alike operators are firstly employed to eliminate the dominated solutions and secondly niching-alike operators are employed to eliminate the crowded solutions. But unfortunately, this kind of MOEAs are not convergent[14] because of fitness deterioration. The schema of ranking-alike and niching-alike MOEAs(RN_MOEA) could be depicted as Figure 4.

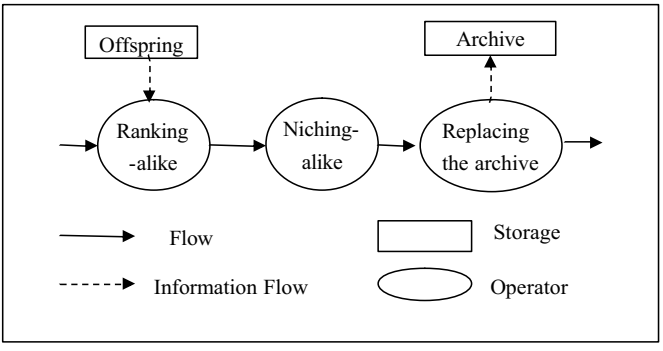


Fig. 4. The Ranking-alike and Niching-alike Schema

Actually, except this schema, there exists another schema. We call it 'the sampling schema'. In this schema, the feasible solution space (includes Pareto optimal front) is divided into grids in advance, when new solution is generated and evaluated, the algorithm firstly computes its coordinate in the grids and compare it with the individual(s) in the right coordinate. Whether updating the archive with the new solution or not just depends on the comparison. Obviously, this kind of methods are very different to ranking-alike and niching-alike methods, they use 'local dominance' instead of 'global dominance', and therefore hold lower time complexity. Some of them do not eliminate the dominated solutions from the archive in the main loop of algorithm, so additional operation(eliminating operator) should be employed after the main loop to cut the dominated solutions off from the archive. But if the archive should only store nondominated solutions, the eliminating operator should be integrated into the main loop. As to the diversity of Pareto optimal front, it depends only on the generator, because the span of cells in the grids has been predefined, may adaptively. The schema of sampling MOEAs(SA_MOEA) can be depicted as Figure 5.

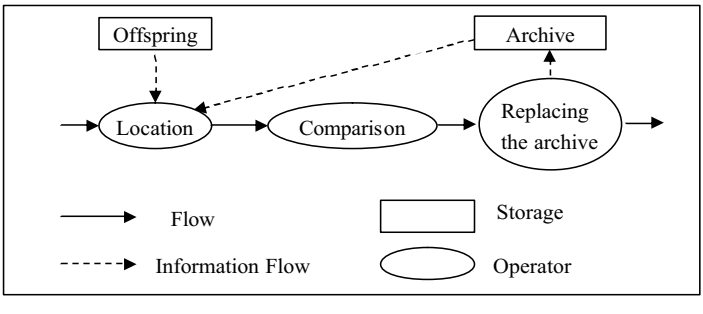


Fig. 5. The Sampling Schema

3 Decomposing Existing MOEAs

In this section, we will use SPEA , AGA and GPS to show how to decompose the existing algorithms.

3.1 SPEA

As to SPEA, it is a classical RN_MOEA. SPEA's mating selection process is a typical selection operator which uses archive and population. clustering method actually is an operator to keep diversity,i.e., niching-alike operator. Moreover, the truncate and update function can be seen as the strategies to update the archive and population.

3.2 AGA

Actually, AGA[8] is a typical archiving algorithm. It must combine a generator to become MOEAs. AGA employ a nondominated vectors archive, so it employs **Is_Dominated** function and **Dominates** function to eliminate the dominated solutions. Moreover, AGA uses adaptive grids, that is, the boundaries should be extended or reduced. However, in spite of these details, AGA samples the feasible solution space in essence. **Reduce_Crowding** function, **Steady_State** function and **Fill** function actually perform the comparison and replacement.

3.3 GPS

GPS is also an archiving algorithm. It employs the **Location** operator to retrieve the right solutions which will be used to compare with the new solution, employs **Comparison** operator to perform the comparison, and at last employs the **Steady_State** operator to update the archive with the new solution if the new solution is better than the original solution in the archive. Moreover, the archive of GPS may store the dominated solution, only when the main loop ends, an additional operation is used to eliminate the dominated solutions.

4 Some Important Issues

According to the model, there are two kinds of archiving algorithms. Based on different schema, MOEAs would behave differently, and they would have different properties.

4.1 Performance Measure

The convergence property is very important to MOEAs. It would theoretically determine the approximation degree to the true Pareto front. But good diversity would be very helpful of the decision-maker. That is, MOEAs had better converge with diversity. So the performance measure should take both the convergence and diversity into considerations.

AGA has been proved convergent with well-distributed solutions under certain strict conditions. GPS also converges to Ray-Pareto optimal front. In contrast to SA_MOEAs, the RN_MOEAs do not converge. Furthermore, the SA_MOEAs could be improved to converge to true Pareto front under certain conditions.

The complexity of MOEAs would be another aspect of performance measure. If two multi-objective approaches have different archivers, their average performance may differ[7]. Because of ranking method, the time complexity of RN_MOEAs would be greater than or equal to $O(MN)$ where computing one new individual, here M is the number of objectives, N is the size of population (or archive). As to AGA and GPS, the time complexity is $O(M)$. Niching-alike operators often hold a space complexity of $O(N^2)$. But AGA's is $O(N^M)$, GPS's is $O(N)$ at the best situation,$O(N^{M-1})$ at the worst situation. Furthermore, we can reduce GPS's space complexity to $O(N)$ by using binary tree with an additional average time complexity of $O(Nlog_2N)$.

4.2 Cooperation Between Generator and Archive

In this model, the generator should cooperate with archive to control the evolving directions. Therefore, considering to deal with difficult objective functions, 'local search' may be used to exploit. As to the archive of SA_MOEAs, 'local dominance' is useful to reduce the time complexity. 'local search' and 'local dominance' are different concepts.

The evolving directions of the population are multi-objective. In one hand, the selection pressure should make the individuals evolving toward the true Pareto front, i.e., depth-first search. In the other hand, the selection pressure should make the individuals spread over the whole Pareto front, i.e., width-first search. How to deal with the conflict between depth-first search and width-first search is still lack of delicate research. As to SA_MOEAs, because of local dominance, the feedback of archive just provide information for depth-first search, less for width-first search.

4.3 Taxonomy

Based the proposed model, we suggest that MOEAs could be categorized into four classes:

1. Non-Pareto MOEAs
   The representative MOEA is VEGA[1,2]. This algorithm employs multiple sub-populations to optimize every single objective separately. This algorithm often converge to special points which often are not Pareto Optimal points, moreover, the diversity is not taken into consideration.
2. Pareto MOEAs (without Elitism)
   The representative MOEAs include Multi-Objective Genetic Algorithm (MOGA) [15], Niched Pareto Genetic Algorithm (NPGA) [16,17] and Non-dominated Sorting Genetic Algorithm (NSGA)[18]. These algorithms employ some strategies to maintain the diversity, but approximation is not good enough.

3. Pareto MOEAs with Elitism
   The representative MOEAs include NSGA-II[4] , Strength Pareto Evolutionary Algorithm and its improvement(SPEA/SPEA2)[5]. These algorithms employ elitism strategy to maintain good approximation. But these algorithms are not convergent.
4. Convergent MOEAs
   Actually, the archiving algorithms determine the convergence property of MOEAs. The representative algorithms include Adaptive Grid Algorithm (AGA) [8], GPS [10,11]. These algorithms should converge/pseudoconverge under certain conditions.

5 Conclusions and Future Work

The proposed model is intended to understand the-state-of-the-art MOEAs and provides a more systematic approach to design more efficient and more customized MOEAs for researchers and possible users.

Our model implies that the ranking-alike and niching-alike schema is very different to the sampling schema,though they both may use archive to store elitist solutions.

In contrast to the previous unified models, the new model can describe the-state-of-the-art MOEAs more accurately and be more atomic. So it is more convenient to use this model for the analysis of the algorithms.

This model provide us many cues to improve the MOEAs, such as the relationship between 'local search' and 'local dominance', the relationship between evaluation operation and re-evaluation operator and the relationship between depth-first search and width-search etc. The future work would try to discover more principles and develop new operators based on this model.

Acknowledgement. The authors gratefully acknowledge the financial support of the National Natural Science Foundation of China under Grant No.60473014 and No.60603008.

References

1. Schaffer, J.: Some Experiments in Machine Learning Using Vector Evaluated Genetic Algorithms. PhD thesis, Vanderbilt University (1984)
2. Schaffer, J.: Multiple objective optimization with vector evaluated genetic algorithms. In: Proceedings of the First International Conference on Genetic Algorithms. (1985) 93–100
3. Laumanns, M., Zitzler, E., Thiele, L.: A unified model for multi-objective evolutionary algorithms with elitism. In: Congress on Evolutionary Computation (CEC 2000), IEEE Press (2000) 46–53
4. Deb, K., Pratap, A., Agarwal, S., Meyarivan, T.: A fast and elitist multiobjective genetic algorithm: Nsga-ii. Evolutionary Computation, IEEE Transactions on **6**(2) (2002) 182–197

5. Zitzler, E., Thiele, L.: Multiobjective evolutionary algorithms: a comparative case study and the strength pareto approach. Evolutionary Computation, IEEE Transactions on **3**(4) (1999) 257–271
6. Knowles, J.D., Corne, D.: Approximating the nondominated front using the pareto archived evolution strategy. Evolutionary Computation **8**(2) (2000) 149–172
7. Corne, D., Knowles, J.: Some multiobjective optimizers are better than others. In: Evolutionary Computation, 2003. CEC '03. The 2003 Congress on. Volume 4. (2003) 2506–2512
8. Knowles, J., Corne, D.: Properties of an adaptive archiving algorithm for storing nondominated vectors. Evolutionary Computation, IEEE Transactions on **7**(2) (2003) 100–116
9. Haiming, L., Yen, G.G.: Rank-density-based multiobjective genetic algorithm and benchmark test function study. Evolutionary Computation, IEEE Transactions on **7**(4) (2003) 325–343
10. Zheng, B., Li, Y., Peng, S.: GPS: A geometric comparison-based pareto selection method. In Kang, L., Cai, Z., Yan, X., eds.: Progress in Intelligence Computation and Applications, International Symposium on Intelligent Computation and its Application, ISICA 2005. Volume 1., Wuhan,China (2005) 558– 562
11. Zheng, B.: Researches on Evolutionary Optimization. PhD thesis, Wuhan University (2006)
12. Bui, L.T., Deb, K., Abbass, H.A., Essam, D.: Dual guidance in evolutionary multiobjective optimization by localization. In: The 6th International Conference on Simulated Evolution and Learning. Volume 4247 of Lecture Notes in Computer Science., Hefei, China, Springer Verlag, Heidelberg, D-69121, Germany (2006) 384–391
13. David, E.G.: Genetic Algorithms in Search, Optimization and Machine Learning, Reading, Massachusetts (1989)
14. Hanne, T.: On the convergence of multiobjective evolutionary algorithms. European Journal of Operational Research **117** (1999) 553–564
15. Tadahiko, M., Hisao, I.: Moga: Multi-objective genetic algorithms. In: Proceedings of the 2nd IEEE International Conference on Evolutionary Computing, Perth, Australia (1995) 289–294
16. Jeffrey, H., Nicholas, N., David, E.G.: A niched pareto genetic algorithm for multiobjective optimization. In: Proceedings of the First IEEE Conference on Evolutionary Computation, IEEE World Congress on Computational Intelligence. Volume 1., Piscataway, New Jersey (1994) 82–87
17. Igor, E.G., Sushil, J.L., Roberto, C.M.: Parallel implementation of niched pareto genetic algorithm code for x-ray plasma spectroscopy. In: Late Breaking Papers at the 2000 Genetic and Evolutionary Computation Conference, Las Vegas, Nevada (2000) 222–227
18. Srinivas, N., Kalyanmoy, D.: Multiobjective optimization using nondominated sorting in genetic algorithms. Evolutionary Computation **2**(3) (1994) 221–248

The Study on a New Immune Optimization Routing Model

Jun Qin, Jiang-qing Wang, and Zi-mao Li

College of ComputerScience, South-central University for Nationalities,Wuhan,430074,China
dr.qinj@gmail.com, wjqing2000@yahoo.com.cn, lizm@sdu.edu.cn

Abstract. The integrated network (mixed with fixed network, cellular wireless network and ad hoc network etc.) would have different dynamic properties to the single-technique-based network, such as the movement of nodes and the change of delay of links. In this paper, we propose a new dynamic multicast routing model with a mechanism called local rearrangement to handle the changes in integrated network. Furthermore, D-IOA, an immune algorithm based on clone process, which employs gene library to improve the effectiveness of D-IOA to meet the real-time requirement in online multicast routing fields, is given to optimize the multicast sub-tree within the range of local rearrangement. The simulation results indicate that our algorithm balances three metrics better compared with other two famous dynamic multicast routing algorithms.

Keywords: Integrated Network, Dynamic Multicast Routing, Immune Algorithm, Gene Library.

1 Introduction

With the proliferation of wireless communication and individual communication, more and more various connecting techniques appear. It can be seen that future network will be an integration of various network techniques such as fixed network, cellular wireless network and Ad-hoc network, which is called "integrated network" in this paper. In order to guarantee the end-to-end Service of Quality (QoS), the seamless communication between users in the integrated network becomes a very challenging issue. Especially, with the importance of routing techniques for multicast connections being emphasized, how to conduct multicast routing in this kind of integrated network environment should be a key research issue.

Multicast now has been viewed as a very important facility in communicating networks with the popularity of multimedia applications such as radio, TV, on-demand video and teleconferences[1]. For dynamic multicast routing, there have been a few of research to do it such as DGA (Dynamic greedy algorithm), SP (Source rooted shortest path), DP(Dynamic Prim), LRA (Lagrangian-Relaxation Algorithm) and DCMTCD, etc [2] which are good inexpensive heuristics for the dynamic multicast routing.

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 1045–1052, 2007.
© Springer-Verlag Berlin Heidelberg 2007

However, considering that there exist many connecting techniques in the integrated network as we mentioned before, there could be many different possible dynamics such as the node's move, the change of state of link etc. All those changes, sometimes even much, affect the multicast tree. However, current dynamic multicasting routing models are not considering these kinds of updates but only the membership change. Accordingly, those models aren't able to match the dynamics of the integrated network very well.

The rest of this paper is organized as follows. In section 2, we will give a new dynamic multicast routing model suitable for the integrated network. In section 3, "local rearrangement" mechanism is introduced. In section 4, an immune algorithm is proposed to optimize the multicast sub-trees within the range of the local rearrangement and in Section 5 and the last section, some simulations and summaries are given.

2 A New Dynamic Multicast Routing Model

In order to simulate the dynamics of integrated network better, in this paper, besides membership updates, we will take another two dynamics: node movement and the delay of link changes into consideration. That is to say, we will focus on a new dynamic multicast routing model specific to integrated network. It is obvious that it is much more challenging than static multicast model. As the first step of our research plan about dynamic routing research in integrated network, we will simplify properly the situation discussed above. We will ignore the network topology change (the link appear and/or disappear) due to the movement of nodes.

Definition 1 (Multiple Dynamic Constrained Multicast Routing problem (MDCMR)). *Given a undirected graph G (V, E); a source node s, a set of initial destination nodes* M_0*, the initial delay of any link e* $Delay_0(e)$*, a series of link delay variation* $U=\{u_0, u_1 \ldots u_i, ..\}$*, where* u_i *is a two-tuple (*e_i*,* $Delay_i$*),* $e_i \in E$*,* $Delay_i$ *is updated delay of* e_i*; a series of connection request* $R=\{r_0, r_1, \ldots r_i \ldots\}$*, where* r_i *is a two-tuple* (y_i, o_i)*,* $y_i \in v$*,* $o_i \in \{join, leave\}$*, a series of movement of node* $H=\{h_0, h_1, \ldots, h_g\}$*, where* h_i *is a two-tuple (*Old_i*,* New_i*),* Old_i *and* New_i *are old and new location of the moving node i .*

As we can see, these three two-tuple series stand for link state change, multicast membership change and node movement, respectively. Supposing M_i is the destination node set after suffering u_i, h_i or r_i , The aim of MDCMR is to find a sequence of tree $\{T_0, T_1, \ldots T_P\}$ where each of them covers $S \cup Mi$ and satisfies:

$$\text{Cost } (T_i) = \text{Min} \sum_{e \in E_{T_i}} \cos t(e)$$

$$\text{Delay } (\text{Path}(s, d)) < \Delta, \forall d \in M_i$$

The dynamics of MDCMR has three possibilities: the link change, the membership change and the movement of node. It is simple logically to analyse under which conditions the current multicast tree T_{i-1}, possibly need to be reconstructed .

3 Local Rearrangement Mechanism

Before we discuss the routing strategy based on our new dynamic multicast routing model, there are two main principles should be considered carefully:

(1) Track the optimal or near-optimal multicast routing during the whole multicast session, which is described as the optimization model defined in this section.

(2) Reduce the extension and frequency of adaptation to current multicast tree. This is because if we adapt the current multicast tree dramatically and frequently, the current multicast session must be interrupted often, and hence it is very hard to guarantee the QoS.

In some sense, although we can consider MDCMR as a series of static multicast trees optimization problem, optimizing a multicast tree from scratch will be too costly time-consumed to satisfy the real- time update. Moreover, the QoS requirement of multicast sessions cannot tolerate frequent dramatic changes. Thus, to achieve this balance between cost, time and extension of changes of the tree, we will adopt an idea [3] of accumulating the impact to a tree and triggering a rearrangement based on a threshold represented by QF which is defined in literature [3] to measure the extension of changes to a multicast tree. Further, even when it's time to rearrange a multicast tree, we don't need to optimize the whole multicast tree from the scratch. Just the part of a tree affected directly by the changes is chosen to rearrange.

The details of how to decide the range of sub-tree and the related changes' handling are the similar to the process used in Ref [3].

4 Optimization of Multicast Sub-tree

As soon as the rearrangement of the sub-Tree is needed, an algorithm called D-IOA is used to reconstruct the sub-Tree and the newly generated sub-tree will be later combined with the left part of original multicast tree to form a complete multicast tree without loops.

Up to now, the problem in front of us has become a kind of static multicast tree construction problem within part of original multicast tree (in the range of sub-Tree). Here, we will utilize immune algorithm frame based on population search to solve the typical optimization problem. Based on our previous successful research in static multicast routing domain [4], still, the main component of D-IOA is clone process, a famous immune process in AIS, which is composed of two phrases: clone (same meaning as copy) and mutation.

But it is noticeable that our task is to compute online the multicast route and hence the real-time requirement is highlight considered by the optimization algorithms. Hence we have to give the revised version of IOA called dynamic IOA (D-IOA) to match the real-time feature. To achieve this, "gene library" will be introduced into our algorithm.

The details of D-IOA are given in Fig. 1.

Input: A given network $G = (V, E)$, the state of network, a revised multicast session request $(s^{'}, M^{'}, \Delta^{'})$
Output: An optimized multicast sub-trees $(s^{'}, M^{'})$
1. Use the Dijkstra K-th shortest algorithm to construct the gene library, and delete the paths violating the delay constraints.
2. Initialize the population P_0, where $\lfloor d_0\% \cdot |P_0| \rfloor$ individuals produced from gene library and the remaining individuals by random construction. Those individuals violating the delay constraint should be deleted and re-produced again. T=0.
3. Delete the ring paths existed in individuals.
4. Compute the cost of each individual as the fitness f_i of it.
5. Order the individuals based on f_i of individuals in non-decreasing order. Suppose the rank of each individual is $Rank_i, 1 \leq i \leq |P(t)|$.
6. For each individual T_i, compute $p_i = \frac{1/f_i}{\sum_{j=1}^{|P(t)|} 1/f_j}$, and then clone (copy) $\delta \cdot |P(t)| \cdot p_i$ individuals to produce clone population Q_i.
7. Conduct mutation operator for each clone population $Q_i$and then select the best multicast tree into $P(t+1)$
8. If the size of population ¿$PopSize$, delete$|P(t)| - PopSize$individuals with biggest f_i value.
9. t = t + 1.
10. If not satisfy the termination condition, goto 4; else goto 11.
11. return the best multicast tree in population

Fig. 1. The pseudocode of D-IOA

In D-IOA, An individual means a multicast sub-tree. There are three important parameters: population size $PopSize$, the size of gene library k, the rate of initial individuals produced from gene library $d_0\%$.

Gene library is composed of good genes. A new antibody produced from gene library has good quality with much higher possibility than produced randomly. In gene library, there is more than one path from source node to destination nodes. Which are limited within the range of sub-tree. Dijsktra K-th shortest path algorithm is used to construct gene library. The details are that for each pair of source node and destination node, Dijkstra K-th shortest path algorithm is used to find K least-delay paths satisfying the delay constraint. The paths set for all pair nodes forms gene library and the size of it should be K* $|M'|$.

5 Experimental Results

In this section, we present the results of the simulation experiments that were conducted to analyze the performance of our algorithm. We compared our algorithm with two other famous algorithms. We will first define the performance metrics, then describe the simulation model, and finally present and discuss the results.

5.1 Performance Metrics

To analyze the performance of our algorithm, we used the following metrics:

(1) Cost Ratio (CR) and Average Cost Ratio (ACR). Those two metrics is a measure of how well a given algorithm performs (as regards cost of tree) in relation to the optimal algorithm. Since the online multicast problem is NP-complete, it will be impractical to use the optimal algorithms as a basis for comparison. Hence, we use the BSMA heuristic proposed by Zhu [5] as our benchmark. Suppose that $T_i, 1 \leq i \leq n$ and $T_i^{'}, 1 \leq i \leq n$ are the sequences of trees respectively constructed by our algorithm and BSMA algorithm after undergoing n changes in network. Then, the CR is given by the expression:$CR_i = \frac{C(T_i)}{C(T_i^{'})}$, where $C(T)$ is the cost of tree T. the ACR is defined as:$ACR = \frac{\sum_{i=1}^{n} CR_i}{n}$.

(2) Tree Change (TC) and Average Tree Change (ATC): we define TC as follow: for a sequence of n update requests, let an algorithm produce the sequence of $T_i, 1 \leq i \leq n$, then TC is given by $TC_i = |(E(T_i) - E(T_{i+1})) \cup (E(T_{i+1}) - E(T_i))|$; and ATC is defined as:$ATC = \frac{\sum_{i=1}^{n-1} TC_i}{n-1}$, where $E(T)$ denotes the set of edges in tree T and $|\bullet|$ denotes the size of the set. A lower value of TC or ATC indicates that tested algorithm is better able to accommodate changes in the group without excessively modifying the tree.

(3) CPU Time (CT) and Average CPU Time (ACT). For our simulation purpose, $CT_i, 1 \leq i \leq n$ (or ACT) is defined as the (average) time taken by a single (or by sequential) update of network for the algorithm.

5.2 Simulation Model

To generate four kinds of possible changes in the network discussed above for our simulation, we have used the probabilistic model similar to the model employed in [3]. In a network of size N, let$k$ represents the number of member nodes in the multicast tree. Supposing the probabilities of an add-request, a remove-request, the movement of a multicast node and a delay change of a link are defined as the way in [3].

5.3 Simulation Parameters and Simulation Results

Our simulation studies were conducted on a set of 100 random networks. A random graph model proposed by Waxman [6] was used to produce the network. This ensures that the simulation results are independent of the characteristics of any particular network topology.

The size of gene library k=10. The dynamic change of those four types are chosen randomly with the probability expression above. The threshold parameter to trigger a rearrangement algorithm is $\rho = 0.6$. In order to verify the performance of D-IOA, we compare our algorithm with DCMTCD [7] and LRA [8] algorithms. The main parameters used in D-IOA is P_0=20, d_0%=80%, θ=0.4. Every simulation result is the average of 50 running for tested algorithms. The

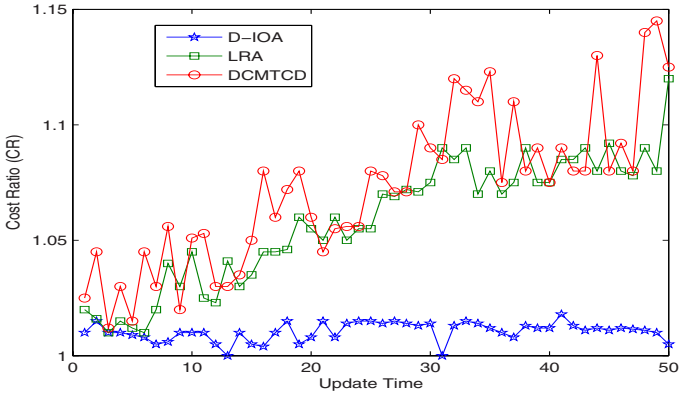


Fig. 2. the dynamic curve of CR

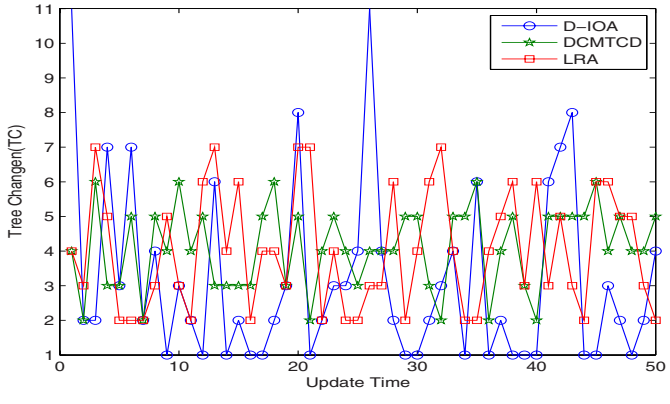


Fig. 3. the dynamic curve of TC

initial multicast group size is 20. Each simulation will be suffered 50 changes with only single change considered each time. All algorithms are implemented on a P4 2.4GHz machine with 256Mb memory. Fig.2, Fig.3 and Fig.4 shows respectively the CR, TC, CT curves along with the 50 dynamic changes for the three algorithms. Table 1 shows the values of ACR, ATC, ACT.

As we can see from Fig. 2 and Table 1, the multicast tree found by D-IOA does not suffer an obvious cost increase along with the dynamic change of network while the DCMTCD and LRA do. The key different Between D-IOA and the latter two algorithms is that the latter two algorithms do not have the rearrangement mechanism, which means that rearrangement mechanism used in D-IOA indeed help a lot to improve the quality of multicast trees. From the Fig.3 and Table1, we can see that the average tree change (ATC) got by D-IOA is still lower than both the DCMTCD and LRA, although TC value become higher at

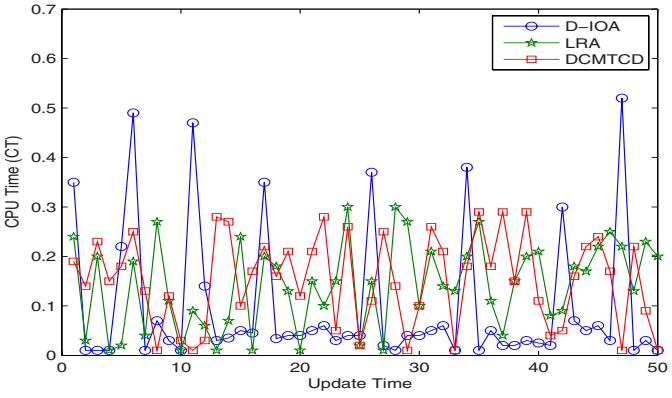


Fig. 4. The dynamic curve of CT

Table 1. The average performance of the three algorithms

	ACR	ATC	ACT
D-IOA	1.013	3.7	0.087
LRA	1.084	5.8	0.153
DCMTCD	1.049	4.9	0.142

the right time rearrangement triggered. Similarly, from Fig.4 and Table 1, we can see that the CT used by D-IOA is shorter than DCMTCD and LRA. Even though our algorithm introduces the rearrangement mechanism, we still have acceptable computation overhead because we limit the rearrangement to a local range rather than a whole multicast tree. Taking all the three aspects discussed here to be considered, our algorithm improves the performance of the dynamic multicast routing problem obviously.

6 Summary

In this paper, we proposed a new dynamic multicast routing model and a new algorithm to solve it. In this new model, we firstly considered another two dynamics ignored before in integrated network: the movement of node and the delay change of links. Both of them happen often in integrated network. In order to meet two important and contradicting goals: tracking the least-cost after each update and minimizing the frequent large changes to the multicast tree, we designed a local rearrangement mechanism and related algorithm to optimize the multicast sub-tree within local rearrangement range. The simulation results revealed that our algorithm balanced the two contradicting goals very well due to the introduction of local rearrangement mechanism in our algorithm. Also, the use of gene library further improves the effectiveness of rearrangement algorithm in that it makes best use of priori knowledge.

Our further research will include: the sensitivity analysis of parameters in our algorithm; the convergence analysis of our algorithm; extensive version of our algorithm with multiple constraints such as delay jitter, bandwidth, packet loss, etc.

Acknowledgement. This study was supported by China NSF grant (No.60603008) and Hubei province PSF grant (No.2004ABA029).

References

1. Ravikumar, C.P., Bajpai, R.: Source-based delay-bounded multicasting in multimedia networks. Computer Communications **21** (1998) 126–132
2. Cobb, J.A.: Dynamic multicast trees. In: Networks, 1999. (ICON '99) Proceedings. IEEE International Conference on. (1999) 29–36
3. Raghavan, S., Manimaran, G., Siva Ram Murthy, C.: A rearrangeable algorithm for the construction of delay-constrained dynamic multicast trees. Networking, IEEE/ACM Transactions on **7**(4) (1999) 514–529
4. Qin, J., Dong, W.Y., Chen, Y.P., Kang, L.S.: An immune-balance model and its preliminary application in apl problems. In Lishan, K., Zhihua, C., Xuesong, Y., eds.: Progress in Intelligence Computation & Applications. (2005) 586–593
5. Hac, A., Zhou, K.L.: A new heuristic algorithm for finding minimum-cost multicast trees with bounded path delay. International Journal of Network Management **9**(3) (1999) 265–278
6. Waxman, B.M.: Routing of multipoint connections. Selected Areas in Communications, IEEE Journal on **6**(9) (1988) 1617–1622
7. Lin, H.C., Lai, S.C.: Vtdm-a dynamic multicast routing algorithm. Volume 3. (1998) 1426–1432 vol.3
8. Hong, S.P., Lee, H., Park, B.H.: An efficient multicast routing algorithm for delay-sensitive applications with dynamic membership. In: INFOCOM '98. Seventeenth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE. Volume 3. (1998) 1433–1440

Pheromone Based Dynamic Vaccination for Immune Algorithms

Yutao Qi, Fang Liu, and Licheng Jiao

Institute of Intelligent Information Processing, Xidian University, Xi'an, China, 710071
{qi_yutao,f63liu}@163.com
lchjiao@mail.xidian.edu.cn

Abstract. The vaccination which is a most effective protection against the virus is also a helpful part of the artificial immune systems (AIS). A vaccination in AIS means modifying some genes of an antibody in accordance with priori knowledge so as to gain higher affinity with greater probability. Since vaccination is problem specific, what can we do if we have no idea about the problem? To address this, we propose a pheromone based dynamic vaccination for real coded immune algorithms in this paper. The pheromone which is a term of the ant colony system is used as a carrier of knowledge learned from population's evolution and acts as a producer of dynamic vaccines. Experiments on numerical optimization problems indicate that the pheromone based vaccination operator has the ability of acquiring useful information about the objective functions and developing effective vaccines dynamically.

1 Introduction

Artificial immune systems (AIS) which can be defined as computational systems inspired by theoretical immunology and observed immune functions has attracted significant research interest over the years [1]. Inspired by the clonal selection theory, De Castro pioneered the Clonal selection algorithm (CSA) [2] in 2000. After that, many clonal selection based artificial immune algorithms have been proposed. The dynamic clonal selection algorithm (DynamiCS) constructed by Kim in 2002 [3] and polyclonal strategy proposed by Licheng Jiao in 2003 [4] are two of the most outstanding contributions. Lei Wang and Licheng Jiao introduced immune concepts and methods into evolutionary algorithm to form immune evolutionary algorithms [5]. In this work, two immune operators of vaccination and immune selection are designed and they have been proven to be able to restrain the degenerate phenomenon of GA.

Vaccination is a means of stimulating the immune system to produce disease-preventing antibodies. Vaccines can be regarded as a priori knowledge about a certain disease. In the study of artificial immune systems, vaccines have been designed for different problems such as numerical optimization problem [6], TSP problem [7], SAR image segmentation [8] and intrusion detection problem [9] and so on. These vaccines are problem specific. What can we do when we have no idea about the problem? This paper is dealing with an approach for preparing vaccines adaptively when we know nothing about the character of the problems dealt with.

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 1053–1060, 2007.
© Springer-Verlag Berlin Heidelberg 2007

This work proposed a pheromone based dynamic vaccination for immune algorithms to learn valuable information from population's evolution by stimulating the pheromone deposition and volatilization mechanism of ant colonies. In addition, the acquired knowledge is used to produce dynamic vaccines which act as a guidance for further evolutions. Experiments on numerical optimization problems indicate that the pheromone based vaccination operator has the ability of acquiring useful information about the objective functions and developing effective vaccines dynamically.

2 Pheromone Based Vaccination

The pheromone is a term from ant colonies. Inspired by the foraging behavior of ants, M. Dorigo pioneered the ant colony algorithm (ACA) [9] and applied it to the traveling salesman problem (TSP) [10]. Ant colony algorithm is a novel category of bionic algorithm for optimization problems. It is suitable for solving combinational optimization problems, but has the limitation of stagnation and is easy to fall into local optimums when dealing with massive problems. Moreover, it is difficult for ACA to deal with optimization problems with continuous searching space. However, ACA provides illuminations that can be utilized effectively to learn the developing tendency of genes with continuous coding spaces. Inspired by ACA, a pheromone based vaccination operator is designed for artificial immune algorithms to solve optimization problems with continuous searching space. The optimization problems with continuous searching space can be described as follows:

$$\min f(\mathbf{x}), \qquad \mathbf{x} = (x_1, x_2, ..., x_n) \in S \tag{1}$$

In formula (1), $f(\mathbf{x})$ is the target function of the optimization problem, $\mathbf{S} \subseteq R^n$ is the n-dimensional continuous searching space with border $\underline{x_i} \le x_i \le \overline{x_i}, i = 1, 2, ..., n$.

2.1 Real Number Coding and Antibody-Antigen Affinity Definition

The proposed vaccination operator is designed for real number coded immune algorithms. For an optimization problem with n variables, the antibody coding is a real number string with length n, described as, $g_i \in [0,1] \quad i = 1, 2, ..., n$. Note $\mathbf{X} = (x_1, x_2, ..., x_n)$ is the variable of the optimization problem dealt with, $\mathbf{A}$ is the antibody coding of the variable $\mathbf{X}$, described by $\mathbf{A} = e(\mathbf{X})$. $\mathbf{X}$ is called the decoding of antibody A, described by $\mathbf{X} = e^{-1}(\mathbf{A})$. Then we can get:

$$x_i = \underline{x_i} + (\overline{x_i} - \underline{x_i}) \times g_i, \; i = 1, 2, ..., n \tag{2}$$

As antibody's affinity must be a positive value, we construct a negative real function $g(\mathbf{X})$ which is consistent with $f(\mathbf{x})$. In other words, for any two variables $\mathbf{X}_1, \mathbf{X}_2 \in S$, if $g(\mathbf{X}_1) > g(\mathbf{X}_2)$ then $f(\mathbf{X}_1) > f(\mathbf{X}_2)$. In this case, the original

optimization problem becomes $\min\{g(e^{-1}(\mathbf{A})):\mathbf{A}\in\mathbf{I}\}$, in which $\mathbf{I}$ is the antibody space. The antibody-antigen affinity is defined by equation (3):

$$Affinity(\mathbf{A}) = -g(e^{-1}(\mathbf{A})) \tag{3}$$

2.2 Pheromone Matrix

From the point of view of each gene location, its value changes from one to another between 0 and 1 to find an appropriate value that maximize affinity of the antibody. Since there are many antibodies in the population, it looks as if there are many ants walk between 0 and 1 to look for food. Therefore, the foraging behavior of ants can be mimicked. The values of genes are continuous, but pheromone concentrations on the pheromone trail can only be recorded discretely. To overcome this issue, we quantize the coding domain of each gene [0,1] into finite number of segments evenly.

For each gene location, an *m*-dimensional real vector is provided to take records of pheromone concentrations, where *m* is the number of discrete domains. The vector can be described as $\mathbf{Ph}_i = (ph_1, ph_2, ..., ph_m)$ $ph_{\min} \le ph_j \le ph_{\max}$ $1 \le j \le m$, in which $ph_{\min}$ and $ph_{\max}$ are the upper and lower limit of pheromone concentration respectively. As the antibody has *n* gene locations, these *n* pheromone vectors make a pheromone matrix with size $n \times m$, noted as $\mathbf{Ph} = (\mathbf{Ph}_1, \mathbf{Ph}_2, \cdots, \mathbf{Ph}_n)^T$. It is fulfilled with an initial value of *w*.

2.3 Vaccination Operators Based on Pheromone

The pheromone based vaccination proposed in this paper has two parts. One of them is the vaccine injecting operator. In this operator, vaccines will be produced and injected using information from the pheromone matrix. The other one is the pheromone updating operator. It learns useful information from the foregoing evolution by manipulation the pheromone concentration of the pheromone matrix.

2.3.1 Vaccine Injecting Operator

The vaccine injection operator will be applied on antibodies of the population after the immune genetic operations as probability p_i. The vaccine injection operator can be described by the following pseudo code:

For each antibody of the population $\mathbf{A}_k = (g^k{}_1, g^k{}_2, \cdots g^k{}_n)$

If { p_i } = TRUE

$i = \text{Random}(n)$ $j = \text{Location}(g^k{}_i)$;

$L = \begin{cases} j-r & \text{if } j-r>0 \\ m+j-r & \text{else} \end{cases}$, $R = \begin{cases} j+r & \text{if } j+r \le m \\ j+r-m & \text{else} \end{cases}$, $r = \lfloor m/8 \rfloor$;

$t = \text{Roulette}(L, R)$;

$$\left(g^k{}_i\right)' = \begin{cases} \text{Gauss}\left(g^k{}_i, \sigma\right) \left| \left[min_j, max_j\right] \right. & if \;\; t = j \\ \text{Latest}\left(t\right) & else \end{cases} \qquad \sigma = 10^{-\frac{1}{3}\lg m} ;$$

$g^k{}_i = \left(g^k{}_i\right)'$;

End If

End For

In which, *pop_size* is the size of the population. The function Random(n) returns a random integer between 1 and n. The function Location($g^k{}_i$) returns the serial number of segment in which $g^k{}_i$ is located. Roulette(L, R) is to select a segment between L and R by mean of the roulette selection according to pheromone concentrations. The mutation radius $r = \lfloor m/8 \rfloor$ is the largest integer smaller than $m/8$. $g^k{}_i$ and $\left(g^k{}_i\right)'$ are respectively the gene value before and after the vaccine injection operation. $\text{Gauss}\left(g^k{}_i, \sigma\right) \left| \left[min_j, max_j\right] \right.$ means gauss mutation with a standard variation of σ, and it returns a value within the j-th segment. In other words, the returned value is between min_j and max_j, where min_j and max_j are respectively the lower limit and the upper limit of the j-th segment. Latest$\left(t\right)$ is the latest gene value that is located in the t-th segment. The midpoint value between min_t and max_t will be returned if no gene has previously been assigned to values located in this segment.

2.3.2 The Pheromone Updating Operator

For each antibody in the population, apply this operator on it after the vaccine injecting operation. Let $\mathbf{A}$ be any antibody in the population and $\mathbf{A}'$ be its offspring after the immune genetic operation and the vaccine injecting operation. The pheromone updating operator can be described by the following pseudo code:

If $\mathbf{A}' = g_1', g_2', ..., g_n'$ is superior to $\mathbf{A} = g_1, g_2, ..., g_n$

For each gene of antibody A

If $g_i \neq g_i'$, $i = 1, 2, ..., n$

$Ls = \text{Location}(g_i)$, $Le = \text{Location}(g_i')$;

Update pheromone of segments between Ls and Le :

$$ph' = \left[(1-\rho)\, ph + \Delta ph\right] \left| \left[min_{ph}, max_{ph}\right] \right., \qquad \Delta ph = \frac{Q}{abs\left(Le - Ls\right)} ;$$

Evaporate pheromone of segments beside Ls and Le :

$$ph' = (1-\rho)\, ph \left| \left[min_{ph}, max_{ph}\right] \right. ;$$

End If

```
    End For
  End If
```

In which, $\rho \in (0,1]$ is the pheromone trail decay coefficient. Q is a real constant and it represents the total amount of pheromone deposed by an ant at a time. The function Location(g) returns the serial number of segment which real number g is located in. The function abs(a) means the abstract of number a. ph and ph' are respectively the pheromone concentrations before and after updating. The value of ph and ph' must be no smaller than the lower limit min_{ph} and no larger than the upper limit max_{ph}.

3 Simulation Experiments

In order to validate the effectiveness of the proposed vaccination approach, the immune algorithms with pheromone based vaccination are executed to solve the following test functions.

F01-Min $f(\boldsymbol{x}) = \sum_{i=1}^{n} x_i^2$; $\mathcal{S} = [-100, 100]^n$; $f_{\min} = 0$;

F02-Min $f(\boldsymbol{x}) = \sum_{i=1}^{n} |x_i| + \prod_{i=1}^{n} |x_i|$; $\mathcal{S} = [-10, 10]^n$; $f_{\min} = 0$;

F03-Min $f(\boldsymbol{x}) = \max_i \{ | x_i |, 1 \le i \le n \}$; $\mathcal{S} = [-100, 100]^n$; $f_{\min} = 0$;

F04-Min $f(\boldsymbol{x}) = \sum_{i=1}^{n} \left(\lfloor x_i + 0.5 \rfloor \right)^2$; $\mathcal{S} = [-100, 100]^n$; $f_{\min} = 0$;

F05-Min $f(\boldsymbol{x}) = \sum_{i=1}^{n} \left(-x_i \sin\left(\sqrt{| x_i |} \right) \right)$; $\mathcal{S} = [-500, 500]^n$; $f_{\min} = -418.983 \times n$;

F06-Min $f(\boldsymbol{x}) = \sum_{i=1}^{n} (x_i^2 - 10\cos(2\pi x_i) + 10)$; $\mathcal{S} = [-5.12, 5.12]^n$; $f_{\min} = 0$;

F07-Min $f(\boldsymbol{x}) = -20\exp\left(-0.2\sqrt{\frac{1}{N}\sum_{i}^{n} x_i^2} \right) - \exp\left(\frac{1}{n}\sum_{i=1}^{n} \cos(2\pi x_i) \right) + 20 + e$; $\mathcal{S} = [-30, 30]^n$; $f_{\min} = 0$;

F08-Min $f(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} \left(x_i^4 - 16x_i^2 + 5x_i \right)$; $\mathcal{S} = [-5, 5]^n$; $f_{\min}$ -78.33236

3.1 Effectiveness of Pheromone Based Dynamic Vaccination

Experiments have been done to investigate the effectiveness of pheromone based dynamic vaccination. Table 1 is comparisons of the performance between the clonal selection algorithm (CSA) [2] and the CSA with pheromone based dynamic vaccination (PHDV_CSA). Table 2 is comparisons between the immune programming (IP) [5] and the IP with pheromone based dynamic vaccination (PHDV_IP). All results presented are averages of fifty independent runs. Each run continues until either an optimal solution has been found or the maximum number of generation 1000 is reached. In each run, the control parameters were set as follows: the population size *pop_size* is 20, the test function dimension *n* is 10, the number of segments *m* is 1000, the pheromone trail decay coefficient ρ is 0.2, the constant of total pheromone deposited in one trip Q is 800, the initial pheromone concentration *w* is 1, the lower and

upper limit of pheromone concentration are respectively 0.5 and 15 and the vaccine injection probability p_i is 0.6. Other parameters for CSA and IP are the same as those used in reference [2] and [5].

Table 1. Comparison between CSA and PHDV_CSA

Test functions	Average function evaluations		The optimal mean (Standard)	
	CSA	PHDV_CSA	CSA	PHDV_CSA
F01	26077	19074	8.13×10^{-4} (4.13×10^{-4})	3.21×10^{-4} (2.36×10^{-4})
F02	26335	22800	2.07×10^{-4} (1.67×10^{-4})	3.19×10^{-5} (4.19×10^{-6})
F03	206936	166911	7.31×10^{-4} (5.66×10^{-4})	3.98×10^{-4} (2.14×10^{-4})
F04	17931	5414	0 (0)	0 (0)
F05	47839	17450	-4189.76 (6.74×10^{-3})	-4189.83 (3.15×10^{-3})
F06	28619	19680	5.76×10^{-4} (3.36×10^{-4})	1.52×10^{-4} (1.17×10^{-4})
F07	30393	25306	7.31×10^{-4} (1.75×10^{-4})	5.18×10^{-5} (3.32×10^{-6})
F08	22786	7279	-78.3304 (3.29×10^{-5})	-78.3323 (1.77×10^{-5})

Table 2. Comparison between IP and PHVV_IP

Test functions	Average function evaluations		The optimal mean (Standard)	
	IP	PHDV_IP	IP	PHDV_IP
F01	29629	18524	2.84×10^{-4} (3.31×10^{-4})	1.95×10^{-4} (1.63×10^{-4})
F02	30852	20459	1.76×10^{-4} (3.35×10^{-4})	2.82×10^{-5} (3.37×10^{-6})
F03	397596	173426	7.06×10^{-4} (2.17×10^{-4})	5.13×10^{-4} (3.39×10^{-4})
F04	22164	7096	0 (0)	0 (0)
F05	56580	18126	-4189.74 (4.11×10^{-3})	-4189.82 (2.76×10^{-3})
F06	32460	18288	3.67×10^{-5} (6.17×10^{-5})	1.27×10^{-4} (5.42×10^{-4})
F07	36799	24421	4.92×10^{-4} (1.97×10^{-4})	4.55×10^{-5} (7.94×10^{-6})
F08	26126	9329	-78.3306(6.71×10^{-4})	-78.3322(2.34×10^{-5})

It can be seen from these data that PHDV_CSA and PHDV_IP perform much better than CSA and IP, as they have found superior solutions with less function evaluations. These data enable us to conclude that the pheromone based dynamic vaccination is effective on improving the searching capability of immune algorithms. Moreover, the comparisons of the standard deviation indicate that the proposed vaccination strategy makes the immune algorithms much more robust.

3.2 Convergence of Pheromone

This experiment is designed to investigate whether the pheromone based vaccination approach has learned the real information about the test functions. Considering the first vector of the pheromone matrix, Fig.1 and Fig.2 show the convergence process of the pheromone distribution while solving the optimization problem of the 30 dimensional test functions F1 and F5 with PHDV_CSA.

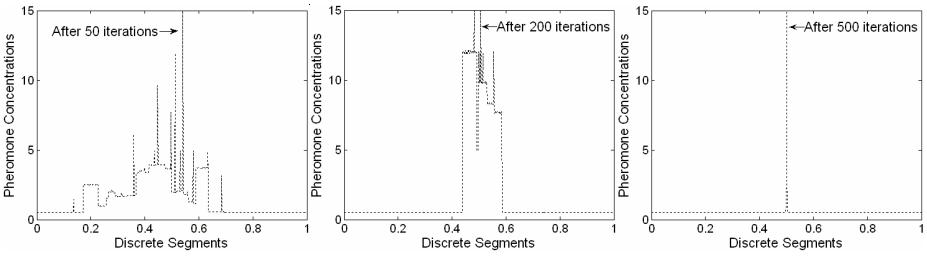


Fig. 1. The change of the pheromone distribution for F1

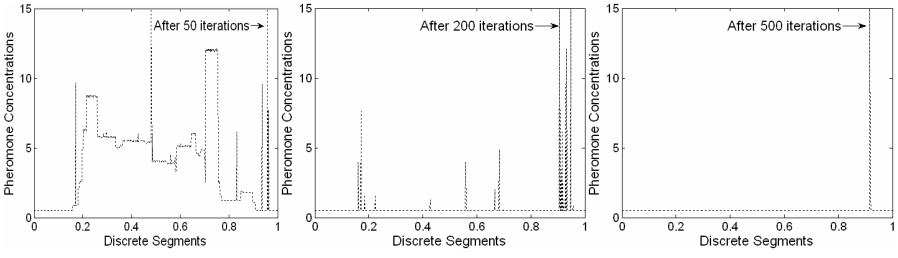


Fig. 2. The change of the pheromone distribution for F5

Take F1 for example, its best value is $x^* = (0,0,\cdots,0)$. The best value in the first dimension is 0 which is mapped to a gene value of 0.5. It can be seen from Fig.3 that, the pheromone concentrations converge to the segment around the very value of 0.5. Another example is F5. The best value of the function F5 is $x^* = (402.9687, 402.9687, \cdots, 402.9687)$. Its best value in first dimension 402.9687 is mapped to a gene value of 0.9029687 in the coding space. Fig.4 indicates that the pheromone concentrations converge to the segment around the very value after 500 iterations. Therefore, we come to the conclusion that the pheromone based vaccination approach has achieved the real information about the objective functions.

4 Concluding Remarks

In this paper, we presented a pheromone based vaccination approach for immune algorithms. Our objective was to design a dynamic vaccination strategy which is independent from prior knowledge about the problem for immune algorithms, so that they could be more self-adaptive and easy to use.

Inspired by the ant colony system, the proposed approach learned useful information from population's evolution by stimulating the pheromone deposition and volatilization mechanism of ant colonies. Experimental results indicate that, the distribution of pheromone concentrations reflects the real information about the objective function, and the pheromone based dynamic vaccination is efficient in improving the performance of immune algorithms.

However, the proposed vaccination is suitable for immune algorithms without crossover operator. How to make a modification so as to cooperate with the crossover operation is one of the future works.

Acknowledgments. This work has been partially supported by the National Natural Science Foundation of China under Grant No. 60372045 and 60133010, and the National Grand Fundamental Research 973 Program of China under Grant No. 2001CB309403.

References

1. L. N. de Castro, J. Timmis. Artificial Immune Systems: A New Computational Intelligence Approach. Springer-Verlag, Heidelberg, Germany, August 2002.
2. L. N. De Castro, F. J. Von Zuben. The Clonal Selection Algorithm with Engineering Applications, Proceedings of Workshop on Artificial Immune Systems and Their Applications, 2000:36-37.
3. Kim, J. and Bentley, P. J. Towards an Artificial Immune System for Network Intrusion Detection: An Investigation of Dynamic Clonal Selection, Proceedings of Congress on Evolutionary Computation, 2002:1015-1020.
4. Ruochen Liu, Haifeng Du, Licheng Jiao, Immunity Clonal Strategies. Proceedings of the Fifth International Conference on Computational Intelligence and Multimedia Applications. 2003:290.
5. Lei Wang, Jin Pan, Licheng Jiao. The Immune Programming. Chinese Journal of Computers. 2000, 23(8): 806-812.
6. Yanjun Li, Tiejun Wu. A Novel Immune Algorithm for Complex Optimization Problems. Proceedings of the International Conference on Control and Automation. Hangzhou China. June, 2004: 2309-2312.
7. Licheng Jiao, Lei Wang. A Novel Genetic Algorithm Based on Immunity. IEEE Transactions on Systems, Man and Cybernetics. Part A, 2000, 30(9): 1-10.
8. Hua Bo, Fulong Ma, Baojun Han, Licheng Jiao. SAR Image Segmentation based on Immune Algorithm. Proceedings of the International Conference on Control and Automation. Budapest, Hungary. June, 2005: 1279-1282.
9. A. Colorni, M. Dorigo, V. Maniezzo. Distributed optimization by ant colonies. Proceedings of the 1st European Conference on Artificial Life. Pans, France. 1991:134-142.
10. M. Dorigo, L. M. Gambardella. Ant Colony System: A cooperative Learning Approach to the Traveling Salesman Problem. IEEE Transactions on Evolutionary Computation. 1997. 1(1):53-66.

Towards a Less Destructive Crossover Operator Using Immunity Theory

Yingzhou Bi [1,2], Lixin Ding[1], and Weiqin Ying[1]

[1] State key laboratory of software engineering, Wuhan University, Wuhan 430072, China
[2] Department of Information Technology, Guangxi Teachers Education University, Nanning 530001, China
byzhou@163.com

Abstract. When searching for good scheme, a good solution can be destroyed by an inappropriate choice of crossover points. Furthermore, because of the randomicity of crossover, mutation and selection, a better solution can hardly reach in last stage in EA, and the solution always traps in local optimal. Faced to "exploding" solution space, it is tough to find high quality solution just by increasing the population size, diversity of searching, and the number of iteration. In this paper, we design the immunity operator to improve the crossover result by utilizing the immunity theory. As the "guided mutation operator", the immunity operator substituted the "blind mutation operator" in normal EA, to restrain the degenerate phenomenon during the evolutionary process. We examine the algorithm with examples of TSP and gain promising result.

Keywords: Crossover operator, Immunity operator, Traveling salesman problem.

1 Introduction

The role of crossover in evolutionary algorithms is to create new individuals from old ones [1]. During the process of stochastic search, crossover is considered to be the major driving force behind EA. However, when searching for good scheme, standard crossover is widely accepted as being a largely destructive operator[2]; a simple one point crossover operator generates an individual by selecting a crossover point randomly from a parent , then splitting both parents at this point and creating two children by exchanging the tails. The randomness of this operator makes it mostly destructive (generates children inferior to their parents).

Classic EA uses little problem specific knowledge, so its' efficiency is usually low. In order to improve EA's efficiency, we try to have other methods or data structures incorporated into it. This category of algorithm is very successful in practice and forms a rapid growing research area with great potential [3-4]. In order to improve the algorithmic efficiency, the immune genetic algorithm (IGA) is presented in [5], which utilizes some characteristics and knowledge in the pending problems for restraining the degenerative phenomena during evolution.

In this paper, we firstly analyze the destructive and constructive aspects of crossover and mutation in section 2, and then design a novel immunity genetic algorithm in section 3: the "blind mutation operator" is substituted by the "guided mutation operator".

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 1061–1067, 2007.
© Springer-Verlag Berlin Heidelberg 2007

Finally, we compare the novel algorithm and canonical genetic algorithm with the example of traveling salesman problem.

2 Disruption and Construction of Crossover

The disruption of crossover refers to the offspring of an individual in a schema H_k no longer belong to schema H_k after crossover. It can be depicted as fig.1: P_1 is a member of a third-order schema H_3, and P_2 is an arbitrary string. If the two parents have not matching alleles at position d_2 or they have not matching alleles at poison d_1 and d_3, the schema H_3 should be disrupted.

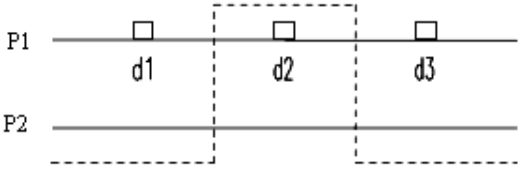


Fig. 1. An example of two-point crossover. P_1 is a member of a third-order schema H_3, and P_2 is an arbitrary string.

The construction of crossover refers to having crossover create an instance of a schema H_k from both parents, where one parent is in schema H_m ,while the other parent is the schema H_n, and $k=m+n$. Figure 2 provides a pictorial example. P_1 is a member of a second-order schema, and P_2 is a member of another second-order schema. A higher schema H_4 is constructed by crossover.

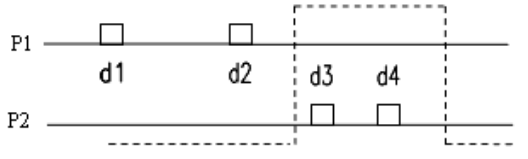


Fig. 2. An example of construction of two-point crossover. P_1 is a member of a two-order schema H_2, and P_2 is a member of another two-order schema H_2.

Based on the systemic analyze of crossover and mutation, literature [6] presents some general results: (1) all forms of crossover are more disruptive of high-order schemas and become less disruptive when the population converges.(2) more disruptive crossover are more likely to construct(and vice versa). This has proven a No-Free-Lunch Theorem with respect to the disruption and construction aspects of crossover. (3) Mutation can achieve any level of disruption that crossover can achieve, however, from a construction point of view; crossover is more powerful than mutation.

3 A Novel Algorithm with Immunity Operator as the "Guided Mutation Operator"

The most remarkable role of immune system is the protection of the organism against the attack of disease, and the elimination of malfunctioning cells. Inspired by the theory of immunity, one main idea of immunity computation is to abstract some useful problem-specific knowledge (vaccine) from the pending problem (antigen), and utilize this knowledge to search the solution (antibody). In TSP, a solution with high fitness should be composed of most short links between nodes, it means the usher and subsequence of a node in a roundtrip is always its near neighbor, and here, we called it "principle of near neighbor". If the principle is disobeyed, the roundtrip will be composed of lots of long links, and the total length must be very long. So we should find the near neighbor of all nodes at the beginning of algorithm and the near neighbor of nodes are selected by the distance between them. Usually, we only consider five nodes as near neighbor. The principle of near neighbor and the near neighbor of every node are the vaccine for TSP.

3.1 Vaccinations

Vaccination means modifying the genes on some bits of the candidate solution with priori knowledge.

1) Analyze all links in the candidate roundtrip: record each pair of nodes of the long links (if a solution composing n links, we usually consider $n/2 \sim n/3$ longer links) and the distance between them with a structure array. Here we called the structure array *update-array*; every element in update-array A has three parts: two nodes and a distance, where one node is called resource-nodes and another node is called destiny-nodes. The update-array A is sorted in descend-order according the distance in its element.

2) Examine all the resource-node in update-array A whether the destiny-node is its near neighbor. For example, for any element in update-array A, suppose its resource-node is V_a ,and its destiny-node is V_b ,if V_b is not near neighbor of V_a ,it means something is wrong in this gene, repairing this gene may increase the fitness of the candidate solution. Inspired by *2-change* and *3-change* algorithms, which have a time complexity of $O(n^2)$ and $O(n^3)$ respectively [7], we present *2-repair* and *3-repair* operators to repair the malfunction genes in candidate solution, however their time complexity are $O(n)$,for every node we only consider its near neighbor and the number of near neighbor is usually 5~8.

3) *2*-repair operator: for a long link (V_a, V_b),if the destiny-node V_b is not 2-nearest neighbor of resource-node V_a, then select a node V_c from the near neighbor of V_a , obviously we have $d(V_a, V_b) > d(V_a, V_c)$. Suppose node V_d is subsequence of node V_c and the solution is expressed as $\pi = \{V_1, V_2, \ldots V_a, V_b \ldots V_c, V_d \ldots V_n\}$,its total length is

$$D(\pi) = \sum_{i=1}^{r} d_i + d(V_a, V_b) + \sum_{i=r+2}^{t} d_i + d(V_c, V_d) + \sum_{i=t+2}^{n} d_i \quad . \tag{1}$$

if $d(V_a,V_b)+ d(V_c,V_d)> d(V_a,V_c)+ d(V_b,V_d)$,then we reverse all the nodes between V_b and V_c(including V_b,V_c), then get a new solution $\pi'= \{V_1,V_2,\ldots V_a, V_c\ldots V_b, V_d\ldots V_n\}$, its total length is

$$D(\pi') = \sum_{i=1}^{r} d_i + d(V_a,V_c) + \sum_{i=r+2}^{t} d_i' + d(V_b,V_d) + \sum_{i=t+2}^{n} d_i \ . \tag{2}$$

for the symmetry TSP, we always have $d_i=d_i'$, so under the condition $d(V_a,V_b)+ d(V_c,V_d)> d(V_a,V_c)+ d(V_b,V_d)$,it is obvious that D(π)>D(π'),which means the new solution is better.

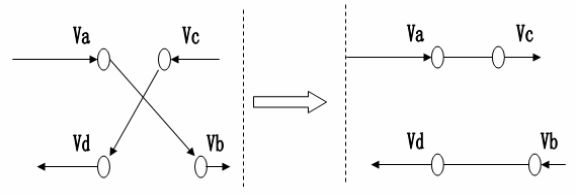


Fig. 3. An example of vaccination of *2*-repair for TSP

4) *3*-repair operator: it is similar to *2*-repair, but it exchange 3 links in every repairing. The mechanism is pictorial in figure 4: the original solution is $\pi=\{V_1,V_2,\ldots V_a,V_b\ldots V_c,V_d,V_e\ldots V_n\}$,

if $d(V_a,V_b)+d(V_c,V_d)+d(V_d,V_e)>d(V_a,V_d)+d(V_d,V_b)+d(V_c,V_e)$ then exchange 3 links: the 3 bold links are substituted by 3 dash links, and the new solution become $\pi'=\{V_1,V_2,\ldots V_a, V_d ,V_b\ldots V_c, V_e\ldots V_n\}$.It is obvious that D(π)>D(π').

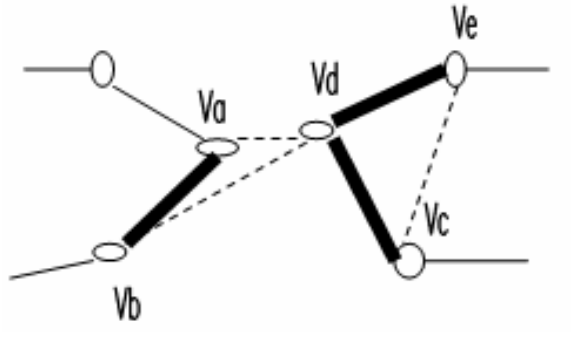


Fig. 4. An example of vaccination of *3*-repair for TSP

3.2 A Novel Algorithm for TSP

According to the analysis in section 2, from a construction point of view, crossover is more powerful than mutation. We present a novel algorithm for TSP: the "blind" mutation operator in canonical genetic algorithm is substituted by the "guided" mutation

operator, which is immunity operator. The aim of leading immune concepts and methods into GA is to utilize the problem-specific knowledge for refraining the destruction of crossover while make useful of its construction during evolutionary process.

Algorithm 1

1) Initialize the population with greedy algorithm.
2) Abstract the vaccine.
3) Evaluate the individuals, if the solution quality is satisfied then stop, or if reach the maximal iteration time but the solution quality is not satisfied , then repair the best individual with 3-repair operator(in section 3.1) and stop; otherwise, continue the following.
4) Crossover: select two parents and perform crossover.
5) Vaccination: repair the resulting offspring with 2-repair operator (in section 3.1) with probability *P*i, and insert the new candidates into next generation.
6) Go to step 3.

4 Experimental Results

We examine algorithm 1 and canonical genetic algorithm(CGA in brief, it's crossover operator is the Order Crossover designed by Davis, and mutation operator is the Inversion Mutation [1].) with 4 TSP instances: two of them are *TSP200* and *TSP500*, where the nodes in *TSP200* and *TSP500* are 200 and 500 respectively; The location of all the nodes are generated randomly in two-dimension space: $x \in [0,1000], y \in [0,1000]$, the distance between nodes is Euclidean distance. Another two instances are *fl417*, *att532*, which come from TSPLIB [8]. The experiments are carried on a PC with PIV 2.93 GHz CPU, 1G RAM and Windows XP operation system.

In order to compare the performance between algorithm 1 and CGA, we design the parameter as following: in CGA, the size of population is 100;the crossover and the mutation probabilities are 0.5 and 0.4,respectively; the number of generations is 20000; in algorithm 1, the size of population is 50; the crossover and the immunity probabilities are 0.8 and 0.85,respectively;the number of generations is 100.The experimental results of CGA and algorithm 1 over 20 independent runs respectively are listed in Table 1, where n is the number of nodes, D_0 is the known minimal tour length: the results are gained by Concorde package [9] (in *att532* the minimal tour length by Concorde package is 86729 while literature [10] take it as 87550)and *time* denotes the average run time, while δ denotes the relative difference[10] which is defined in Equation (3)

$$\delta\% = \left(\sum_{i=1}^{T} (D_i - D_0) \right) / (T \times D_0) \times 100\% \tag{3}$$

Where D_i denotes the minimal tour length obtained from the i-th run, T is the number of independent runs.

Table 1. Compare the performance between CGA and algorithm 1

TSP Instances	n	D_0	δ		*Time*(second)	
			CGA	Algorithm1	CGA	Algorithm 1
TSP200	200	10610	7.9	**0.3**	136	**10**
TSP500	500	16310	4.4	**1.1**	425	**99**
Fl417	417	11861	2.9	**1.2**	315	**39**
Att532	532	86729	6.4	**1.7**	473	**108**

For the analysis of the role of crossover and immunity operator in immunity-evolutionary algorithm, we perform the algorithm 1 with different crossover probabilities P_c while the immunity probability P_i is always 0.85, given population size =50, iterative times= 100. The experimental results of algorithm 1 over 20 independent runs respectively are shown in table 2, where δ denotes the relative difference [10] which is defined in Equation (3).

Table 2. Compare the role of crossover and immunity operator in algorithm 1

TSP Instances	δ			
	P_c=0.00	P_c=0.50	P_c=0.80	P_c=0.99
TSP200	7.1	1.0	0.3	0.3
TSP500	3.1	1.7	1.1	1.1
Fl417	6.1	2.1	1.2	1.1
Att532	3.9	2.4	1.7	1.5

Based on the results in table 1 to table 2, we can make some general observations:

Firstly, evolutionary algorithm can reach a relatively good solution from a bad solution after a stochastic trial-and-error. Because of the destruction of crossover and mutation, a better solution quality can hardly reach in last stage in EA, and the solution always trap in local optimal. Faced to "exploding" solution space, it is tough to find high quality solution just by increasing the population size (in CGA the population size is 100, while it is only 50 in algorithm 1), diversity of searching, and the number of iteration (the maximum number of generations in CGA is 20000, while it is only 100 in algorithm 1). However we can restrain the destruction of crossover by repairing the malfunction genes in candidate solution, and utilize its construction in our novel algorithm with immunity operator as "guided" mutation.

Given different crossover probabilities, algorithm 1 has different performance: the bigger is crossover probability, the better is performance, however, when the crossover probability reaches 0.8, the performance will increase little, while it need more time. If the crossover probability is 0, it means the crossover is turned off, the performance is the worst. The experimental results validate the construction of crossover. Furthermore, when compared with the pure artificial immunity algorithm [10] in TSP instance *att532,* the result of algorithm 1 is better: if consider the minimal tour length is 87550, the relative difference is 0.7%, while the relative difference is 2.21% in reference [10].

5 Conclusions and Future Work

By utilizing the problem-specific knowledge, immunity operator can restrict the degenerative phenomena arising from the evolutionary process, and guaranteeing that the search will concentrate in the regions containing high quality solutions. Artificial immunity theory can contribute with the already established methodologies in order to mutually improve their performances and application domains. It shows that the combination of an evolutionary and a heuristic method--a hybrid evolutionary algorithm--performs better than either of its "parent" algorithms alone.

For future work, we will use the proposed method on other combinational problems or even real optimization problems.

Acknowledgments. This work is supported in part by the National Natural Science Foundation of China (Grant no. 60204001), and Natural Science Foundation of Guangxi (Grant no. 0679018).

References

1. Eiben, A.E, Smith, J.E.: Introduction to Evolutionary Computing. Berlin Heidelberg New York Press, Springer-Verlag (2003)
2. Majeed, H., Ryan,C.: A Less Destructive, Context-Aware Crossover Operator for GP.Lecture Notes in Computer Science, Vol.3905.Berlin /Heidelberg, Springer-Verlag (2006) 36–48
3. Yao, X., Xu, Y.: Recent Advance in Evolutionary Computation. Journ of Comput Sci & Technol, **21**(1) (2006) 1–18
4. Blum C., Roli A.: Metaheuristics in Combinatorial Optimization: Overview and Conceptual Comparison. ACM Computing Surveys, **35**(3)(2003) 268–308
5. Jiao, L., Wang,L.: A novel genetic algorithm based on immunity. IEEE Transactions on System s, M an, and Cybernetics, **30**(5) (2000) 552–561
6. Spears, W.M.: The role of Mutation and Recombination in Evolutionary Algorithms. Virginnia: George Mason University (1998)
7. Helsgaun, K.: An Effective Implementation of the Lin-Kernighan Traveling Salesman Heuristic. http://www.akira.ruc.dk/~keld/
8. http://www.iwr.uni-heidelberg.de/groups/comopt/software/TSPLIB95/
9. Concorde TSP solver for windows. http://www.tsp.gatech.edu/concorde/index.html
10. Gong, M., Jiao, L., Zhang, L.: Solving Tranveling Salesman Problem by Artificial Immunity Responese. T.-D Wang et al.(Eds.):SEAL 2006. Lecture Notes in Computer Science, Vol. 4247. Springer-Verlag, Berlin Heidelberg (2006) 64–71

Studying the Performance of Quantum Evolutionary Algorithm Based on Immune Theory

Xiaoming You[1,2], Sheng Liu[1,2], and Dianxun Shuai[2]

[1] College of Electronic and Electrical Engineering, Shanghai University of Engineering Science ,200065 Shanghai , China
yxm6301@163.com
[2] Dept of Computer Science and Technology, East China University of Science and Technology, 200237 Shanghai, China
ls6601@sina.com

Abstract. A novel quantum evolutionary algorithm based on immune operator (MQEA) is proposed. The algorithm can find out optimal solution by the mechanism in which antibody can be clone selected, immune cell can accomplish cross-mutation and Self-adaptive mutation, memory cells can be produced and similar antibodies can be suppressed. It not only can maintain quite nicely the population diversity than the classical evolutionary algorithm, but also can help to accelerate the convergence speed. The technique for improving the performance of MQEA has been described and its superiority is shown by some simulation experiments in this paper.

Keywords: Quantum evolutionary algorithm, Immune theory, Self-adaptive mutation, Cross-mutation, Performance.

1 Introduction

Evolutionary algorithms have received a lot of attention regarding their potential as global optimization techniques for complex optimization problems. Research on merging evolutionary algorithms with quantum computing[1] has been developed since the end of the 90's, this research can be divided in two different groups: one that focus on developing new algorithms[2]; and another which focus on developing quantum-inspired evolutionary algorithms with binary and real representations [3] which can be executed on classical computers.

Han proposed the quantum-inspired evolutionary algorithm (QEA)[3], he applied the QEA to some optimization problems and the performance of the QEA is better than classical evolutionary algorithms in many fields [4]. Although quantum evolutionary algorithms are considered powerful in terms of global optimization, they still have several drawbacks regarding local search (i) lack of local search ability, and (ii) premature convergence.

In recent years, the study on the novel algorithms based on biological immune mechanisms has become an active research field. A number of researchers

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 1068–1075, 2007.
© Springer-Verlag Berlin Heidelberg 2007

have experimented with biological immunity-based optimization approaches to overcome these particular drawbacks implicit in evolutionary algorithms.

In this paper, a quantum evolutionary algorithm based on the adaptive immune operator (MQEA) is proposed. MQEA can find out optimal solution by the mechanism in which antibody can be clone selected, immune cell can accomplish cross-mutation and Self-adaptive mutation, memory cells can be produced and similar antibodies can be suppressed. We describe technique for improving the performance of MQEA and its superiority is shown by some simulation experiments. In order to evaluate MQEA, a set of standard test functions were used and its performance is compared with that of QEA. Specifically, Section 2 then proposes a novel quantum evolutionary algorithm based on the adaptive immune operator. The following section then analyzes and presents performance results. In Section 4 the performance of MQEA is evaluated by some well-known test functions.

2 Modified Quantum Evolutionary Algorithm Based on Immune Operator

Conventional Quantum Evolutionary Algorithm (CQEA) [3], [4] is efficacious, in which the probability amplitude of Q-bit was used for the first time to encode the chromosome and the quantum rotation gate was used to implement the evolving of population. Quantum evolutionary algorithm has the advantage of using a small population size and the relatively smaller iterations number to have acceptable solution, but they still have several drawbacks such as premature convergence.

2.1 Immune Systems Mechanism Analysis

Immune Algorithms (IAs) are evolutionary algorithms [5],[6], [8] based on physiological immune systems. Physiological immune systems have mechanisms [8] that enable cells to exhibit and recognize foreign substances. The mechanisms work by first recognizing foreign substances known as antigens, the immune systems then generate a set of antibodies to interact with the antigens, these antibodies interact with the antigens to produce diverse results. The mechanisms are able to recognize which antibodies are better at interacting with the antigens and produce those antibodies as memory cells in the next generation of antibodies. Those mechanisms are able to distinguish which antibodies are overly dominant. They suppress the growth of these dominant antibodies so as to diversify the types of antibodies that are tested against the antigens in the exploring and exploiting different results.

Physiological immune systems have affinity maturation mechanism and the immune selection mechanism [7]. Affinity maturation conduces to immune system self-regulates the production of antibodies and diverse antibodies, these higher-affinity matured cells are then selected to enter the pool of memory cells. The merits of IA are as follows:

– IA operates on the memory cell, which guarantees fast convergence toward the global optimum.
– IA has an affinity maturation routine, which guarantees the diversity of the immune system.
– The immune response can enhance or suppress the production of antibodies.

2.2 Quantum Evolutionary Algorithm Based on Immune Operator

The flowchart of quantum evolutionary algorithm based on immune operator (MQEA):

```
MQEA()
{ t=0;
  Initialize Q(0) ;
  Make P(0) by observing the states of Q(0) ;
  Evaluate P(0) ;
  Store the optimal solutions among P(0) ;
  While (not termination - condition) do
     { t=t+1;
       Update Q(t-1) using Q-gates U(t-1);
       Make P(t) by observing the states of Q (t);
       Evaluate P(t);
       Store the optimal solutions among P(t);
       Implement the immune operator for Q (t), P (t):
      {
         Clonal proliferation;
         Self-adaptively cross-mutate each cell;
         Suppress similar antibody;
         Select antibody with higher affinity as memory cells;
      }
    }
 }
```

Quantum gate (Q-gate) U(t) is a variable operator of QEA, it can be chosen according to the problem. A common rotation gate used in QEA is as follows [4]:

$$U(\theta) = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & cos(\theta) \end{bmatrix}$$

where θ represent the rotation angle, $Q(t) = (q_t^1, q_t^2 ... q_t^n)$

In the step "make $P(0)$ by observing the states of $Q(0)$", generates binary solutions in $P(0)$ by observing the states of $Q(0)$, where $P(0) = \{x_0^1, x_0^2, ..., x_0^n\}$ at generation $t = 0$. One binary solution, x_0^j, is a binary string of m, which is formed by selecting either 0 or 1 for each bit by using the probability, either $|\alpha_{0i}^j|^2$ or $|\beta_{0i}^j|^2$ of q_0^j, respectively. In the while loop, q_t^jindividuals in $Q(t)$ are updated by applying Q-gates $U(t)$ defined as a variation operator of QEA, binary

solutions in $P(t)$ are formed by observing the states of $Q(t)$as the aforementioned method and each binary solution is evaluated for the fitness value. It should be noted that x_t^j in $P(t)$ can be formed by multiple observations of q_t^j in $Q(t)$.

2.3 Immune Operator

The clonal selection and affinity maturation principles are used to explain how the immune system improves its capability of recognizing and eliminating pathogens. Clonal selection states that antigen can selectively react to the antibodies, if the antibody matches the antigen sufficiently well, its B cell becomes stimulated and can produce related clones. The cells with higher affinity to the invading pathogen differentiate into memory cells, This whole process of mutation plus selection is known as affinity maturation. Inspired by the above clonal selection and affinity maturation principles, the cross-mutation operator could be viewed as a self-adaptive mutation operator. Self-adaptive mutation plays the key role in MQEA, generally, cells with low affinity are mutated at a higher rate, whereas cells with high affinity will have a lower mutation rate. This mechanism offers the ability to escape from local optima on an affinity landscape.

Cross-mutation operator can act as follow: give a randomly position j of the chromosome

$$q = \begin{pmatrix} \alpha_1|\alpha_2|...\alpha_j|...\alpha_m \\ \beta_1|\beta_2|...\beta_j|...\beta_m \end{pmatrix}$$

if $|\beta_j|^2 < p(p$ is mutation rate), let(α_j, β_j) transfer to (β_j, α_j), we can have low affinity cells with higher mutation rate ph, whereas high affinity cells with lower mutation rate $pl, pl < ph$. The immune system self-regulates the production of antibodies and diverse antibodies. In order to maintain diversity, the similar antibody whose fitness value is larger is suppressed and a new antibody is generated randomly. If the difference of fitness between two antibodies is less than the suppression threshold, these two antibodies are called similar ones.

3 Performance Estimation of Algorithm

3.1 Algorithm Convergence

Theorem 1. *Population sequence of quantum evolutionary algorithm based on immune operator (MQEA)* $\{A(n),\ n \geq 0\}$ *are finite stochastic Markov chain.*

We assume that S is the feasible solutions space and f^* is the optimal solutions of S, let $A^* = \{A|max(f(A)) = f^*, \forall A \in S\}$

Definition 1. $\{A(n), n \geq 0\}$ *are stochastic states,* $S_0 \in S, S_0$ *is the initial solution , If*

$$\lim_{k \to \infty} P\{A(k) \in A^*|A(0) = S_0\} = 1,$$

Then the $\{A(n), n \geq 0\}$*is called convergence with probability one[9].*

Let P_k denote $P\{A(k) \in A^*|A(0) = S_0\}$,then $P_k = \sum_{i \in A^*} P\{A(k) = i|A(0) = S_0\}$.
Let $P_i(k)$ denote $P\{A(k) = i|A(0) = S_0\}$, then

$$P_k = \sum_{i \in A^*} P_i(k) \tag{1}$$

Let $P_{ij}(k) = P\{A(k) = j|A(0) = i\}$. Under elitist approach (the best individual survives with probability one), we have two special equations [9]:
When $i \in A^*, j \notin A^*$,

$$P_{ij}(k) = 0 \tag{2}$$

When $i \in A^*, j \in A^*$,

$$P_{ij}(k) = 1 \tag{3}$$

Theorem 2. *MQEA is convergence with probability one.*

Proof. From the above Eq. (1). $P_k = \sum_{i \in A^*} P_i(k)$.
From $\sum_{j \in A^*} P_{ij}(1) + \sum_{j \notin A^*} P_{ij}(1) = 1$, thus $P_k = \sum_{i \in A^*} P_i(k)$
$= \sum_{i \in A^*} P_i(k)(\sum_{j \in A^*} P_{ij}(1) + \sum_{j \notin A^*} P_{ij}(1))$
$= \sum_{i \in A^*} \sum_{j \in A^*} P_i(k)P_{ij}(1) + \sum_{i \in A^*} \sum_{j \notin A^*} P_i(k)P_{ij}(1)$
From above Eq. (2). $\sum_{i \in A^*} \sum_{j \notin A^*} P_i(k)P_{ij}(1) = 0$,
so $P_k = \sum_{i \in A^*} \sum_{j \in A^*} P_i(k)P_{ij}(1)$.
$\{A(n), n \geq 0\}$ of MQEA is finite stochastic Markov chain (By Theorem 1).
Thus $P_{k+1} = \sum_{i \in A^*} \sum_{j \in A^*} P_i(k)P_{ij}(1) + \sum_{i \notin A^*} \sum_{j \in A^*} P_i(k)P_{ij}(1)$,
so $P_{k+1} = P_k + \sum_{i \notin A^*} \sum_{j \in A^*} P_i(k)P_{ij}(1) > P_k$,
thus $1 \geq P_{k+1} > P_k > P_{k-1} > ... > 0$,
therefore

$$\lim_{k \to \infty} P_k = 1$$

By definition 1, MQEA is convergence with probability one. □

3.2 Guidelines of Parameter Selection

The MQEA is a very compact and fast adaptive search algorithm based on the adaptive immune operator taking a but delicate balance between explorations, i.e., global search, and exploitation, i.e., local search. A major factor contributing to evolution is mutation, which can be caused by spontaneous misreading of bases . The cross-mutation operator could be viewed as a self-adaptive mutation operator. Self-adaptive mutation plays the key role in MQEA, because of the different mutation rates, thus diversity of MQEA is maintained in a population as generations proceed. Generally, cells with low affinity receptors are mutated

at a higher rate (ph), whereas cells with high affinity receptors will have a lower mutation rate (pl). This mechanism offers the ability to escape from local optima on an affinity landscape. To explore the role of mutation on the quality of the memory cells evolved, we modified the mutation routine so that the rate of mutation is dictated by the cells affinity value. Specifically, the higher the normalized affinity value, the smaller the rate of mutation allowed. The rate of mutation for a given cell is as follows:

$$p_s = 1.0 - \frac{|fitness\ of\ given\ cell|}{|maxfitness\ of\ population|} \tag{4}$$

Mutation operator is then a really Self-adaptive, this, in a sense allows for tight exploration of the space around high quality cells, but allows lower quality cells more freedom to exploit widely. in this way, both local refinement and diversification through exploration or exploition are achieved.

Considering time efficiency, we also can divide the population into several groups($g_1, g_2 ... g_r$) according to cells fitness, fitness of cell in group g_i is bigger than that of cell in $g_j (i < j)$, each group g_i has one mutation rate p_{mi}.

$$p_{mi} = 1.0 - \frac{|maxfitness\ of\ group\ g_i|}{|maxfitness\ of\ population|} \tag{5}$$

so $p_{m1} < p_{m2} < ... < p_{mr}$. Generally,the parameter r is set at 3 or 4. The experiment results of stochastic simulation are given to show how the selection of the parameter value influences the convergence of the population in MQEA.

4 Experimental Study

In this section, MQEA is applied to the optimization of well-known test functions and its performance is compared with that of QEA algorithm. Guidelines of parameter selection are evaluated by experiment results of stochastic simulation. The test examples used in this study are listed below:

$$f2(x_1, x_2) = 100(x_1^2 - x_2)^2 + (1 - x_1)^2, -2 \le x_i \le 2 \tag{6}$$

$$f1(x_1, x_2) = -20 * e^{-0.2* \sqrt{\frac{x_1^2 + x_2^2}{2}}} - e^{\frac{\cos(2\pi x_1) + \cos(2\pi x_2)}{2}} + 20 + e \tag{7}$$

f_2(Rosenbrock function): Rosenbrocks valley is a classic optimization function. The global optimum is inside a long, narrow, parabolic shaped flat valley. To find the valley is trivial, however convergence to the global optimum is difficult and hence this problem has been repeatedly used to assess the performance of optimization algorithms. The results for the case of Rosenbrocks function with three variables, averaged over 20 trials, are shown in Fig.1.,where solid line denotes MQEA and dot line denotes QEA. Comparison of the results indicates

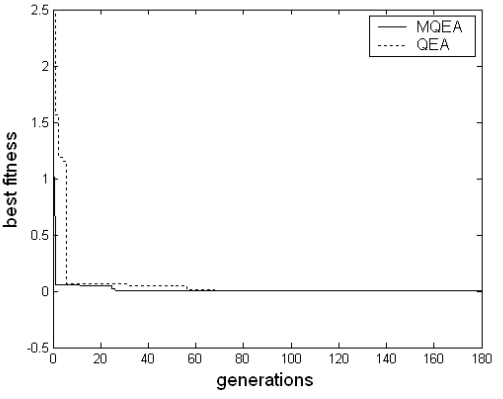


Fig. 1. 25 populations, 180 generations, average 20 trials

that MQEA offers a significant improvement in the results compared to the conventional QEA. In Fig.1., for 25 populations, the optimization value was obtained with QEA after 70 generations, whereas the hybrid method MQEA was able to obtain after 25 generations. This certainty of convergence of the MQEA may be attributed to its ability to maintain the diversity of its population. As a result, fresh feasible antibodies are constantly introduced, yielding a broader exploration of the search space and preventing saturation of the population with identical antibodies. In fact, the IAs superior performance may be attributed to its ability to generate many good antibodies, a larger pool of feasible solutions enhances the probability of finding the optimum solution.

f_1(Ackley function): It is multimodal function with many local minima, one of them is global minimum, $f_{min} = f(0,0) = 0$. It was used to evaluate how the selection of the parameter value influences the convergence of the population in MQEA by using an adaptive immune operator. The result for the case of Ackley function, averaged over 20 trials, are shown in Tabel 1. Comparison of the result indicates that adaptive immune operator can keep the individual diversity and control the convergence speed. For 20 populations, the optimization value was obtained with MQEA-1(adaptive mutation rate, $r = 3, p_{m1} = 0, p_{m2} = 0.1, p_{m3} = 0.2$) after 40 generations, whereas MQEA-2(fixed mutation rate, $p = 0.1$) was able to obtain after 125 generations.

Table 1. Comparison of the performance of MQEA by the selection of the parameter value

MQEA	adaptive mutation rate($r = 3$)	fixed mutation rate
parameter value	$p_{m1} = 0, p_{m2} = 0.1, p_{m3} = 0.2$	$p = 0.1$
iterations number to have solution	40	125

5 Conclusions

In this study, a novel quantum evolutionary algorithm based on immune operator has been presented by using an immune algorithm to imitate the features of a biological immune system. The balance between exploration and exploitation of solutions within a search space are realized through the integration of clonal proliferation, clonal selection, memory antibodies, and the adaptive immune response associated with several diversification schemes, the efficiency of quantum evolutionary algorithm is enhanced by using the immune operator. By combining the two methods, the advantages of both methods are exploited to produce a hybrid optimization method which is both robust and fast, the immune operator is used to improve the convergence of the quantum evolutionary algorithm in search for global optimum. we estimate the performance of algorithm, we also describe technique for improving the performance of MQEA. The efficiency of the approach has been illustrated by applying to a number of test cases. The results show that integration of the adaptive immune algorithm in the quantum evolutionary algorithm procedure can yield significant improvements in both the convergence rate and solution quality. The further work is exploiting more reasonable parameter used in evolutionary model.

References

1. Narayanan,A., Moore,M.: Genetic Quantum Algorithm and its Application to Combinatorial Optimization Problem. In: Proc. IEEE International Conference on Evolutionary Computation (ICEC96), IEEE Press, Piscataway , (1996)61-66
2. Grover, L.K.: A Fast Quantum Mechanical Algorithm for Database Search. In: Proceedings of the 28th Annual ACM Symposium on the Theory of Computing (STOC), ACM Press (1996) 212-219
3. Han,K.H., Kim,J.H.: Quantum-Inspired Evolutionary Algorithms with a New Termination Criterion, $H\varepsilon$ Gate, and Two-Phase Scheme. IEEE Transactions on Evolutionary Computation,IEEE Press, 8(2004) 156-169
4. Han,K.H., Kim,J.H.: Quantum-inspired Evolutionary Algorithm for a Class of Combinatorial Optimization. IEEE Transactions on Evolutionary Computation ,6(2002) 580-593
5. Fukuda,T., Mori,K., Tsukiyama,M.: Parallel Search for Multi-modal Function Optimization with Diversity and Learning of Immune Algorithm.Artificial Immune Systems and Their Applications, Spring-Verlag, Berlin, (1999) 210-220
6. Mori,K., Tsukiyama,M., Fukuda,T.: Adaptive Scheduling System Inspired by Immune Systems. In: Proc. IEEE International Conference on Systems, Man, and Cybernetics, San Diego CA, 12-14 October (1998) 3833-3837
7. Ada,G.L., Nossal,G.J.V.: The Clonal Selection Theory. Scientific American, 257(1987) 50-57
8. Dasgupta,D.: Artificial Immune Systems and Their Applications. Springer-Verlag, Berlin, Germany (1999)
9. Pan,Z.J., Kang,L.S., Chen,Y.P.:Evolutionary Computation. Tsinghua University Press, Bei-jing (1998)

Design of Fuzzy Set-Based Polynomial Neural Networks with the Aid of Symbolic Encoding and Information Granulation

Sung-Kwun Oh, In-Tae Lee, and Hyun-Ki Kim

Department of Electrical Engineering, The University of Suwon,
San 2-2 Wau-ri, Bongdam-eup, Hwaseong-si, Gyeonggi-do, 445-743, South Korea
ohsk@suwon.ac.kr

Abstract. In this paper, we introduce fuzzy-neural networks– Fuzzy Polynomial Neural Networks (FPNN) with fuzzy set-based polynomial neurons (FSPN) whose fuzzy rules include the information granules (about the real system) obtained through Information Granulation. We have developed a design methodology (genetic optimization using Genetic Algorithms) to find the optimal structure for fuzzy-neural networks that expanded from Group Method of Data Handling (GMDH). It is the number of input variables, the order of the polynomial, the number of membership functions, and a collection of the specific subset of input variables that are the parameters of FPNN fixed by aid of genetic optimization that has search capability to find the optimal solution on the solution space. We adopt fuzzy set-based fuzzy rules as substitute for fuzzy relation-based fuzzy rules and apply the concept of Information Granulation to the proposed fuzzy set-based rules. The performance of genetically optimized FPNN (gFPNN) with fuzzy set-based polynomial neurons (FSPN) composed of fuzzy set-based rules is quantified through experimentation where we use a number of modeling benchmarks data which are already experimented with in fuzzy or neurofuzzy modeling.

1 Introduction

It is expected that efficient modeling techniques should allow for a selection of pertinent variables and a formation of highly representative datasets. Furthermore, the resulting models should be able to take advantage of the existing domain knowledge (such as a prior experience of human observers or operators) and augment it by available numeric data to form a coherent data-knowledge modeling entity. Most recently, the omnipresent trends in system modeling are concerned with a broad range of techniques of computational intelligence(CI) that dwell on the paradigm of fuzzy modeling, neurocomputing, and genetic optimization[1, 2]. The list of evident landmarks in the area of fuzzy and neurofuzzy modeling [3, 4] is impressive. While the accomplishments are profound, there are still a number of open issues regarding structure problems of the models along with their comprehensive development and testing.

As one of the representative advanced design approaches comes a family of self-organizing networks with fuzzy polynomial neuron (FPN) (called "FPNN" as a new

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 1076–1083, 2007.
© Springer-Verlag Berlin Heidelberg 2007

category of neuro-fuzzy networks) [5, 8, 12]. The design procedure of the FPNNs exhibits some tendency to produce overly complex networks as well as comes with a repetitive computation load caused by the trial and error method being a part of the development process.

In this paper, in considering the above problems coming with the conventional FPNN [9, 8, 12], we introduce a new structure of fuzzy rules as well as a new genetic design approach. The new structure of fuzzy rules based on the fuzzy set-based approach changes the viewpoint of input space division. In other hand, from a point of view of a new understanding of fuzzy rules, information granules seem to melt into the fuzzy rules respectively. The determination of the optimal values of the parameters available within an individual FSPN leads to a structurally and parametrically optimized network through the genetic approach.

2 The Architecture and Development of FSPNN

The Fuzzy Set-based Polynomial Neuron(FSPN) encapsulates a family of nonlinear "if-then" rules. When put together, FSPNs results in a self-organizing Fuzzy Set-based Polynomial Neural Networks (FSPNN). Each rule reads in the form.

$$\begin{aligned} &\text{if } x_p \text{ is } \boldsymbol{A_k} \text{ then } z \text{ is } P_{pk}(x_i, x_j, \boldsymbol{a}_{pk}) \\ &\text{if } x_q \text{ is } \boldsymbol{B_k} \text{ then } z \text{ is } P_{qk}(x_i, x_j, \boldsymbol{a}_{qk}) \end{aligned} \tag{1}$$

where $\boldsymbol{a}_{qk}$ is a vector of the parameters of the conclusion part of the rule while $P(x_i, x_j, \boldsymbol{a})$ denoted the regression polynomial forming the consequence part of the fuzzy rule. The activation levels of the rules contribute to the output of the FSPN being computed as a weighted average of the individual condition parts (functional transformations) P_K. (note that the index of the rule, namely "k" is a shorthand notation for the two indices of fuzzy sets used in the rule (1), that is $K=(l,k)$).

$$\begin{aligned} z &= \sum_{l=1}^{\text{total inputs}} \left(\sum_{k=1}^{\text{total_rules related to input } l} \mu_{(l,k)} P_{(l,k)}(x_i, x_j, \boldsymbol{a}_{(l,k)}) \Big/ \sum_{k=1}^{\text{total_rules related to input } l} \mu_{(l,k)} \right) \\ &= \sum_{l=1}^{\text{total inputs}} \left(\sum_{k=1}^{\text{rules related to input } l} \tilde{\mu}_{(l,k)} P_{(l,k)}(x_i, x_j, \boldsymbol{a}_{(l,k)}) \right) \end{aligned} \tag{2}$$

In the above expression, we use an abbreviated notation to describe an activation level of the "k"th rule to be in the form

$$\tilde{\mu}_{(l,k)} = \frac{\mu_{(l,k)}}{\sum_{k=1}^{\text{total rule related to input } l} \mu_{(l,k)}} \tag{3}$$

When developing an FSPN, we use genetic algorithms to produce the optimized network. This is realized by selecting such parameters as the number of input variables, the order of polynomial, and choosing a specific subset of input variables. Based on the genetically optimized number of the nodes (input variables) and the polynomial order, refer to Table 1, we construct the optimized self-organizing network architectures of the FSPNNs.

Table 1. Different forms of the regression polynomials forming the consequence part of the fuzzy rules

No. of inputs / Order of the polynomial	1	2	3
0 (Type 1)	Constant	Constant	Constant
1 (Type 2)	Linear	Bilinear	Trilinear
2 (Type 3)	Quadratic	Biquadratic-1	Triquadratic-1
2 (Type 4)		Biquadratic-2	Triquadratic-2

1: Basic type, 2: Modified type

3 Information Granulation Through Hard C-Means Clustering Algorithm

We assume that given a set of data $X=\{\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n\}$ related to a certain application, there are some clusters which are capable of being found through HCM. The center point and the membership elements represent each cluster. The set of membership elements is crisp. To construct a fuzzy mode, we should transform the crisp set into the fuzzy set. The center point means the apex of the membership function of the fuzzy set. Let us consider building the consequent part of fuzzy rule. We can think of each cluster as a sub-model composing the overall system. The fuzzy rules of Information Granulation-based FSPN are as followings.

$$\begin{aligned} &\text{if } x_p \text{ is } \boldsymbol{A}^*_k \text{ then } z\text{-}m_{pk} = P_{pk}((x_i - v^i_{pk}),(x_j - v^j_{pk}),a_{pk}) \\ &\text{if } x_q \text{ is } \boldsymbol{B}^*_k \text{ then } z\text{-}m_{qk} = P_{qk}((x_i - v^i_{qk}),(x_j - v^j_{qk}),a_{qk}) \end{aligned} \tag{4}$$

Where, $\boldsymbol{A}^*_k$ and $\boldsymbol{B}^*_k$ mean the fuzzy set, the apex of which is defined as the center point of information granule (cluster) and m_{pk} is the center point related to the output variable on cluster$_{pk}$, v^i_{pk} is the center point related to the *i-th* input variable on cluster$_{pk}$ and $\boldsymbol{a}_{qk}$ is a vector of the parameters of the conclusion part of the rule while $P((x_i - v^i),(x_j - v^j),a)$ denoted the regression polynomial forming the consequence part of the fuzzy rule which uses several types of high-order polynomials (linear, quadratic, and modified quadratic) besides the constant function forming the simplest version of the consequence; refer to Table 1. If we are given m inputs and one output system and the consequent part of fuzzy rules is linear, the overall procedure of modification of the generic fuzzy rules is as followings. The given inputs are $\mathbf{X}=[\mathbf{x}_1\ \mathbf{x}_2\ \ldots\ \mathbf{x}_m]$ related to a certain application, where $\mathbf{x}_k=[x_{k1}\ \ldots\ x_{kn}]^T$, n is the number of data and m is the number of variables and the output is $\mathbf{Y}=[y_1\ y_2\ \ldots\ y_n]^T$.

Step 1) build the universe set
Universe set $U=\{\{x_{11}, x_{12}, \ldots, x_{1m}, y_1\}, \{x_{21}, x_{22}, \ldots, x_{2m}, y_2\}, \ldots, \{x_{n1}, x_{n2}, \ldots, x_{nm}, y_n\}\}$
Step 2) build m reference data pairs composed of $[\mathbf{x}_1;Y]$, $[\mathbf{x}_2;Y]$, and $[\mathbf{x}_m;Y]$.
Step 3) classify the universe set U into $\boldsymbol{l}$ clusters such as c_{i1}, c_{i2}, …, c_{il} (subsets) by using HCM according to the reference data pair $[\mathbf{x}_i;Y]$. Where c_{ij} means the *j-th* cluster (subset) according to the reference data pair $[\mathbf{x}_i;Y]$.

Step 4) construct the premise part of the fuzzy rules related to the *i-th* input variable (x_i) using the directly obtained center points from HCM.
Step 5) construct the consequent part of the fuzzy rules related to the *i-th* input variable (x_i). On this step, we need the center points related to all input variables. We should obtain the other center points through the indirect method as followings.

Sub-step1) make a matrix as equation (5) according to the clustered subsets

$$A_j^i = \left[\begin{array}{cccc|c} x_{21} & x_{22} & \cdots & x_{2m} & y_2 \\ x_{51} & x_{52} & \cdots & x_{5m} & y_5 \\ x_{k1} & x_{k2} & \cdots & x_{km} & y_k \\ \vdots & \vdots & \cdots & \vdots & \vdots \end{array}\right] \tag{5}$$

Where, $\{x_{k1}, x_{k2}, \ldots, x_{km}, y_k\} \in c_{ij}$ and A_j^i means the membership matrix of ***j-th*** subset related to the ***i-th*** input variable.
Sub-step2) take an arithmetic mean of each column on A_j^i. The mean of each column is the additional center point of subset c_{ij}. The arithmetic means of column is equation (6)

$$center\ point = \left[\begin{array}{ccc|c} v_{ij}^1 & v_{ij}^2 & \cdots & v_{ij}^m & m_{ij} \end{array}\right] \tag{6}$$

Step 6) if i is m then terminate, otherwise, set i=i+1 and return step 3.

4 The Design Procedure of Genetically Optimized FSPNN

The framework of the design procedure of the genetically optimized Fuzzy Polynomial Neural Networks (FPNN) with fuzzy set-based PNs (FSPN) comprises the following steps

[Step 1] *Determine system's input variables*
[Step 2] *Form training and testing data*
[Step 3] *specify initial design parameters*
[Step 4] *Decide upon the FSPN structure through the use of the genetic design*
[Step 5] *Carry out fuzzy-set based fuzzy inference and coefficient parameters estimation for fuzzy identification in the selected node(FSPN)*
[Step 6] *Select nodes (FSPNs) with the best predictive capability and construct their corresponding layer*
[Step 7] *Check the termination criterion*
[Step 8] *Determine new input variables for the next layer*

5 Experimental Studies

We demonstrate how the IG-based gFSPNN can be utilized to predict future values of a chaotic Mackey-Glass time series. This time series is used as a benchmark in fuzzy

Table 2. Summary of the parameters of the genetic optimization

	Parameters	1st layer	2nd to 3rd layer
GA	Maximum generation	150	150
	Total population size	300	300
	Selected population size (*W*)	30	30
	Crossover rate	0.65	0.65
	Mutation rate	0.1	0.1
	String length	Max*2+1	Max*2+1
FSPNN	Maximal no.(Max) of inputs to be selected	$1 \le l \le$ Max(2~5)	$1 \le l \le$ Max(2~5)
	Polynomial type (Type T) of the consequent part of fuzzy rules	$1 \le T \le 4$	$1 \le T \le 4$
	Consequent input type to be used for Type T (*)	Type T	Type T
		Type T*	Type T
	Membership Function (MF) type	Triangular	Triangular
		Gaussian	Gaussian
	No. of MFs per input	2 ~ 5	2 ~ 5

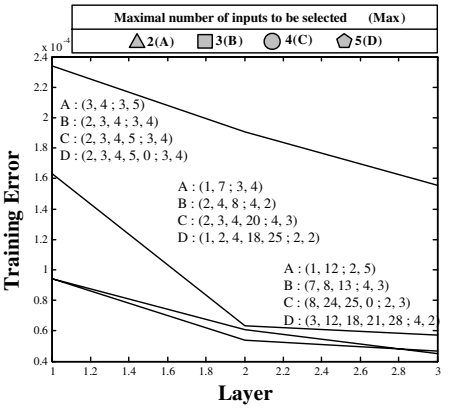
l, T, Max: integers, T* means that entire system inputs are used for the polynomial in the conclusion part of the rules.

and neurofuzzy modeling. The time series is generated by the chaotic Mackey-Glass differential delay equation. To come up with a quantitative evaluation of the network, we use the standard RMSE performance index. Table 2. depicts the parameters of genetic optimization related to the proposed network.

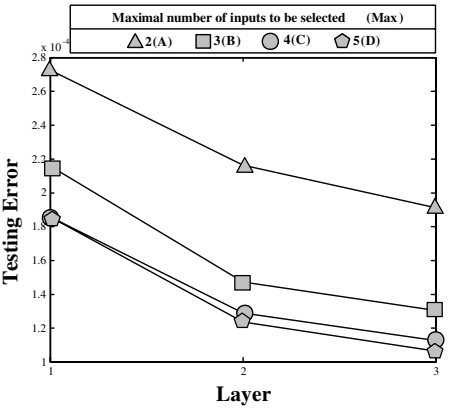
The population size being selected from the total population size (300) is equal to 30. The process is realized as follows. 150 nodes are generated in each layer of the network. The parameters of all nodes generated in each layer are estimated and the network is evaluated using both the training and testing data sets. Then we compare these values and choose 30 FSPNs that produce the best (lowest) value of the performance index. The maximal number (Max) of inputs to be selected is confined to two to five (2-5) entries. The polynomial order of the consequent part of fuzzy rules is chosen from four types, that is Type 1, Type2, Type 3, and Type 4.

Fig. 1 depicts the performance index of each layer of gFSPNN according to the increase of maximal number of inputs to be selected. In Fig. 1, A(•)- D(•) denote the optimal node numbers at each layer of the network.

Fig. 2 illustrates the detailed optimal topologies of gFSPNN for three layer when using Max=5. As shown in Fig 2, the proposed network enables the architecture to be a structurally more optimized and simplified network than the conventional FPNN. In nodes (FSPNs) of Fig. 2, 'FSPNn' denotes the n^{th} FSPN (node) of the corresponding layer, the number of the left side denotes the number of nodes (inputs or FSPNs)

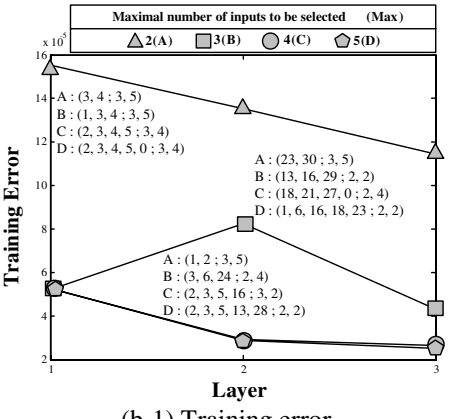


(a-1) Training error

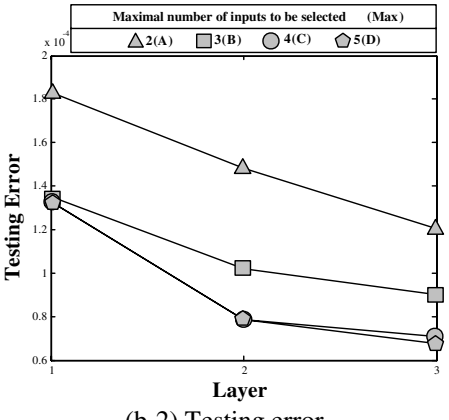


(a-2) Testing error

(a) Triangular membership function



(b-1) Training error



(b-2) Testing error

(b) Gaussian-like membership function

Fig. 1. Performance index of IG-gFSPNN (with Type T*) with respect to the increase of number of layers

coming to the corresponding node, the number of the center denote the polynomial order of conclusion part of fuzzy rules used in the corresponding node and the number of the right side denotes the number of membership functions.

Table 3 summarizes a comparative analysis of the performance of the network with other models. The experimental results clearly reveal that it outperforms the existing models both in terms of significant approximation capabilities (lower values of the performance index on the training data, PI_s) as well as superb generalization abilities (expressed by the performance index on the testing data EPI_s). The proposed IG based genetically optimized FSPNN (IG-based gFSPNN) can be much lower in comparison to the conventional optimized FPNN as shown in Table 3.

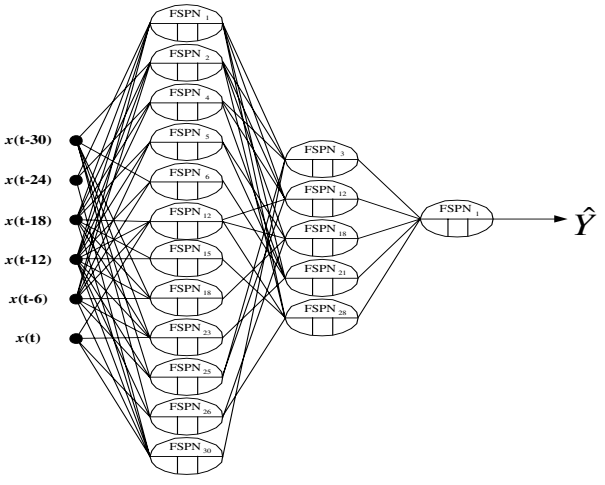


Fig. 2. Optimal networks structure of GAs-based FPNN (for 3 layers)

Table 3. Comparative analysis of the performance of the network; considered are models reported in the literature

Model				Performance index		
				PI	PI_s	EPI_s
Wang's model[9]				0.044		
				0.013		
				0.010		
ANFIS[10]					0.0016	0.0015
FNN model[11]					0.014	0.009
Proposed IG-gFSPNN	Type T*	Triangular (3^{nd} layer)	Max=5		4.65e-5	1.06e-4
		Gaussian (3^{nd} layer)	Max=5		2.49e-5	6.76e-5

6 Concluding Remarks

In this study, we have surveyed the new structure and meaning of fuzzy rules and investigated the GA-based design procedure of Fuzzy Polynomial Neural Networks (FPNN) along with its architectural considerations. The whole system is divided into some sub-systems that are classified according to the characteristics named information granules. Each information granule seems to be a representative of the related sub-systems. A new fuzzy rule with information granule describes a sub-system as a stand-alone system. A fuzzy system with some new fuzzy rules depicts the whole system as a combination of some stand-alone sub-system.

The GA-based design procedure applied at each stage (layer) of the FSPNN leads to the selection of the preferred nodes (or FSPNs) with optimal local characteristics (such as the number of input variables, the order of the consequent polynomial of fuzzy rules, and input variables) available within FSPNN. The comprehensive

experimental studies involving well-known datasets quantify a superb performance of the network in comparison to the existing fuzzy and neuro-fuzzy models.

Acknowledgements. This work was supported by the Korea Research Foundation Grant funded by the Korean Government (MOEHRD)(KRF-2006-311-D00194, Basic Research Promotion Fund).

References

1. Pedrycz, W.: Computational Intelligence: An Introduction, CRC Press, Florida, (1998)
2. Peters, J.F, Pedrycz, W.: Computational intelligence, In Encyclopedia of Electrical and Electronic Engineering, (Edited by J.G. Webster). John Wiley & Sons, New York. 22 (1999)
3. Pedrycz, W., Peters, J.F. (ed.): Computational Intelligence in Software Engineering, World Scientific, Singapore. (1998)
4. Takagi, H., Hayashi, I.: NN-driven fuzzy reasoning, Int. J. of Approximate Reasoning. 5 (3) (1991), 191-212
5. S.-K. Oh and W. Pedrycz, "Self-organizing Polynomial Neural Networks Based on PNs or FPNs : Analysis and Design", *Fuzzy Sets and Systems*, 2003(in press).
6. Z. Michalewicz, "*Genetic Algorithms + Data Structures = Evolution Programs*", Springer-Verlag, Berlin Heidelberg, 1996.
7. D. Jong, K. A., "*Are Genetic Algorithms Function Optimizers?*", Parallel Problem Solving from Nature 2, Manner, R. and Manderick, B. eds., North-Holland, Amsterdam.
8. S.-K. Oh and W. Pedrycz, "Fuzzy Polynomial Neuron-Based Self-Organizing Neural Networks", *Int. J. of General Systems*, Vol. 32, No. 3, pp. 237-250, May, 2003.
9. L. X. Wang, J. M. Mendel, "Generating fuzzy rules from numerical data with applications", *IEEE Trans. Systems, Man, Cybern.*, Vol. 22, No. 6, pp. 1414-1427, 1992.
10. J. S. R. Jang, "ANFIS: Adaptive-Network-Based Fuzzy Inference System", *IEEE Trans. System, Man, and Cybern*., Vol. 23, No. 3, pp. 665-685, 1993.
11. L. P. Maguire, B. Roche, T. M. McGinnity, L. J. McDaid, "Predicting a chaotic time series using a fuzzy neural network", *Information Sciences*, Vol. 112, pp. 125-136, 1998.
12. S.-K. Oh, W. Pedrycz and T.-C. Ahn, "Self-organizing neural networks with fuzzy polynomial neurons", *Applied Soft Computing*, Vol. 2, Issue 1F, pp. 1-10, Aug. 2002.
13. L. A. Zadeh, "Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic", *Fuzzy Sets Syst*, vol. 90, pp. 111-117, 1997
14. Park, B.J., Lee, D.Y., Oh, S.K.**:** Rule-based Fuzzy Polynomial Neural Networks in Modeling Software Process Data. Int. J. of Control Automation and Systems. **1**(3) (2003) 321-331

An Heuristic Method for GPS Surveying Problem

Stefka Fidanova

IPP – Bulgarian Academy of Sciences, Acad. G. Bonchev str. bl.25A, 1113 Sofia, Bulgaria
stefka@parallel.bas.bg

Abstract. This paper describes metaheuristic algorithm based on simulated annealing method,which is a nature-inspired method, to analyze and improve the efficiency of the design of Global Positioning System (GPS) surveying networks. Within the context of satellite surveying, a positioning network can be defined as a set of points which are coordinated by placing receivers on these point to determine sessions between them. The problem is to search for the best order in which these sessions can be observed to give the best possible schedule. The same problem arise in Mobile Phone surveying networks. Several case studies have been used to experimentally asses the performance of the proposed approach in terms of solution quality and computational effort.

1 Introduction

The continuing research on naturally occurring social systems offers the prospect of creating artificial systems that generate practical solutions to many Combinatorial Optimization Problems (COPs). Metaheuristic techniques have evolved rapidly in an attempt to find good solutions to these problems within a desired time frame [9]. They attempt to solve complex optimization problems by incorporating processes which are observed at work in real life [2,4]. The paper proposes simulated annealing algorithm which provide optimal or near optimal solution for large networks with an acceptable amount of computational effort.

The Global Positioning System is a satellite-based radio-navigation system that permits land, sea and airborne users to determine their three-dimensional position, velocity and time. This can be achieved 24 hours a day in all weather. In addition the satellite navigation system has an impact in all related fields in geoscience and engineering, in particular on surveying work in quickly and effectively determining locations and changes in locations network. The purpose of surveying is to determine the locations of points on the earth. Measuring tapes or chains require that the survey crew physically pass through all the intervening terrain to measure the distance between two points. Surveying methods have undergone a revolutionary change over that last few years with the deployment of the satellite navigation systems. The most widely known space systems are: the American Global Positioning System (GPS), the Russian GLObal Navigation Satellite System (GLONASS), and the forthcoming European Satellite

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 1084–1090, 2007.
© Springer-Verlag Berlin Heidelberg 2007

Navigation System (GALILEO). In this paper, it is the use of GPS to establish surveying networks that is being investigated. GPS satellites continuously transmit electrical signals to the earth while orbiting the earth. A receiver, with unknown position on the earth, has to detect and convert the signals transmitted from all of the satellites into useful measurements. These measurements would allow a user to compute a three-dimensional coordinates position for the location of the receiver.

The rest of the paper is organized as follows. The general framework for GPS surveying network problem as a combinatorial optimization problem is described in Section 2. Then the search strategy of the simulated annealing is explained in Section 3. The metaheuristic algorithm applied to GPS surveying network is outlined and the general case of the problem is addressed, by presenting several case studies and the obtained numerical results in Section 4. The paper ends with a summary of the conclusions and directions for future research.

2 Formulation of the GPS Surveying Network Problem

A GPS network is distinctly different from a classical survey network in that no inter-visibility between stations is required. In GPS surveying, after defining the locations of the points for an area to be surveyed, GPS receivers will be used to map this area by creating a network of these coordinated points. These points, control stations within the context of surveying, are fixed on the ground and located by an experienced surveyor according to the nature of the land and the requirements of the survey [7]. At least two receivers are required to simultaneously observe GPS satellites, for a fixed period of time, where each receiver is mounted on each station. The immediate outcome of the observation is a session between these two stations. After completing the first stage of sessions observation and defining the locations of the surveyed stations, the receivers are moved to other stations for similar tasks till the whole network is completely observed according to an observation schedule. The total cost of carrying out the above survey, which is computed upon the criteria to be minimized, represents the cost of moving receivers between stations. The problem is to search for the best order, with respect to the time, in which these sessions can be observed to give the cheapest schedule V, i.e.:

$$Minimize: \; C(V) = \sum_{p \in R} C(S_p)$$

where: $C(V)$ is the total cost of a feasible schedule V;
S_p is the route of the receiver p in a schedule;
R is the set of receivers $R = \{1, \ldots, r\}$;

3 Simulated Annealing Method

The Simulated Annealing (SA) is a generalization of an optimization method for examine the equations of state and frozen states of n-body systems. The

original Metropolis scheme [8] was that an initial state of a thermodynamic system was chosen at energy E and temperature T, holding T constant the initial configuration is perturbed and the change of energy dE is computed. The current state of the thermodynamic system is analogous to the current solution to the combinatorial problem, the energy equation for the thermodynamic system is analogous to the objective function and ground state is analogous to the global minimum. The key objective of this paper is to find an effective solution in a short period of time with close to least cost for a given GPS network using Simulated Annealing method.

SA is a heuristic method that has been implemented to obtain good solutions of an objective function defined on a number of discrete optimization problems. This method has proved to be a flexible local search method and can be successfully applied to the majority of real-life problems [1,3,5,6,10,12]. The origin of the algorithm is in statistical mechanics. The fundamental idea is to allow moves resulting in solutions of worse quality than the current solution in order to escape from local minimum. The probability of doing such a move is declared during the search. The algorithm starts by generating an initial solution and by initializing the so-called temperature parameter T. The temperature is decreased during the search process, thus at the beginning of the search the probability of accepting uphill moves is high and it gradually decreases. This process is analogous to the annealing process of metals and glass, which assume a low energy configuration when cooled with an appropriate cooling schedule. Regarding the search process, this means that the algorithm is the result of two combined strategies:

- random walk;
- iterative improvement.

The first phase permits the exploration of the search space. The advantages of the algorithm are:

- Simulated Annealing is proved to converge to the optimal solution of the problem;
- An easy implementation of the algorithm makes it very easy to adapt a local search method to a simulated annealing algorithm.

In order to implement SA for GPS surveying problem a number of decisions and choices have to be made. Firstly , the problem specific choices, which determine the way in which the GPS surveying is modeled in order to fit into the SA framework. In other words, it involves the definition of the solution space Q and neighborhood structure I, the form of the objective function $C(V)$ and the way in which a starting solution V is obtained. Secondly, the generic choices which govern the working of the algorithm itself, are mainly concerned with the components of the cooling parameters: control parameter T and its initial starting value, the cooling rate F and the temperature update function, the number of iterations between decreases L and the condition under which the system will be terminated [11]. The performance of the achieved result is highly dependent on the right choice of both specific and generic desisions.

The SA procedure used in this work is designed and developed essentially from practical experience and the requirement of the GPS surveying. A simple constructive procedure is proposed to obtain an initial (starting) feasible scheduling V for the GPS surveying. The aim of this simple procedure is to obtain quickly an initial schedule. The structure of the SA algorithm is shown below:

1. The Problem Specific Decisions
Step 1. Formulation the problem parameters;

Step 2. Determination of the initial schedule, generate a feasible solution V;

2. The Problem Generic Decisions
Initialization the cooling parameters:
Set the initial starting value of the temperature parameter, $T > 0$;
Step 3. Set the temperature length, L;
Set the cooling ratio, F;
Set the number of iterations, $K = 0$.

3. The Generation Mechanism, Selecting and Acceptance Strategy of Generated Neighbors
Select a neighbor V' of V where $V' \in I(V)$
Step 4. Let $C(V') =$ the cost of the schedule V'
Compute the move value $\Delta = C(V') - C(V)$

If $\Delta \leq 0$ accept V' as a new solution and set $V = V'$
ELSE
Step 5. IF $e^{-\Delta/T} > \Theta$ set $V = V'$,
where Θ is a uniform random number $0 < \Theta < 1$
OTHERWISE retain the current solution V.

4. Updating the Temperature
Step 6. Updating the annealing schedule parameters using the cooling schedule $T_{k+1} = F * T_k$ $k = \{1, 2, \ldots\}$.

5. Termination of the Solution
Step 7. IF the stopping criteria is met THEN

Show the output
Step 8. Declare the best solution
OTHERWISE Go to step 4.

4 Simulated Annealing Algorithm for GPS Surveying

To understand the simulated annealing method, the Local Search (LS) strategy is throughly introduced. It is based on the concept of searching the local neighborhood of the current schedule [14]. The LS procedure perturb a given solution to generate different neighborhoods using a move generation mechanism. In

general, neighborhood for large-size problems can be much complicated to search. Therefor, LS attempts to improve an initial schedule v to a GPS network by a series of local improvements. A move generation is a transition from a schedule V to another one $V' \in I(V)$ in one step (iteration). These schedules are selected and accepted according to some pre-defined criteria. The returned schedule V may not be optimal, but it is the best schedule in its local neighborhood $I(V)$. A local optimal schedule is a schedule with the local minimal cost value. The basic steps for the LS procedure are as follows:

- Compute the cost value $C(V)$ of the current schedule V;
- Generate a schedule $V' \in I(V)$ and compute its cost value $C(V')$;
- If $C(V) < C(V')$ then, V' replaces V as a current schedule;
- Otherwise, retain V and generate other moves until $C(V') < C(V)$ for all $V' \in I(V)$;
- Terminate the search and return V as the local optimal schedule.

The local search is an essential part of the simulated annealing method. In Saleh [13], a practical local search procedure that satisfies the GPS requirements has been developed. In this procedure, which is based on the sequential session interchange the potential pair-swaps are examined in order $(1, 2)$, $(1, 3)$,..., $(1, n)$, $(2, 3)$, $(2, 4)$, ..., $(n-1, n)$, n is the number of sessions. The solution is represented by graph. The nodes correspond to the sessions and the edges correspond to the observation order.

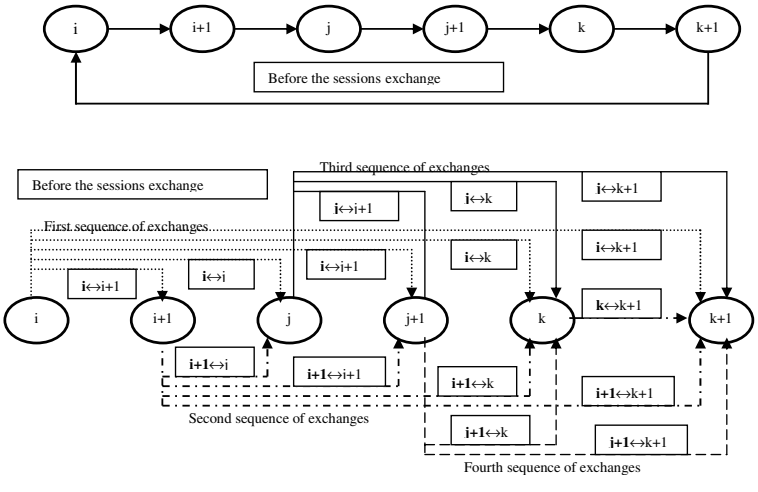


Fig. 1. The sequential local search procedure

In our algorithm we apply different local search procedure. We choose the potential pair-swaps randomly, in [13] they are chosen sequentially. When the best neighbor solution is equal to the global best solution, we choose randomly

new current solution from the set of neighbors. The aim is to prevent $\Delta = 0$ and thus to prevent repetition of the same solutions. It is a way for diversification of the search process.

5 Experimental Results

In this section some experimental results are reported. We compare simulated annealing algorithm used in [13] with our variant of the simulated annealing. Like a test problems we use real data of Malta and Seychelles GPS networks. The Malta GPS network is composed of 38 sessions and the Seychelles GPS network is composed of 71 sessions. We use 6 large test problems too from **http://www.informatik.uni-heidelberg.de/ groups/ comopt/ software/ TSLIB95/ ATSP.html**. These test problems consist from 100 to 443 sessions.

Table 1. Comparison of simulated annealing algorithms applied to various types of GPS networks

Test	Nodes	sequential LS	random LS
Malta	38	1345	1035
Seychelles	71	986	965
rro124	100	125606	59033
ftv170	170	6942	5179
rbg323	323	3787	1867
rbg358	358	2762	1749
rbg403	403	4852	2757
rbg443	443	6509	3111

We use the following parameters for both algorithms: initial temperature is 20% of the result, final temperature is 3, cooling parameter is 0.99, number of iterations is 10 times the number of the nodes, number of swaps is 120% of the number of the nodes. The programs are written on C++ language and are run on Pentium 4 PC with 2.8 GHz 512 RAM. The reported results are average over 25 runs of the algorithms.

Comparing the both simulated annealing algorithms, our algorithm outperforms other for all tested problems. In our algorithm we include random search and we prevent repeating of same solutions. It is a way for diversification of the search, to climb the hills and thus the algorithm has a greater possibility to find new and better solutions.

6 Conclusion

In this paper a new simulated annealing algorithm for GPS surveying problem has been developed and compared with the algorithm used by Saleh in [13]. A comparison of the performance of the both simulated annealing algorithms applied on various GPS networks is reported. The obtained results are encouraging and the ability of the developed techniques to generate rapidly high-quality solutions for observing GPS networks can be seen. The problem is important because it arises in wireless communications like GPS and mobile phone communications and can improve the services in the networks. Thus the problem has an economical importance. For future work we will investigate the influence of the parameter settings and their dependence of the number of the nodes and the expected results.

Acknowledgments. The author is supported by European community grant BIS 21++.

References

1. Arts E., Van Loarhoven P.: Statistical Cooling: A General Approach to Combinatorial Optimization Problems, Phillips Journal of Research 40, (1985), 193–226
2. Bonabeau E., Dorigo M., Theraulaz G.: Swarm Intelligence: From Natural to Artificial Systems, Oxford University Press, New York.(1999)
3. Cerny V.: A Thermodynamical Approach to the Traveling Salesman Problem: An Efficient Simulated Annealing Algorithm, J. of Optimization Theory and Applications 45, (1985), 41–51
4. Corne D., Dorigo M., Glover F. (eds.): New Ideas in Optimization, McCraw Hill, London.(1999)
5. Dowsland K.: Variants of Simulated Annealing for Practical Problem Solving In Rayward-Smith V. ed., Applications of Modern Heuristic Methods, Henley-on-Thames: Alfred Walter Ltd. in association with UNICOM, (1995)
6. Kirkpatrick S., Gelatt C.D., Vecchi P.M.: Optimization by Simulated Annealing, Science 220, (1983),671–680
7. Leick, A.: GPS Satellite Surveying, 2nd. edition. Wiley, Chichester, England.(1995)
8. Metropolis N., Rosenbluth A., Rosenbluth M., Teller A., Teller E.: Equation of State Calculations by Fast Computing Machines, J. of Chem Phys. 21(6), (1953), 1087–1092
9. Osman I.H., Kelly J.P.: Metaheuristics:An overview. In Osman I.H., Kelly J.P. eds. Metaheuristics:Theory and Applications, Kluwer Acad. Publisher.(1996)
10. Osman I.H., Potts C.N.: Simulated Annealing for Permutation Flow-Shop Scheduling, Omega 17, (1989), 551–557
11. Reeves C. R. (ed.): Modern Heuristic Techniques for Combinatorial Problems, Oxford, England, Blackwell Scientific Publications, (1993)
12. Rene V.V.: Applied Simulated Annealing, Berlin, Springer, (1993)
13. Saleh H. A., and Dare P.: Effective Heuristics for the GPS Survey Network of Malta: Simulated Annealing and Tabu Search Techniques. Journal of Heuristics **N 7** (6), (2001), 533-549
14. Schaffer A. A., Yannakakis M.: Simple Local Search Problems that are Hard to Solve, Society for Industrial Applied Mathematics Journal on Computing, Vol 20, (1991), pp. 56-87

Real-Time DOP Ellipsoid in Polarization Mode Dispersion Monitoring System by Using PSO Algorithm

Xiaoguang Zhang [1,2], Gaoyan Duan [1,2], and Lixia Xi [1,2]

[1] Department of Physics, Beijing University of Posts and Telecommunications, Beijing 100876, P.R. China
zhang.x.g@263.net
[2] Key Laboratory of Optical Communication and Lightwave Technologies, Ministry of Education, Beijing University of Posts and Telecommunications, P.R. China

Abstract. In high bit rate optical fiber communication systems, Polarization mode dispersion (PMD) is one of the main factors to signal distortion and needs to be compensated. PMD monitoring system is the key integral part of an adaptive PMD compensator. The degree of polarization (DOP) ellipsoid obtained by using a polarization scrambler can be used as a feedback signal for automatic PMD compensation. Generally, more than several thousands of sampling data of states of polarization (SOP) must be collected to insure getting a correct DOP ellipsoid. This would result in an unacceptable time consuming for adaptive PMD compensation. In this paper, we introduce the particle swarm optimization (PSO) algorithm in determining the real-time DOP ellipsoid with high precision, requiring only 100 sampling data of SOPs. Experimental results confirm that the PSO algorithm is effective for ellipsoid data fitting with high precision within 250ms using our hardware environment.

Keywords: polarization mode dispersion, monitoring techniques, degree of polarization ellipsoid, particle swarm optimization algorithm.

1 Introduction

Polarization mode dispersion (PMD) is one of the main factors to signal distortion and one of the main limiting factors preventing capacity increase of optical fiber communication systems. PMD compensation has become one of hot topics in recent years [1-5]. An ordinary automatic PMD compensator can be divided into three important subparts: the PMD monitoring unit, the compensation unit and the control algorithm. An effective PMD monitoring system allows its monitoring signal to be highly correlated to PMD. An ideal PMD monitoring technique can reveal as much information as possible about the PMD vector, such as both the differential group delay (DGD), which is the magnitude of the vector, and the principal states of polarization (PSPs), the direction of the vector. The DOP ellipsoid obtained by using a polarization scrambler can determine both the DGD and PSP orientation by its three radii and the orientation angle of ellipsoid [2]. The polarization scrambler placed at the fiber input

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 1091–1098, 2007.
© Springer-Verlag Berlin Heidelberg 2007

makes the input random states of polarization (SOPs) of light distribute uniformly over the whole Poincaré sphere in Stokes space. After propagation in a fiber link with PMD, all the output SOP points of light will constitute in Stokes space an ellipsoid (DOP ellipsoid) whose shape is determined by DGD and PSP of the fiber link. In order to extract the PMD information such as DGD or PSP, we need to get the analytical DOP ellipsoid equation from all these measured discrete points of output SOPs. The more points of the output SOPs we measure, the more accurate ellipsoid equation we get, and much time elapse. Generally, an ellipsoid can be distinguished after several thousands of SOPs points are measured, whereas the real-time adaptive PMD compensator requires a real-time feedback signal. In this paper we introduce the particle swarm optimization (PSO) as a powerful data fitting algorithm for getting a precise analytical DOP ellipsoid equation from only 100 measured SOP samples with fast speed, and getting real-time characteristic information of PMD such as DGD and PSP.

2 A Brief Introduction to Polarization Mode Dispersion and PMD Compensation

2.1 Polarization Mode Dispersion

Polarization mode dispersion has its origins in optical birefringence [3]. In a single mode fiber, an optical wave traveling in the fiber can be represented as the linear superposition of two orthogonal polarized HE_{11} modes. In an ideal fiber, with a perfect circular symmetric cross-section, the two modes HE_{11}^{x} and HE_{11}^{y} are indistinguishable (degenerate) with the same time group delay. However, real fibers have some amount of asymmetry due to imperfections in manufacturing process or mechanical stress on the fiber after manufacture as shown in Fig. 1. The asymmetry breaks the degeneracy of the HE_{11}^{x} and HE_{11}^{y} modes, resulting in birefringence with a difference in the phase and group velocities of two modes.

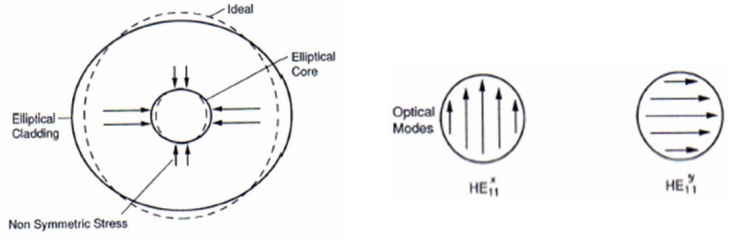


Fig. 1. Asymmetry of a real fiber and degeneracy of two orthogonal HE_{11} modes

If a pulsed optical wave that is linearly polarized at 45° to the birefringence axis is launched into a birefringent fiber, the pulse will be splitted and separated at output end of the fiber due to the different group velocities between two HE_{11} modes, as shown in Fig. 2, resulting in a signal distortion in optical transmission system. The time separation between two modes is defined as differential group delay (DGD) $\Delta\tau$, which is the magnitude of PMD vector. Roughly speaking, the fast and slow axes are

called principal states of polarization, one of which is defined as the direction of the PMD vector. This phenomenon is called polarization mode dispersion.

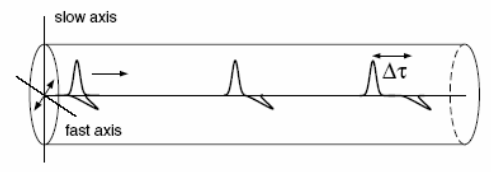


Fig. 2. Pulse splitting due to fiber birefringence

2.2 The PMD Compensator and Monitoring Technique

Fig. 3 shows a PMD compensation system. The optical signal is generated by the transmitter. The polarization scrambler placed just after the transmitter generates the random SOPs distributed over the whole Poincaré sphere in Stokes space. The optical signal will be distorted because of PMD in transmission fiber or made by PMD emulator. The control unit in PMD compensator makes the PMD compensation by adjusting the compensation unit according to the feedback signal generated by the PMD monitoring unit.

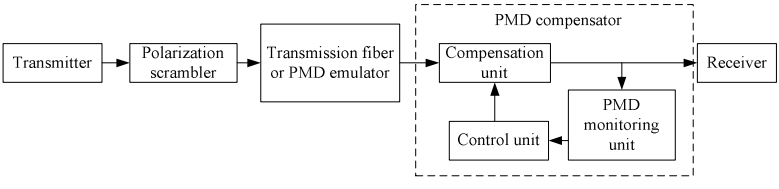


Fig. 3. The PMD compensation system

The PMD monitoring system is the key integral part of an adaptive PMD compensator. The degree of polarization (DOP) ellipsoid obtained by using a polarization scrambler can be used as a feedback monitoring signal for automatic PMD compensation.

2.3 The Theory of DOP Ellipsoid

DOP is defined by the four intensity related Stokes parameters S_0, S_1, S_2, S_3 according to the following equation:

$$\mathrm{DOP} = \frac{\sqrt{S_1^2 + S_2^2 + S_3^2}}{S_0} \tag{1}$$

The states of polarization (SOP) of the optical signals in the fiber link are represented by their Stokes parameters. The DOP ellipsoid obtained by using a polarization scrambler can determine both DGD and PSP orientation [5]. The

polarization scrambler placed at the fiber input generates the random input SOP uniformly distributed over the whole Poincaré sphere. If there is no PMD in the transmission fiber link, the Stokes parameters of the output signals satisfy the relation $S_0^2 = S_1^2 + S_2^2 + S_3^2$, and hence DOP=1. All the output SOPs form a perfect ball sphere in Stoke space with unity radius. In the case of first-order PMD in the fiber link, the Stokes parameters satisfy the relation $S_0^2 > S_1^2 + S_2^2 + S_3^2$, and DOP<1, except that when the input SOP is aligned with one of the PSPs (at this point DOP=1). So all the output SOPs form an ellipsoid whose longest radius has a length of unity and point into the direction of the PSPs.

For a system with only a first-order PMD, a DOP ellipsoid whose major axis is aligned with the PSP can be described as follows [5]

$$S_1^2 + \frac{S_2^2 + S_3^2}{R^2(\tau)} = 1 \tag{2}$$

where τ is the DGD and $R(\tau)$ is the normalized autocorrelation of the signal. Also its longest axis points along the S_1 axis. For general case, first-order PMD with PSP pointing in an arbitrary direction, have an ellipsoid with unity value for the longest radius ($r_{max}=1$) and the same value of $R(\tau)$ for the shortest radius ($r_{mid} = r_{min}$), and with its longest axis oriented in the corresponding direction in Stokes space, is shown in Fig. 4.

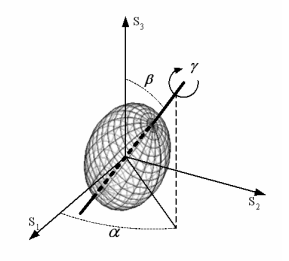


Fig. 4. The orientation of DOP ellipsoid in Stokes space

From the discussion above, we can see that the DOP ellipsoid is a good PMD monitoring signal, from which we can read out the detailed information related to PMD. Thus it can be used in feedback PMD compensation systems.

3 PSO Technique Used to Obtain Real-Time DOP Ellipsoid

The characteristic parameters for a given ellipsoid are its three radii and its orientation angles. We can get only the discrete sampling data of output SOPs from experiments. It is important to find a good algorithm to get the analytical ellipsoid equation from measured discrete data of output SOPs as fast as possible, in order to obtain instantly the parameters of the DOP ellipsoid such as the three radii of the ellipsoid r_{max}, r_{mid}, r_{min} and its orientation angles α, β, γ, as shown in Fig.4. The angles α and β determine

the direction of the longest axis, γ determines the rotation angle around ellipsoid axis itself. This algorithm to be found is required to determine the correct ellipsoid with high precision using the fewest amounts of data, thus ensuring the speed requirement of PMD compensation to be satisfied.

We have previously introduced the particle swarm optimization (PSO) algorithm into automatic PMD compensation as a feedback control algorithm [4]. The PSO algorithm can be described as a procedure aimed at finding the global maximum or global minimum of a function in a multi-dimensional hyperspace by adjusting multi-control parameters, shown mathematically as follows

$$\underset{parameters}{\text{MAX}}\,(function) \tag{3}$$

or

$$\underset{parameters}{\text{MIN}}\,(function) \tag{4}$$

where the number of parameters is the number of dimensions of hyperspace.

In this paper we describe how, by introducing PSO technique into the PMD monitoring unit, we constructed an experiment for obtaining a real-time DOP ellipsoid by using the PSO algorithm in the form of expression (4).

A normal ellipsoid without tilt in the principal axis coordinate $S_1''' - S_2''' - S_3'''$ satisfies

$$\frac{S_1'''^2}{r_1^2} + \frac{S_2'''^2}{r_2^2} + \frac{S_3'''^2}{r_3^2} = 1 \tag{5}$$

In order to obtain a correct DOP ellipsoid, we firstly transform the measured sample data of Stokes parameters S_{1n}, S_{2n}, S_{3n} that constitute a tilt ellipsoid into the S_{1n}''', S_{2n}''', S_{3n}''' that are related to a normal ellipsoid in the principal axis coordinate, by three rotations $-\alpha$, $-\beta$, $-\gamma$. Secondly, we will continually adjust 6 parameters (r_1, r_2, r_3, α, β, γ) until following function in bracket is minimized:

$$\underset{(r_1,r_2,r_3,\alpha,\beta,\gamma)}{\text{MIN}}\left(\sum_{n=1}^{N}\left|\frac{S_{1n}'''^2}{r_1^2} + \frac{S_{2n}'''^2}{r_2^2} + \frac{S_{3n}'''^2}{r_3^2} - 1\right|\right) \tag{6}$$

where N is the number of sampling data points used for ellipsoid data fitting. Once the global minimum of the function in the bracket is found, we obtain the correct ellipsoid parameters (r_1, r_2, r_3, α, β, γ). This is a global minimum searching problem in the 6-dimentional hyperspace. High number of degree of freedom will let us encounter the problem of being trapped in local sub-minima, which will result in the wrong ellipsoid. With the PSO algorithm, we can easily tackle the problem at a fast speed. The detail of the process of implementation of PSO can be found in reference [6].

4 The Experiment and Results

Fig. 5 shows the experimental setup for obtaining the DOP ellipsoid. The laser pulses with linear polarization state were generated by a fiber ring laser. The pulsewidth is

7-ps, and the central wavelength is 1560.5-nm. A polarization controller controlled by the computer was used as a polarization scrambler to randomly transform the SOPs of the laser signal to be uniformly distributed on entire Poincaré sphere, by randomly adjusting the three cells of the polarization controller. A computer controlled air gap time delay line was used as the first-order PMD emulator. A polarimeter detected the SOPs of the output optical signals, by getting the Stokes parameters S_0, S_1, S_2, S_3, and fed them into the computer through a 4-channel A/D. In the computer, the program with PSO algorithm performed the data fitting from sampled data, in order to obtain the DOP ellipsoid and hence the PMD information.

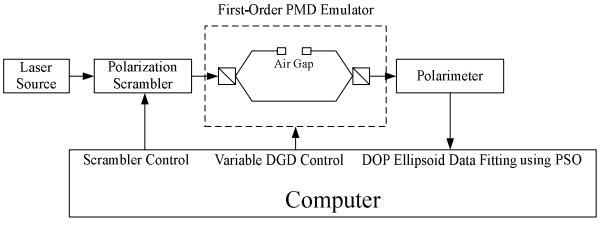
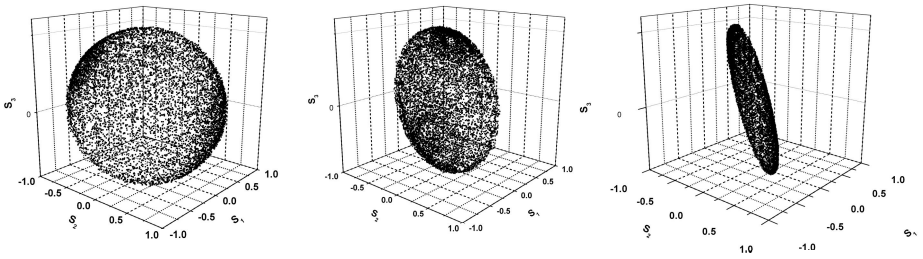


Fig. 5. Experiment scheme for DOP ellipsoid collection

Fig. 6 shows the 8000 sampling point graphs demonstrating the SOPs of the output optical signals obtained in our experiment, with various DGD values. These graphs just show the 8000 discrete output SOPs without any data fitting. It can be seen that, all the output SOPs form a ball sphere in Stokes space with zero DGD (Fig.6(a)), and a needle-like spheroid is formed with larger DGD (Fig.6(c)). The larger DGD is, the smaller value of the minimum radius of the ellipsoid is. Furthermore, according to the PMD theory, the PSPs of the PMD vector do not change for the first-order case. This accords with the results in Fig. 6(b) and Fig. 6(c) which show that the longest radius of ellipsoid remains oriented in the same direction with various DGDs.

Generally, the greater the size of the sampling data, the more accurate the data fitting ellipsoid is. But getting 8000 sampling data points combined with the 6-dimentional adjusting parameter data fitting would be too time consuming using the hardware in our experiment (such as the scrambler, the 4-channel A/D, computer, etc), which is not acceptable for real-time PMD compensation. Just getting 8000 sampling points without any ellipsoid data fitting already consumed the time of 3815-ms in the experiment. In order to get the right PMD information at a instant time, it is necessary to obtain a precise analytical ellipsoid equation to get the three radii r_{max}, r_{mid}, r_{min} and the orientation angles α, β, γ through data fitting from as little data sampling as possible. In this experiment, we made the data fitting from only 100 data samples using the PSO algorithm, and obtained the DOP ellipsoid shown in Fig. 7.



(a) DGD=0ps (b) DGD=4ps (c) DGD=10ps

Fig. 6. 8000 sampling points of output SOPs with various DGD values

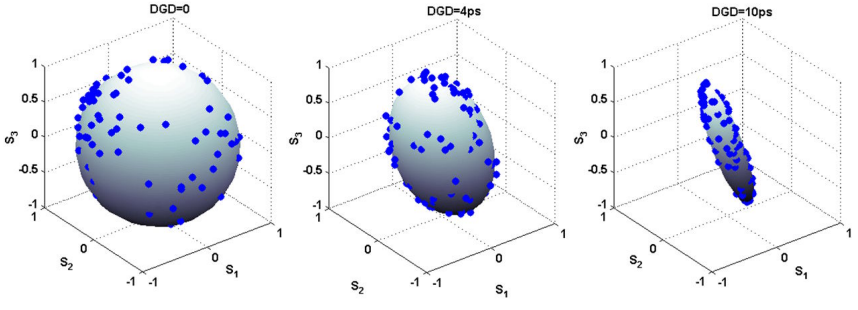


Fig. 7. Data fitting ellipsoid from 100 sampling SOP data using PSO algorithm

Comparing Fig. 6 and Fig. 7, we can see that, the data fitting ellipsoids with 100 samplings nearly fit the ellipsoids formed by 8000 sampling points with high precision. It can be concluded that the PSO algorithm is a powerful one to do the job of data fitting with multi-adjusting parameters, without being trapped in local sub-optima. The whole procedure of getting 100 sampling points and ellipsoid data fitting using the PSO algorithm consumed less than 250-ms. The time can be shorter when using higher speed hardware.

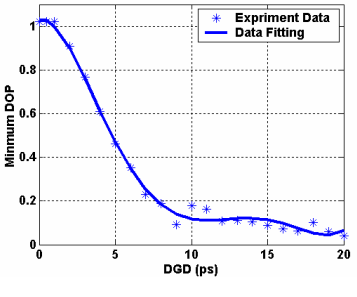


Fig. 8. The length of radius $r_{\min}$ vs. DGD

We also recorded the relationship between the shortest radius of ellipsoid r_{min} (minimum DOP) and DGD as shown in Fig. 8. It can be seen that r_{min} decreases nearly monotonically as DGD increases. So it can be used as the feedback signal for PMD compensation.

5 Conclusion

Using the PSO algorithm, we have realized an experiment for determination of the DOP ellipsoid by data fitting using a small number of sampled SOP data. By using the PSO algorithm we can determine the DOP ellipsoid from only 100 samplings with high precision. The 100 data sampling and fitting were achieved within 250ms. The experiment showed that minimum DOP of the ellipsoid decreased monotonically with DGD, and the direction of ellipsoid orientation was unchanged for the first-order PMD, in accordance with the fact that the PSPs of the PMD vector remain unchanged.

Acknowledgement. This work was supported partly by the National Natural Science Foundation of China (No. 60577046), and Corporative Building Project of Beijing Educational Committee (No. XK100130637)**.**

References

1. Noé, R., Sandel, D., Yoshida-Dierolf, M., Hinz, S., Mirvoda, V., Schöpflin, A., Glingener, C., Gottwald, E., Scheerer, C., Fischer, G. Weyrauch, T., Haase, W.: Polarization Mode Dispersion Compensation at 10, 20, and 40Gb/s with Various Optical Equaliziers. J. Lightwave Technol. 17 (1999) 1602-1616
2. Rosenfeldt, H., Knothe, Ch., R. Ulrich et al.: Automatic PMD compensation at 40Gbit/s and 80Gbit/s using a 3-dimentional DOP evaluation for feedback. OFC2001, (2001) Postdeadline paper PD27-1
3. Kogelnik, H., Jopson, R. M., Nelson, L.: Polarization-Mode Dispersion, In: Kaminow, I. P., Li, T. (eds): Optical Fiber Telecommunications, IV B. Academic Press, San Diego San Francisco New York Boston London Sydney Tokyo, (2002) 725-861
4. Zhang, X., Yu, L., Zheng, Y., Shen, Y., Zhou, G., Chen, L., Xi, L., Yuan, T., Zhang, J., Yang, B.: Two-Stage Adaptive PMD Compensation in 40Gb/s OTDM Optical Communication System Using PSO Algorithm. Opt. Quantum Electron. 36 (2004) 1089-1104
5. Zheng, Y., Zhang, X., Chen, L., Yang, B.: Analysis of degree of polarization ellipsoid as feedback signal for polarization mode dispersion compensation in NRZ, RZ and CS-RZ systems. Opt. Commun. 234 (2004) 107-117
6. Kennedy, J., Eberhart, R. C.: Paticle Swarm Optimization. In: Proc. of IEEE International Conference on Neural Networks. Piscataway, NJ, USA, (1995) 1942-1948

Fast Drug Scheduling Optimization Approach for Cancer Chemotherapy

Yong Liang[1], Kwong-Sak Leung[2], and Tony Shu Kam Mok[3]

[1] Department of Computer Science, Shantou University, China
[2] Department of Computer Science & Engineering,
The Chinese University of Hong Kong, HK
[3] Department of Clinical Oncology,
The Chinese University of Hong Kong, HK
yliang@stu.edu.cn, ksleung@cse.cuhk.edu.hk, tony@clo.cuhk.edu.hk

Abstract. In this paper, we propose a novel fast evolutionary algorithm — cycle-wise genetic algorithm (CWGA) based on the theoretical analyses of a drug scheduling mathematical model for cancer chemotherapy. CWGA is more efficient than other existing algorithms to solve the drug scheduling optimization problem. Moreover, its simulation results match well with the clinical treatment experience, and can provide much more drug scheduling policies for a doctor to choose depending on the particular conditions of the patients. CWGA also can be widely used to solve other kinds of the real dynamic systems.

Keywords: Genetic algorithm, Drug scheduling model.

1 Introduction

An important target for cancer chemotherapy is to maximally kill tumor cells for a fixed treatment period. So drug scheduling is essential in cancer chemotherapy. Martin [7] have proposed the optimal drug scheduling model by the following differential equations:

$$\frac{dx_1}{dt} = -\lambda x_1 + k(x_2 - \beta)H(x_2 - \beta) \tag{1}$$

$$\frac{dx_2}{dt} = u - \gamma x_2 \tag{2}$$

$$\frac{dx_3}{dt} = x_2 \tag{3}$$

with the initial state $x^T(0) = [ln(100),\ 0,\ 0]$, the parameters $\lambda = 9.9 \times 10^{-4}$, $k = 8.4 \times 10^{-3}$, $\beta = 10$, $\gamma = 0.27$ and:

$$H(x_2 - \beta) = \begin{cases} 1, & \text{if} \quad x_2 \geq \beta \\ 0, & \text{if} \quad x_2 \leq \beta \end{cases} \tag{4}$$

where x_1 is a transformed variable that is inversely related to the mass of the tumor. The tumor mass is given by $N = 10^{12} \times exp(-x_1)$ cells, and the initial

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 1099–1107, 2007.
© Springer-Verlag Berlin Heidelberg 2007

tumor cell population is set at 10^{10} cells [7]. The variable x_2 is the drug concentration in the body in drug units (D) and x_3 is the cumulative drug toxicity in the body. Equation (1) describes the net change in the tumor cell population per unit time. The first term on the right-hand side of Equation (1) describes the increase in cells due to cell proliferation and the second term describes the decrease in cells due to the drug. The parameter λ is a positive constant related to the growth speed of the cancer cells, and k is the proportion of tumor cells killed per unit time per unit drug concentration which is assumed to be a positive constant. The implication of the function described in Equation (4) is that there is a threshold drug concentration level, β below which the number of the killed tumor cells is smaller than the number of the reproduced tumor cells, and the drug is not efficient. Equation (2) describes the net increase in the drug concentration at the cancer site. The variable u is the rate of delivery of the drug, and the half-life of the drug is $ln(2)/\gamma$, where γ is the biochemical character parameter of the drug. It is assumed that the drug is delivered by infusion, and there is an instantaneous mixing of the drug with plasma, as well as an immediate delivery of the drug to the cancer site. These assumptions represent approximations based on the relative amount of time. It takes for the aforementioned activities to occur with respect to the total amount of time over which the treatment is administered. Equation (3) relates the cumulative drug toxicity to the drug concentration, e.g., the cumulative effect is the integral of the drug concentration over the period of exposure.

The performance index [7] to be maximized is:

$$I = x_1(t_f) \tag{5}$$

where the final time $t_f = 84$ days. The control optimization is performed subject to constraints on the drug delivery: $u \geq 0$, and on the state variables: $x_2 \leq 50$, $x_3 \leq 2.1 \times 10^3$.

Cancer chemotherapy is a systemic treatment, so the action of the chemotherapeutic agent is not restricted to the tumor site. Any of the body organs are liable to injury. This is on contrast to the localized treatments, such as surgery or radiotherapy. Therefore, the constraints on the drug concentration x_2 and the cumulative drug toxicity x_3 are to ensure that the patient can tolerate the toxic side effects of the drug. Drug resistance is considered to be a significant factor in chemotherapeutic failure [8] and it has been shown that the drug resistant cells are likely to increase as the tumor burden increases [6]. In order to reduce the likelihood of the emergence of drug resistant cells, the tumor size is forced to reduce by at least 50% every 3 weeks, so that: $x_1(21) \geq ln(200)$, $x_1(42) \geq ln(400)$, $x_1(63) \geq ln(800)$.

Many researchers have applied different optimization methods to improve the results of the drug scheduling model [1] [3] [4] [6] [7] [8] [10] [11]. Through analyzing the experimental results from the existing model, there are two obvious unreasonable outcomes in the optimal drug scheduling policies: (i) unreasonable timing for the first treatment; and (ii) three point constraints cannot improve the efficiency of the cancer treatment. To solve the unreasonable problems and

accurately describe the time course of the cumulative drug toxicity x_3 in the body, in [5], we have modified the third equation in Martin's model as follows:

$$\frac{dx_3}{dt} = x_2 - \eta_L \times x_3(1 - \frac{x_3}{2\theta}) - \eta_G \times x_3 \ln(\frac{2\theta}{x_3}) - \eta_E \times x_3, \tag{6}$$

where η_L, η_G, η_E are nonnegative constants. We set the constraint of the cumulative drug toxicity to $x_3 < \theta$.

Equation (6) describes the net change of the cumulative drug toxicity x_3 per unit time. In the right-hand side of this equation, the first term x_2 describes the increase of the cumulative drug toxicity x_3 due to the drug concentration x_2. The second, third and fourth terms can be called logistic, Gompertz and exponential drug toxicity clearance functions respectively, which describe the decreases in the drug toxicity due to the clearance in the body. As you know, the liver and kidney are the primary detoxification and elimination organs — they eliminate the drug toxicity from our body. Drugs passing through the liver are eliminated in a chemically altered (metabolized) form in the bile. Whenever, drug metabolism or movement across the liver involves an active process, then the likelihood of saturable kinetics exists. Thus, the logistic and Gompertz functions give the fit for saturable toxic elimination processes in the liver. On the other hand, drugs, particularly water-soluble drugs and their metabolites, are also eliminated by the kidney in urine. The kidney is a filter that cleanses toxins from our blood and its ability to excrete the toxic compounds depends on the amount of the drug toxicity in the bloodstream. Therefore, the renal excretion (i.e., filtration and passive reabsorption) may be best considered to the nonsaturable mechanisms and described by the exponential function. We combine Equation (6) with Equations (1) and (2) to construct a renewed drug scheduling model. It matches well with the clinical treatment knowledge and experience.

The rest of this paper is organized as follows. Section 2 analyzes the optimal solutions obtained by the renewed drug scheduling model. Section 3 presents the theoretical analysis of the drug scheduling model. Section 4 describes a fast evolutionary algorithm — cycle-wise genetic algorithm (CWGA) to solve the drug scheduling problem. Section 5 provides the simulation results under the new optimization approach. The paper conclusion and future work are summarized in Section 6.

2 Analysis of the Experimental Results of the Renewed Model

In our previous work, we use our proposed adaptive elitist-population based genetic algorithm (AEGA), which is an efficient multimodal optimization algorithm [2], to implement the optimization of the drug scheduling model for exploring multiply efficient drug scheduling policies. In the drug scheduling problem as described in Section 1, there are 84 control variables to be optimized, which represent the dosage levels for the 84 days. The drug scheduling model is a high dimensional and multimodal optimization problem. Here we have proposed a cycle-wise variable representation to accurately and efficiently describe the drug scheduling policy.

The cycle-wise variable representation includes two parts: a front and a cyclic parts. For example, the cycle-wise variable representation is $\{57.05, 13.5, 0, 10.8, 80 \times \overline{21.5}\}$ that means giving the drug doses 57.05, 13.5, 0, 10.8 at the first four days respectively, and then repeated 80 times giving the drug dose 21.5 every two days. Its front part is $\{57.05, 13.5, 0, 10.8\}$. Its cyclic part is $80 \times \overline{21.5}$, which consists of the number of cycles 80 and the repetend $\overline{21.5}$. The cycle-wise variable representation is very suitable for the drug scheduling problem. Because in the first few days of the treatment period, the patient's body may not have adapted to the drug, but it is important to kill as much tumor cells as possible, the drug doses will be adjusted day by day. We use the front part to represent the drug doses in this initial period. Then when the patient's body gradually gets used to the drug, the drug administration schedule will follow a fixed cycle and a fixed dose pattern, which is suitably represented by the cycle-part.

Fig.1-(a) and (b) show the multiple optimal solutions, which are explored by AEGA from our renewed drug scheduling model. We also used other optimization algorithm — iterative dynamic programming (IDP) without the cycle-wise variable representation [6] to solve the renewed model. The global optimal solution obtained by IDP is as same as Fig.1-(a) shown.

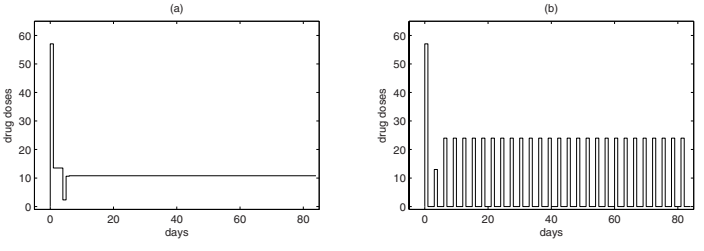


Fig. 1. The multiple optimal solutions obtained by the renewed drug scheduling model ($\eta_L = 0, \eta_G = 0, \eta_E = 0.4$). (a): global optimal solution; (b): second optimal solution.

We believe that the cycle-wise variable representation is suitable and consistent with the renewed drug scheduling model. So we do the theoretical analysis of this model in the next section. The results show that the cycle-wise variable representation can be used to predict the treatment results of the drug and significantly reduce the computational complexity of the optimization algorithm. Moreover, it is more consistent with the clinical experience.

3 Theoretical Analysis of the Renewed Drug Scheduling Model

In the renewed drug scheduling model, Equation (6) is a higher order differential equation. We can write it as a system with a very simple change of variable. We define the following three new functions.

$$y_1 = \eta_L \times x_3(1 - \frac{x_3}{2\theta}) \tag{7}$$

$$y_2 = \eta_G \times x_3 \ln(\frac{2\theta}{x_3}) \tag{8}$$

$$y_3 = \eta_E \times x_3 \tag{9}$$

Equation (6) can then be wrote the following system of differential equations.

$$\frac{dy_i}{dt} = \tau_i \times x_2 - y_i, \quad i = 1, 2, 3 \tag{10}$$

We combine Equations (1) and (2) with the differential equations (10) instead of Equation (6). The renewed drug scheduling model can then be considered as the first order, linear and homogeneous systems of differential equations.

Theorem 1. In the drug scheduling model, suppose the giving drug doses u is a cycle-wise variable with a cycle Δt (giving drug at t_0, $t_0+\Delta t$, $t_0+2\Delta t$,..., days), then the maximal values of the drug concentration x_2 and the drug cumulative toxicity $y_1+y_2+y_3$ are approximately equal in each Δt respectively (e.g., $x_2(t_0) \approx x_2(t_0 + \Delta t) \approx x_2(t_0 + 2\Delta t) \ldots$).

Proof : Suppose giving drug at t_0, $t_0 + \Delta t$, $t_0 + 2\Delta t$,..., days in the cycle part, then the drug concentration x_2 or the drug cumulative toxicity $y_1 + y_2 + y_3$ will meet its constraint at these days. If supposing that x_2 meets its constraint, this means that x_2 monotonously increases and reaches its maximal value at times t_0, $t_0 + \Delta t$, $t_0 + 2\Delta t$,..., then,

$$\frac{dx_2}{dt}(t_0) \approx \frac{dx_2}{dt}(t_0 + \Delta t) \approx \frac{dx_2}{dt}(t_0 + 2\Delta t) \approx 0 \tag{11}$$

$$\begin{aligned} u(t_0) - \lambda x_2(t_0) &\approx u(t_0 + \Delta t) - \lambda x_2(t_0 + \Delta t) \\ &\approx u(t_0 + 2\Delta t) - \lambda x_2(t_0 + 2\Delta t) \end{aligned}$$

Since $u(t_0) = u(t_0 + \Delta t) = u(t_0 + 2\Delta t)$, then

$$x_2(t_0) \approx x_2(t_0 + \Delta t) \approx x_2(t_0 + 2\Delta t) \tag{12}$$

In the same way, we can get the same results of the drug toxicity $y_1 + y_2 + y_3$ as the drug concentration x_2. ♢

In the drug scheduling model, there is not the constraint for the number of cancer cells x_1. In equation (1), x_2 must be larger than β for any efficient drug scheduling policy, then $k(x_2 - \beta) \gg \lambda x_1$. We can evaluate the number of cancer cells at the end of the treatment period using the formula $x_1(t_0 + k \times \Delta t) = x_1(t_0) + k \times (x_1(t_0 + \Delta t) - x_1(t_0))$. Based on the theoretical analysis of the drug scheduling model, we can only calculate the values of x_1, x_2 and $y_1 + y_2 + y_3$ in front part and first two cycles of the cycle part, then using them evaluate the values of x_1, x_2 and $y_1 + y_2 + y_3$ at the end of the treatment period. Thus, the computational complexity of the algorithm will be significantly reduced to solve the drug scheduling model.

4 Optimization of Drug Scheduling Model Via CWGA

In this section we propose a new genetic algorithm — a cycle-wise genetic algorithm (CWGA) to implement the optimization of the drug scheduling model for exploring multiply efficient drug scheduling policies.

In order to successfully explore multiple optimal solutions of the drug scheduling model, several rules for applying the CWGA based on our proposed adaptive elitist- population search technique [2] are made as follows:

- Use the cycle-wise representation to keep the scheduling freedom and improve the efficiencies of algorithm.
- Use the front part and the first two cycles of the cycle-wise representation to evaluate the fitness of the drug scheduling policy to significantly reduce the computational complexity of the algorithm.
- Use the adaptive elitist-population search technique in the crossover operator to reduce the redundancy of the population.
- Use the adaptive elitist-population search technique in the mutation operator to increase the diversity of the population.
- Adaptively adjust the population size to optimally use the elitist individuals to explore multiple optima.

Following these rules, the CWGA for the drug scheduling model is implemented as follows:

1. Set $t = 0$ and initialize a chromosome population $P(t)$ (uniform random initialization within the bounds);
2. Evaluate $P(t)$ using the fitness measure;
3. **While** (termination condition not satisfied) **Do**
   (a) Elitist crossover operation to generate $P(t+1)$;
      i. check the dissimilarity of the randomly selected parents p_i and p_j;
      ii. **if** the parents p_i and p_j are similar, the elitist operation conserves the better one of them for the next generation; **else**, according to multi-point crossover operation, generates 6 offspring, and selects the better two from the parents and their offspring to the next generation;
   (b) Elitist mutation operation to generate $P(t+1)$;
      i. according to the one-point mutation operation, generate the offspring c_i from the parent p_i;
      ii. **if** p_i and c_i are dissimilar, the elitist operation conserves p_i and c_i together for the next generation; **else**, selects the better one of them to the next generation;
4. Evaluate $P(t+1)$;
5. Stop if the termination condition is satisfied; otherwise, go to Step 3.

5 Experimental Results and Comparisons

The drug scheduling problem were simulated using the CWGA with the following parameters: initial population size=100; maximal number of generations=10000; crossover rate=1.0; mutation rate=1.0 and the distance threshold σ_s=10. We set the parameters $\eta_L = 0.01, \eta_G = 0.01, \eta_E = 0.38$ in the model. The drug scheduling model was simulated using numerical differentiation method of Runge-Kutta [9], with a small time interval of 0.1 day for good accuracy.

Optimizing the renewed drug scheduling model via CWGA for 50 times can consistently obtain 6 most efficient drug scheduling policies. These results are listed in Table 1. For example, Fig. 2 and 3 show the control variable u, the number of cancer cells (inversely related to the best performance index x_1), the change processes of the drug concentration x_2 and the cumulative drug toxicity x_3 of the first and sixth optimal policies. The 6 best results all satisfy the three point constraints, and therefore it is not necessary to find the special solutions for the new model with the three point constraints separately.

Table 1. The most efficient drug scheduling policies obtained by the renewed model

The most efficient drug scheduling policies	Cancer cells
(1): $\{92.21,(83 \times \overline{10.8})\}$	21
(2): $\{132.69, 0, 22.57, 0, 8.63, 0, 24.52, (38 \times \overline{0, 21.02}), 8.22\}$	32
(3): $\{132.69, (2 \times \overline{0}), 31.5, (2 \times \overline{0}), 24.5, (25 \times \overline{(2 \times \overline{0}), 29.89}), 0, 25.24\}$	68
(4): $\{132.69, (3 \times \overline{0}), 42.6, (3 \times \overline{0}), 34.9, (18 \times \overline{(3 \times \overline{0}), 38.3}), 0, 0, 26.1\}$	113
(5): $\{132.69, (4 \times \overline{0}), 50.1, (4 \times \overline{0}), 46.9, (14 \times \overline{(4 \times \overline{0}), 49.3}), 0, 20.9\}$	376
(6): $\{132.69, (13 \times \overline{(5 \times \overline{0}), 54.24}), 3 \times \overline{0}, 47.53\}$	2125

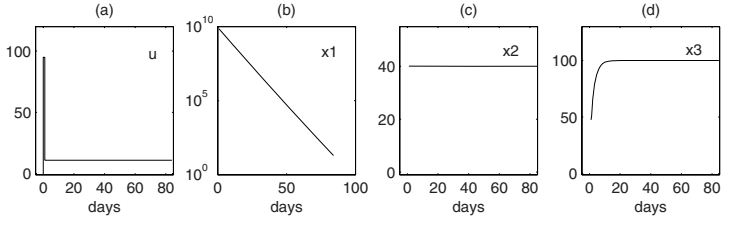


Fig. 2. The first efficient drug scheduling policy under our renewed model

In [5], we use our proposed adaptive elitist-population genetic algorithm (AEGA) to solve the drug scheduling optimization problem. Comparing with AEGA, CWGA can explore same multiple optimal solutions for the renewed model, but its efficiency (CPU time) is at least 14 times higher than that of AEGA because AEGA need to calculate the values of x_1, x_2 and x_3 under the differential equation system in the whole treatment period. We also combine the

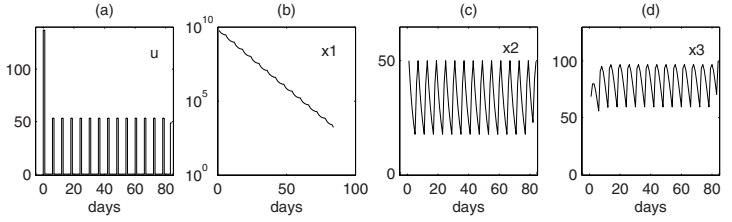


Fig. 3. The sixth efficient drug scheduling policy under our renewed model.

fitness measure of CWGA into iterative dynamic programming (IDP) to solve the renewed drug scheduling problem. IDP can find the global optimal solution and its efficiency is improved almost 9 times.

On the other hand, the multiple efficient drug scheduling policies under the new model match well with the clinical experience. In the clinical treatment, generally the drug scheduling policies include two kinds: continuous and repeated. The drug scheduling policy (1) and the drug scheduling policies (2)-(6) represent these two kinds respectively. In some patients, the aim of treatment may be to reduce the tumor size with minimum toxicity and the drug scheduling policy (6) is suitable because its cumulative drug toxicity is low and often decreases to 60. For other patients, they may be cure despite higher toxicity, the drug scheduling policy (1) is suitable because this policy is most efficient but with high toxicity. So these multiple efficient drug scheduling policies obtained by the new model are more useful. According to the different conditions of the patients, the doctor can select the suitable drug scheduling policy to treat a cancer and get the best efficiency.

6 Conclusion

This paper has presented the novel fast evolutionary algorithm — cycle-wise genetic algorithm (CWGA) based on the theoretical analyses of a drug scheduling mathematical model for cancer chemotherapy. CWGA is more efficient than other existing algorithms to solve the drug scheduling optimization problem. Moreover, its simulation results match well with the clinical treatment experience, and can provide much more drug scheduling policies for a doctor to choose depending on the particular conditions of the patients. CWGA also can be widely used to solve other kinds of the real dynamic systems.

Acknowledgment

This research was partially supported by RGC Earmarked Grant 4173/04E of Hong Kong SAR and Research Grant Direct Allocation of the Shantou University.

References

1. Carrasco, E.F., Banga, J.R.: Dynamic Optimization of Batch Reactors Using Adaptive Stochastic Algorithms. Ind. Eng. Chem. Res. **36** (1997) 2252-2261
2. Leung, K.S., Liang, Y: Genetic Algorithm with Adaptive Elitist-population Strategies for Multimodal Function Optimization. Proc. of Int. Conf. GECCO-2003. (2003) 1160-1171
3. Liang, Y., Leung, K. S., Mok, S. K., A Novel Evolutionary Drug Scheduling Model in Cancer Chemotherapy, IEEE Transactions on Information Technology in BioMedicine, **10**(2), (2006) 237-245
4. Liang, Y., Leung, K. S., Mok S. K., Optimal Control of a Cancer Chemotherapy Problem with Different Toxic Elimination Processes, Proc. of IEEE WCCI. (2006) 8644-8651
5. Liang, Y., Leung, K.S., Mok S.K., Automating the Drug Scheduling with Different Toxicity Metabolism in Cancer Chemotherapy via Evolutionary Computation, Proc. of ACM GECCO-2006. (2006) 1705-1712
6. Luus, R., Harting, F., Keil, F.J.: Optimal Drug Scheduling of Cancer Chemotherapy by Direct Search Optimization. Hung. J. Ian. Chen. **23** (1995) 55-58
7. Martin, R.B.: Optimal Control Drug Scheduling of Cancer Chemotherapy. Automatica. **28** (1992) 1113-1123
8. Murray, J.M.: Optimal Control of a Cancer Chemotherapy Problem with General Growth and Loss Functions. Math. Biosci. **38** (1990) 273-287
9. Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P.: Numerical Recipes in C: he Art of Scientific Computing. 2nd edn. Cambridge University Press (1992)
10. Tan, K.C., Khor, E.F., Cai, J., Heng, C.M., Lee, T.H.: Automating the Drug Scheduling of Cancer Chemotherapy via Evolutionary Computation. Artificial Intelligence in Medicine. **25** (2002) 169-185
11. Wheldon, T.E.: Mathematical Models in Cancer Research. Bristol. Adam Hilger. (1998)

Optimization of IG-Based Fuzzy System with the Aid of GAs and Its Application to Software Process

Sung-Kwun Oh[1], Keon-Jun Park[1], and Witold Pedrycz[2]

[1] Department of Electrical Engineering, The University of Suwon, San 2-2 Wau-ri, Bongdam-eup, Hwaseong-si, Gyeonggi-do, 445-743, South Korea
ohsk@suwon.ac.kr
[2] Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB T6G 2G6, Canada and Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland

Abstract. We introduce an optimization of information granules (IG)-based fuzzy model with the aid of genetic algorithms (GAs) to describe projects in terms of complexity and development time in experimental software datasets. The proposed fuzzy model implements system structure and parameter identification with the aid of IG and GAs. To identify the structure and the parameters of fuzzy model we use genetic algorithms. The concept of information granulation was coped with to enhance the abilities of structural optimization of fuzzy model. Granulation of information realized with Hard C-Means clustering help determine the initial parameters of fuzzy model such as the initial apexes of the membership functions in the premise part and the initial values of polynomial functions in the consequence part of the fuzzy rules. And the initial parameters are tuned effectively with the aid of the GAs and the least square method. An aggregate objective function is constructed in order to strike a sound balance between the approximation and generalization capabilities of the fuzzy model. The experimental results include well-known software data such as medical imaging system (MIS).

1 Introduction

Fuzzy modeling has been studied to deal with complex, ill-defined, and uncertain systems in many other avenues. The researches on the process have been exploited for a long time. Linguistic modeling [2] and fuzzy relation equation-based approach [3] were proposed as primordial identification methods for fuzzy models. The general class of Sugeno-Takagi models [4] gave rise to more sophisticated rule-based systems where the rules come with conclusions forming local regression models. While appealing with respect to the basic topology (a modular fuzzy model composed of a series of rules) [5] these models still await formal solutions as far as the structure optimization of the model is concerned, say a construction of the underlying fuzzy sets—information granules being viewed as basic building blocks of any fuzzy model. Some enhancements to the model have been proposed by Oh and Pedrycz [6], yet the problem of finding "good" initial parameters of the fuzzy sets in the rules remains open.

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 1108–1115, 2007.
© Springer-Verlag Berlin Heidelberg 2007

This study concentrates on the central problem of fuzzy modeling that is a development of information granules-fuzzy sets. Taking into consideration the essence of the granulation process, we propose to cast the problem in the setting of clustering techniques and genetic algorithms. Information granulation with the aid of C-Means clustering help determine the initial parameters of fuzzy model such as the initial apexes of the membership functions in the premise part and the initial values of polynomial function in the consequence part. And the initial parameters are tuned (adjusted) effectively by means of the genetic algorithms and the least square method. An aggregate objective function with some weighting factor is proposed so that we could achieve a sound balance between accuracy and generalization abilities of the fuzzy model. The model is applied to the medical imaging system (MIS) being widely used in quantitative software engineering.

2 Information Granules

Roughly speaking, information granules (IG) [7], [8] are viewed as related collections of objects (data point, in particular) drawn together by the criteria of proximity, similarity, or functionality. Granulation of information is an inherent and omnipresent activity of human beings carried out with intent of gaining a better insight into a problem under consideration and arriving at its efficient solution. In particular, granulation of information is aimed at transforming the problem at hand into several smaller and therefore manageable tasks. In this way, we partition this problem into a series of well-defined subproblems (modules) of a far lower computational complexity than the original one. The form of information granulation themselves becomes an important design feature of the fuzzy model, which are geared toward capturing relationships between information granules.

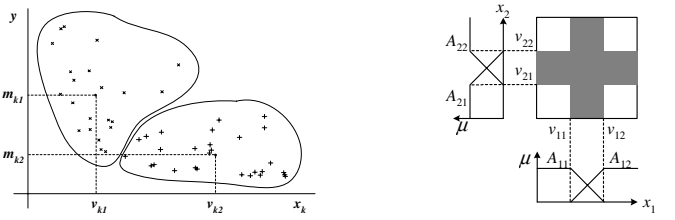
Clustering is often regarded as a synonym of information granulation. The intent of clustering is to find a structure in the data and reveal clusters – information granules in the data set. The clustering algorithms have been used extensively not only to organize and categorize data, but it becomes useful in data compression and model construction. The C-Means clustering [9] has been applied to a variety of areas, including image and speech data compression, data preprocessing of system modeling.

3 IG-Based Fuzzy Inference System (IG_FIS)

The identification procedure for fuzzy models is usually split into the identification activities dealing with the premise and consequence parts of the rules. The identification completed at the premise level consists of two main steps. First, we select the input variables x_1, x_2, …, x_k of the rules. Second, we form fuzzy partitions of the spaces over which these individual variables are defined. The identification of the consequence part of the rules embraces two phases, namely 1) a selection of the consequence variables of the fuzzy rules, and 2) determination of the parameters of the consequence (conclusion part). And the least square error method used at the parametric optimization of the consequence parts of the successive rules.

3.1 Premise Identification

In the premise part of the rules, we confine ourselves to a triangular type of membership functions whose parameters are subject to some optimization. The C-Means clustering helps us organize the data into cluster so in this way we capture the characteristics of the experimental data. In the regions where some clusters of data have been identified, we end up with some fuzzy sets that help reflect the specificity of the data set. In the sequel, the modal values of the clusters are refined (optimized) using genetic optimization, and genetic algorithms (GAs), in particular.



(a) Clusters formed by C-Means clustering (b) Fuzzy partition and resulting MFs

Fig. 1. Identification of the premise part of the rules of the system

The identification of the premise part is completed in the following manner.

Given is a set of data $\mathbf{U}=\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k ; \mathbf{y}\}$, where $\mathbf{x}_k =[x_{1k}, \ldots, x_{mk}]^T$, $\mathbf{y} =[y_1, \ldots, y_m]^T$, k is the number of variables and , m is the number of data.

[Step 1] Arrange a set of data **U** into data set $\mathbf{X}_k$ composed of respective input data and output data.

$$\mathbf{X}_k=[\mathbf{x}_k ; \mathbf{y}] \tag{1}$$

[Step 2] Determine the centers (prototypes) $\mathbf{v}_{kg}$ with data set $\mathbf{X}_k$ using C-Means clustering algorithms

[Step 2-1] Categorize data set $\mathbf{X}_k$ into c-clusters (in essence this is effectively the granulation of information)

[Step 2-2] Calculate the center vectors $\mathbf{v}_{kg}$ of each cluster.

$$\mathbf{v}_{kg} = \{v_{k1}, v_{k2}, \ldots, v_{kc}\} \tag{2}$$

[Step 3] Partition the corresponding input space using the prototypes of the clusters $\mathbf{v}_{kg}$. Associate each cluster with some meaning (semantics), say Small, Big, etc.

[Step 4] Set the initial apexes of the membership functions using the prototypes $\mathbf{v}_{kg}$.

3.2 Consequence Identification

We can also identify the structure of the consequence parts of rules by considering the initial values of polynomial functions based upon the information granulation.

[Step 1] Find a set of data included in the fuzzy space of the j-th rule.

[Step 2] Compute the prototypes $\mathbf{V}_j$ of the data set by taking the arithmetic mean of each rule.

$$\mathbf{V}_j = \{V_{1j}, V_{2j}, \ldots, V_{kj}; M_j\} \tag{3}$$

[Step 3] Set the initial values of polynomial functions with the center vectors $\mathbf{V}_j$.

The identification of the conclusion parts of the rules deals with a selection of their structure that is followed by the determination of the respective parameters of the local functions occurring there.
The conclusion is expressed as follows.

$$R^j : If\ x_1\ is\ A_{1c}\ and \cdots and\ x_k\ is\ A_{kc}\ then\ y_j - M_j = f_j(x_1, \cdots, x_k) \tag{4}$$

Type 1 (Simplified Inference): $f_j = a_{j0}$

Type 2 (Linear Inference): $f_j = a_{j0} + a_{j1}(x_1 - V_{j1}) + \cdots + a_{jk}(x_k - V_{jk})$ (5)

Type 3 (Quadratic Inference):
Type 4 (Modified Quadratic Inference):

The calculations of the numeric output of the model, based on the activation (matching) levels of the rules there, rely on the expression

$$y^* = \frac{\sum_{j=1}^{n} w_{ji} y_i}{\sum_{j=1}^{n} w_{ji}} = \frac{\sum_{j=1}^{n} w_{ji}(f_j(x_1, \cdots, x_k) + M_j)}{\sum_{j=1}^{n} w_{ji}} = \sum_{j=1}^{n} \hat{w}_{ji}(f_j(x_1, \cdots, x_k) + M_j) \tag{6}$$

If the input variables of the premise and parameters are given in consequence parameter identification, the optimal consequence parameters that minimize the assumed performance index can be determined. In what follows, we define the performance index as the mean squared error (MSE).

$$\mathrm{PI} = \frac{1}{m} \sum_{i=1}^{m} (y_i - y_i^*)^2 \tag{7}$$

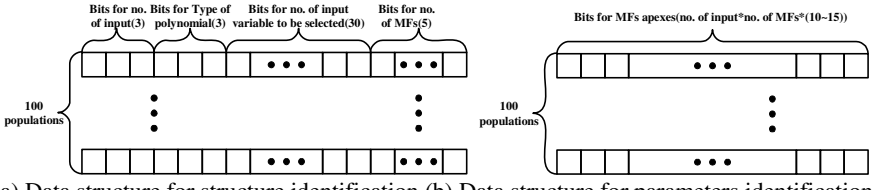
4 Optimization of IG-Based FIS

4.1 Genetic Algorithms

It has been demonstrated that genetic algorithms [10] are useful in a global optimization of such problems given their ability to efficiently use historical information to produce new and improved solutions. GAs are shown to support robust search in complex search spaces. In particular, they are stochastic and less likely to get trapped in local minima as we can witness quite often when dealing with gradient-descent techniques. GAs are population-based optimization techniques. The search of the solution space is completed with the aid of several genetic operators. There are three generic genetic operators such as reproduction, crossover, and mutation. Reproduction is a process in which the mating pool for the next generation is chosen. Individual strings are copied into the mating pool according to their fitness function values. Crossover usually proceeds in two steps. First, members from the mating pool

are mated at random. Second, each pair of strings undergoes crossover as follows: a position l along the string is selected uniformly at random from the interval [1, l-1], where l is the length of the string. Swapping all characters between the positions k and l creates two new strings. Mutation is a random alteration of the value of a string position. In a binary coding, mutation means changing a zero to a one or vice versa. Mutation occurs with small probability. Those operators, combined with the proper definition of the fitness function, constitute the main body of the genetic computing.

In this study, in order to identify the fuzzy model we determine such a structure as the number of input variables, input variables being selected and the number of the membership functions standing in the premise part and the order of polynomial (Type) in conclusion. The membership parameters of the premise are genetically optimized. Figure 2 shows an arrangement of the content of the string to be used in genetic optimization. Here, parentheses denote the number of chromosomes allocated to each parameter.



(a) Data structure for structure identification (b) Data structure for parameters identification

Fig. 2. Data structure of genetic algorithms used for the optimization of fuzzy model

For the optimization of the fuzzy model, genetic algorithms use a binary bit string, roulette-wheel in the selection operator, one-point crossover in the crossover operator, and invert in the mutation operator. We also apply elitism to keep the best individual across generations. Here, we use 150 generations and run the GA of a size of 100 individuals for structure identification. For parameter estimation, GA was run for 300 generations and the population was of size 100. We set up the crossover rate and mutation probability to be equal to 0.65, and 0.1, respectively.

4.2 Objective Function with Weighting Factor

The objective function (performance index) is a basic mechanism guiding the evolutionary search carried out in the solution space. The objective function includes both the training data and testing data and comes as a convex combination of the two components.

$$f(PI,\ E_PI) = \theta \times PI + (1 - \theta) \times E_PI \tag{8}$$

Here, PI and E_PI denote the performance index for the training data and testing (validation) data, respectively. θ is a weighting factor that allows us to form a sound balance between the performance of the model for the training and testing data. Depending upon the values of the weighting factor, several specific cases of the objective function are worth distinguishing.

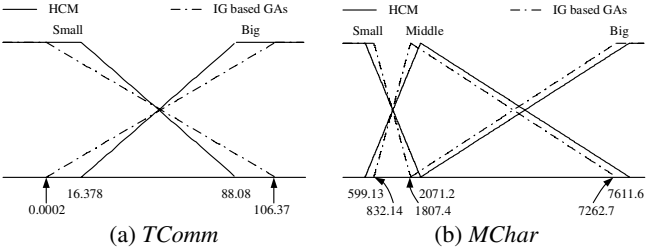
5 Experimental Studies

In this section, we consider a medical imaging system [11] subset of 390 software modules written in Pascal and FORTRAN for modeling. These modules consist of approximately 40,000 lines of code. To design an optimal model from the MIS, we study 11 system input variables such as *LOC*, *CL*, *TChar*, *TComm*, *MChar*, *DChar*, *N*, $\hat{N}$, N_F, $V(G)$ and *BW*. The output variable of the model is the number of changes *Changes* made to the software module during its development. When applying any modeling technique, an assessment of predictive quality is important. Data splitting is a modeling technique that is often applied to test predictive quality. Applying this technique, one randomly partitions the data set to produce two data sets. The first 60% data set is used for fitting the models. The remaining 40% data set, the testing data set, provides for quantifying the predictive quality of the fitted models. We carried out the structure and parameters identification on a basis of the experimental data using GAs to design IG-based fuzzy model. Table 1 summarizes the performance index for IG-based fuzzy model.

Table 1. Performance index of IG-based fuzzy model

Model	θ	Identification	Input variable	No. Of MFs	Type	M_PI	PI	E_PI
IG_FIS	0.0	Structure	*Tcomm* $\hat{N}$ *V*(*G*)	2x3x2	Type 1	33.867	54.841	33.867
		Parameters				26.836	63.005	26.836
	0.5	Structure	*TComm* *MChar*	2x3	Type 3	35.137	29.957	40.316
		Parameters				30.257	30.519	29.996

Figure 3 shows the partition of the spaces and their optimization for the IG-based fuzzy model with MFs 2x3 and Type 3 for input variables *TComm* and *MChar*. Figure 4 depicts the values of the performance index produced in successive generations of the genetic optimization at the same case.



(a) *TComm* (b) *MChar*

Fig. 3. Initial and optimized membership functions for the IG-based fuzzy model

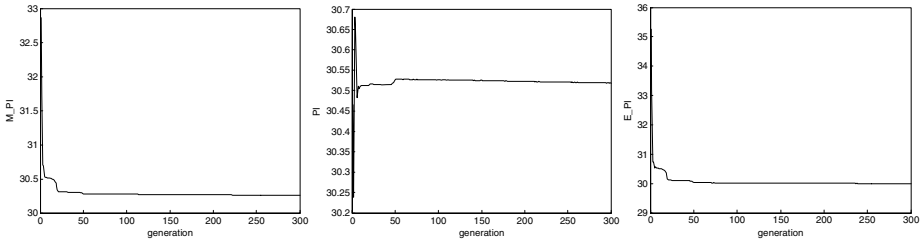


Fig. 4. Optimal convergence process of performance index for IG-based fuzzy model

Table 2 contains a comparative analysis including the previous model. Regression models are constructed by a linear equation. The comparative analysis reveals that the proposed model comes with high accuracy and improved prediction (generalization) capabilities with smaller rules.

Table 2. Comparison of identification error with previous models

Model	Selected inputs	No. of MFs(Rules)	Conse-quence. Type	PIs	E_PI
Regression Model [11]	All			40.056	36.322
	TComm, MChar, DChar, N			43.849	38.917
SONFN [12]	*TComm, MChar, DChar, N*	2x2x2x2 (16)	2	39.179	23.864
Our model	*TComm, MChar*	2x3 (6)	3	30.519	29.996

6 Conclusions

In this paper, we have developed a comprehensive framework for IG-based fuzzy system. We also showed how this model is used to apply software data. The underlying idea deals with an optimization of information granules by exploiting techniques of clustering and genetic algorithms. We defined some initial membership functions and the polynomial functions by means of information granulation realized with the C-Means clustering. The genetic algorithm was used afterwards to tune the initial values of the membership functions. Genetic algorithms were also used for further structural and parametric optimization of the fuzzy model. The experimental studies show that the model is compact while its performance is better than some other models previously discussed in the literature. Through the use of the certain form of the performance we were able to achieve a balance between the approximation and generalization abilities of the resulting model. While the detailed discussion has been exclusively focused on triangular fuzzy sets, the developed methodology applies equally well to other classes of fuzzy sets as well as various types of nonlinear local models. The proposed models scale up quite easily and do not suffer from the curse of dimensionality encountered in some other architecture of rule-based systems.

Acknowledgements. This work was supported by the Korea Research Foundation Grant funded by the Korean Government (MOEHRD)(KRF-2006-311-D00194, Basic Research Promotion Fund).

References

1. Zadeh, L.A.: Fuzzy sets. Information and Control. 8 (1965) 338-353
2. Tong, R.M.: Synthesis of fuzzy models for industrial processes. Int. J Gen Syst. 4 (1978) 143-162
3. Pedrycz, W.: Numerical and application aspects of fuzzy relational equations. Fuzzy Sets Syst. 11 (1983) 1-18
4. Takagi, T., Sugeno, M.: Fuzzy identification of systems and its applications to modeling and control. IEEE Trans Syst, Cybern. SMC-15(1) (1985) 116-132
5. Sugeno, M., Yasukawa, T.: Linguistic modeling based on numerical data. In: IFSA'91 Brussels, Computer, Management & System Science. (1991) 264-267
6. Oh, S.K., Pedrycz, W.: Identification of fuzzy systems by means of an auto-tuning algorithm and its application to nonlinear systems. Fuzzy Sets and Syst. 115(2) (2000) 205-230
7. Zadeh, L.A.: Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic. Fuzzy Sets and Syst. 90 (1997) 111-117
8. Pderycz, W., Vukovich, G.: Granular neural networks. Neurocomputing. 36 (2001) 205-224
9. Krishnaiah, P.R., Kanal, L.N., editors.: Classification, pattern recognition, and reduction of dimensionality, volume 2 of Handbook of Statistics. North-Holland Amsterdam (1982)
10. Golderg, D.E.: Genetic Algorithm in search, Optimization & Machine Learning, Addison Wesley (1989)
11. Lyu, M.R.: Handbook of Software Reliability Engineering. McGraw-Hill, New York. 1995 510-514
12. Oh, S.K., Pderycz, W., Park, B.J.: Self-organizing neurofuzzy networks in modeling software data. Fuzzy Sets and Systems. 145 (2004) 165-181
13. Park, H.S., Oh, S.K.: Fuzzy Relation-based Fuzzy Neural-Networks Using a Hybrid Identification Algorithm. International Journal of Control Automation and Systems. 1(3) (2003) 289-300

Evolvable Face Recognition Based on Evolutionary Algorithm and Gabor Wavelet Networks

Chuansheng Wu[1], Yong Ding[1], and Lishan Kang[2]

[1] School of Science, Wuhan University of technology, Wuhan, 430070, China
[2] State Key lab of software engineering, Wuhan University, Wuhan, 430072, China
Chuansheng.Wu@gmail.com, acerearth@163.com, Lishan.Kang@gmail.com

Abstract. Gabor Wavelet Networks (GWN) is a method for face recognition. Evolutionary Algorithms are proved to be efficient to deal with GWN optimization problems. Inver-over Evolutionary Algorithm (IOEA) mixed with GWN performs well in face recognition because of its considerable effects on complex functions optimizations, which is called IOEA-GWN put forward by us, and shows higher recognition rate than Simple Genetic Algorithm mixed with GWN (called SGA-GWN) in the same experiment conditions according to our practices.

Keywords: Face Recognition, Evolutionary Algorithm, Wavelet Networks.

1 Introduction

GWN for face recognition [1], [2], [3], [4]: Gabor wavelets functions and relative weight coefficients (multi-parameters) can be used to describe 2-Dimension faces. But optimizations of objective functions and relative parameters have a direct impact on the results of face recognition. However, in real-world applications of face recognition, it is impossible that efficient selections of initial values can be insured constantly considering robustness. Methods insensitive to initial values should be used to improve the effects of face images reconstruction and recognition.

Levenberg-Marquardt Algorithm (LMA) [3] is sensitive to initial values and relative results are easy to be made an impacted on by variant initial values [5]; SGA which is used to optimize GWN [5] (called SGA-GWN) can reduce the susceptibility to initial values. However, the results of SGA-GWN show slow convergence and easy to converge into local extreme values [5]. Inver-over Evolutionary Algorithm (IOEA) has considerable effects on functions and parameters optimizations [6], [7], [8], [9]. We respectively use IOEA and SGA to optimize GWN in the experiments.

2 Face Images Database

We select Yale face database as a tested object.

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 1116–1123, 2007.
© Springer-Verlag Berlin Heidelberg 2007



Fig. 1. Examples in Yale face database

Yale face database can be described here: 165 images (all are grayscale images), 15 subjects, 128×128 (image resolution), normal faces, different expressions, few covers, variant illuminant directions. One image of each subject is called sample-image (I) and the rest are called test-images (I').

3 Face Images Pre-processing

Face images pre-processing including lightness normalization is useful for following face recognition [10], [11].

We derive the whole expression of lightness normalization as follows during the pre-processing.

Gray value $f(x, y) \in [0, 255]$. We assume that the image-height is h, the image-width is w, and the default average lightness value $\bar{v}$ is 128.

Average lightness value $v = \frac{1}{wh} \sum_{x=1}^{w} \sum_{y=1}^{h} f(x, y)$.

$f'(x, y)$ denote images (after we do lightness normalization).

$$f'(x,y) = g(\bar{v} f(x,y)/v), \quad g(x) = \begin{cases} 0, & x < 0 \\ x, & 0 \le x \le 255 \\ 255, & x > 255 \end{cases}$$

Therefore the whole expression is:

$$f'(x,y) = \begin{cases} 0, & f(x,y) < 0 \\ \frac{128hwf(x,y)}{\sum_{x=1}^{w} \sum_{y=1}^{h} f(x,y)}, & 0 \le f(x,y) \le \frac{255 \sum_{x=1}^{w} \sum_{y=1}^{h} f(x,y)}{128hw} \\ 255, & f(x,y) > \frac{255 \sum_{x=1}^{w} \sum_{y=1}^{h} f(x,y)}{128hw} \end{cases}$$

4 GWN Process Analysis

There are four steps in GWN process analysis.

4.1 Gabor Wavelets Optimizations

GWN is expressed by Gabor wavelets functions and relative weight coefficients. Gabor wavelets functions and multi-parameters optimizations' purpose is to describe face sample images and test images, and to search the effective reconstruction ($\hat{I}$) of image I. As to same subjects' different images, Gabor wavelets functions are close in values but relative weight coefficients are different. According to our experience, the number of Gabor wavelets funtions is selected 52 in each objective function (i=1,2,3,4,...,52):

$$\min g(G) = \min g(c_{x_i}, c_{y_i}, \theta_i, s_{x_i}, s_{y_i}) = \min \left[\left\| I - \sum_{i=1}^{N} w_i \psi_{n_i} \right\|_2^2 \right]$$

We expand the expression so that it is suitable for Matlab:

$$\begin{aligned} g(c_{x_i}, c_{y_i}, \theta_i, s_{x_i}, s_{y_i}) &= \left\| I - \sum_{i=1}^{52} [w_i \psi_{n_i}(c_{x_i}, c_{y_i}, \theta_i, s_{x_i}, s_{y_i})] \right\|_2^2 \\ &= \left\langle (I - \sum_{i=1}^{N} w_i \psi_{n_i}), (I - \sum_{i=1}^{N} w_i \psi_{n_i}) \right\rangle \\ &= \int_{-\infty}^{\infty} (I - \sum_{i=1}^{N} w_i \psi_{n_i}) \overline{(I - \sum_{i=1}^{N} w_i \psi_{n_i})} dx \end{aligned}$$

Gabor wavelets functions are not orthogonal. Therefore the weight coefficient w_i can not be directly determined but can be derived from $\bar{\psi}$ (dual wavelet).

$$\left\langle \psi_{n_i}, \bar{\psi}_{n_j} \right\rangle = \delta_{i,j} = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases},$$

$$\bar{\psi}_{n_i} = \frac{1}{\sqrt{2\pi}} \int e^{jwt} \psi_{n_i}(t) dt, \quad w_i = \left\langle I, \overline{\psi}_{n_i} \right\rangle.$$

n_i=(c_{x_i}, c_{y_i}, θ_i, s_{x_i}, s_{y_i}) denote parameters of Gabor wavelets functions.

c_{x_i}, c_{y_i} denote translation factors. θ_i denote rotation factors. s_{x_i}, s_{y_i} denote scaling factors [5]. Odd Gabor wavelets functions reflect face structures:

$$\begin{aligned} \psi_{n_i}(x, y) = & \frac{2}{\pi^3} \exp \Big\{ - \frac{1}{2\pi^2} [s_{x_i}((x - c_{x_i}) \cos\theta_i + (y - c_{y_i}) \sin\theta_i)]^2 \\ & + [s_{y_i}(-(x - c_{x_i}) \sin\theta_i + (y - c_{y_i}) \cos\theta_i)]^2 \Big\} \\ & \cdot \sin [s_{x_i} ((x - c_{x_i}) \cos\theta_i - (y - c_{y_i}) \sin\theta_i)] \end{aligned}$$

4.2 Super Gabor Wavelets Optimizations

Super Gabor wavelets are the results of Gabor wavelets translation, scaling, and rotation. The Super Gabor wavelets optimizations' purpose is to transform sample images to align with test images.

4.3 Gabor Wavelets Parameters Expressed in IOEA

T denotes transposition of a vector or a matrix.

$$G_i = (c_{x_i}, c_{y_i}, \theta_i, s_{x_i}, s_{y_i})^T \in R^n, \quad c_{x_i}, c_{y_i}, \theta_i, s_{x_i}, s_{y_i} \in D.$$

According to our experience, the range of parameters is:

$$D = \left\{ G_i \middle| -64 \le c_{x_i}, c_{y_i} \le 63, 0 \le \theta_i \le 2\pi, 0 \le s_{x_i}, s_{y_i} \le 10 \right\}$$

M points in D are: $G'_j = (x_{j1}, x_{j2}, \ldots, x_{jm})^T$, $j = 1, 2, \ldots, m$.
They span a subspace S:

$$S = \left\{ G \in D \middle| G = \sum_{i=1}^{m} a_i G'_i \right\}. \quad \sum_{i=1}^{m} a_i = 1, \quad -0.5 \le a_i \le 1.5$$

4.4 Contrasts of Various GWNs

We compare test-images' GWN with Super-GWN derived from results of Super Gabor wavelets optimizations. Various GWNs show distinctions between different subjects and same subjects under different conditions (illumination, expression, etc.). GWNs contrasts index are:

1) Euclidean distances [5] (V and W are weight coefficients vectors):

$$d^2_{\psi}(I, I') = \left\| \hat{I} - \hat{I}' \right\|_2 = (V - W)^T \langle \psi_i, \psi_j \rangle (V - W)$$

2) Cross-correlation factors [5]:

$$d^c_{\psi}(I, I') = \frac{V^T \langle \psi_i, \psi_j \rangle W}{\sqrt{V^T \langle \psi_i, \psi_j \rangle V} \sqrt{W^T \langle \psi_i, \psi_j \rangle W}}$$

Various GWNs tend to be more close when $d^c_{\psi}(I, I') \to 1$.

5 Algorithms and Experiments

We especially use IOEA to optimize GWN as follows (IOEA-GWN).
{ begin

1 Initialize Population $P = (G_1, G_2, \ldots, G_{52})^T$, $G_i \in D$;
2 $t := 0$;
3 $G_{best} = \arg \min_{1 \le i \le 52} (g(G_i))$; ("arg" denotes relative parameters G when Gabor wavelets functions converge into overall optimum);
4 $G_{worst} = \arg \max_{1 \le i \le 52} (g(G_i))$;

5 *While* $(|g(G_{best}) - g(G_{worst})| > \varepsilon)$ *do*
6 {Select M points $G_1, G_2, \ldots, G_m$ randomly from P to span subspace S;
7 Select one point $G_{crossover}$ randomly from S;
8 *If* $(g(G_{crossover}) < g(G_{worst}))$ *then* $G_{worst} = G_{crossover}$;
9 $t := t + 1$;
10 $G_{best} = \arg \min_{1 \le i \le 52} (g(G_i))$;
11 $G_{worst} = \arg \max_{1 \le i \le 52} (g(G_i))$;
12 end do}
13 output t, P;
14 end }

We use IOEA-GWN and SGA-GWN respectively with 5×52 parameters by real number programming. Experiment Machine: CPU Celeron 1.7G, Memory 512M. Programming tool are both Matlab 6.5. Popular sizes are both 100 [12]. The Maximum numbers of generations are both 100. Selections are both roulette wheel selection.

1) In IOEA-GWN: According to our experience, the value of M is 10 and Multi-parent's crossover-probability is 0.6.
2) In SGA-GWN: Crossover-probability is 0.6. In our experiments, "multi-point crossover" destroys correct solutions, and "one-point crossover" shows low diversity of solutions. Therefore we use "two-point crossover" finally. Mutation-probability is 0.01.



Fig. 2. An original image

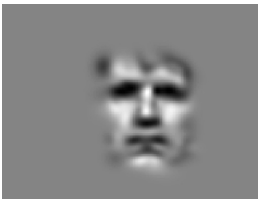


Fig. 3. Gabor wavelets optimizations

We select 55 images (including: 5 subjects and 11 images per subject) from Yale face database. Then we determine a sample image per subject and other 50 images (named test-images) for recognition.

The recognition rate of SGA-GWN is not same as the high rate 97.8% [5] because of different numbers of test-images and databases.

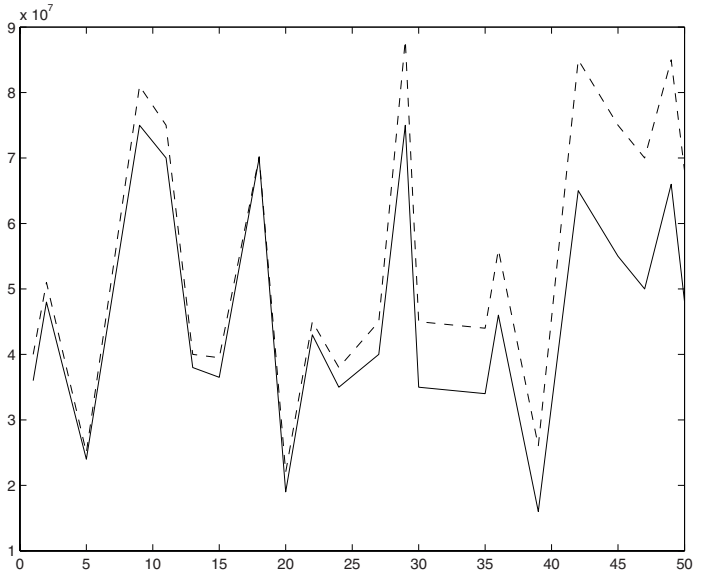


Fig. 4. Results of SGA-GWN (*broken line*) and IOEA-GWN (*real line*)

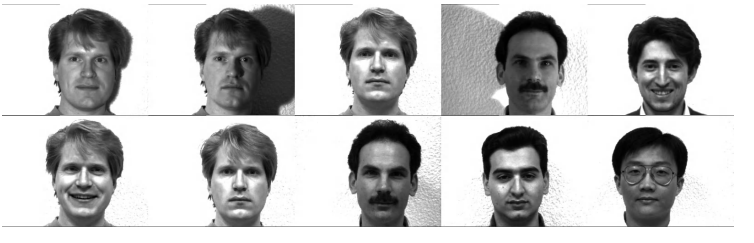


Fig. 5. Images marked 1-5 and 6-10 (The fifth is a sample image)

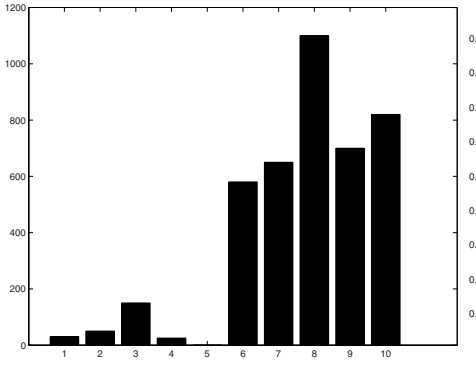


Fig. 6. $d^2_{\psi}(I, I')$

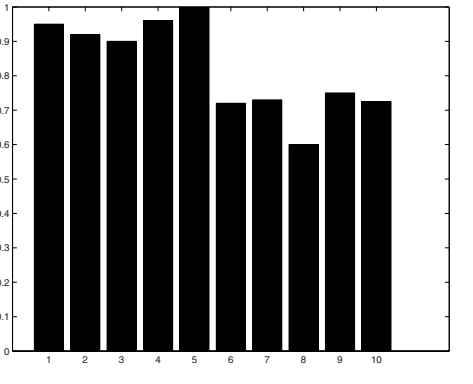


Fig. 7. $d^c_{\psi}(I, I')$

Table 1. Contrasts of Algorithms

Algorithm	SGA-GWN	IOEA-GWN
Test set	50	50
Recognition Images	41	46
Time(minutes)	22	16
Recognition rate	82%	92%

6 Conclusion and Future Work

Following methods can be used to improve convergent rate and continuous recognition rate.

1) Improved IOEA can be used to optimize GWN.
2) Gene Expression Programming (GEP) can be used to optimize GWN (called GEP-GWN). Because GEP has some features of dealing with high dimension parameters optimizations problems, and GEP can also be used for face modeling.
3) All these methods can be used in 2D/3D face recognition [13] and may be more useful for face tracking.

Therefore, the new branch "Evolvable Face Recognition, or Evolutionary Face Recognition" is considerable to research including "Evolutionary face modeling", "Evolutionary face algorithms", "Evolutionary automatic face programming", and "Evolutionary face hardware". All of these are considerable to be used to build "Evolvable Face Systems".

Acknowledgments. acknowledge the financial support of the National Natural Science Foundation of China under Grant No.60473081.

References

1. Phillips, P., Moon, H., Rizvi, S., Rauss, P.: The feret evaluation methodology for face-recognition algorithms. Pattern Analysis and Machine Intelligence, IEEE Transactions on. **22** (2000) 1090–1104
2. Shan, S.: Study on Some Key Issues in Face Recognition. PhD thesis, Chinese Academy of Sciences (2004)
3. Krueger, V., Sommer, G.: Gabor wavelet networks for object representation. Journal of the Optical Society of America. **19**(6) (2002) 1112–1119
4. Lee, T.: Image representation using 2d gabor wavelets. Pattern Analysis and Machine Intelligence, IEEE Transactions on. **18** (1996) 959–971
5. Wang, B.: Global-Feature-based face recognition technology using genetic algorithms. PhD thesis, Beijing University of posts and telecommunications (2004)
6. Zhan, W., Dai, G., Gong, W.: A high-efficiency hybrid evolutionary algorithm for solving function optimization problem. Computer Engineering and Applications. (2006) 1–3

7. Li, Y., Kang, Z., Liu, P.: Guo's algorithm and its application. Journal of Wuhan automotive polytechnic university. (2000) 1–3
8. Wang, J., Chen, J., Wei, W., Li, Z.: Based on improved gt algorithm for tsp. Computer Engineering and Design. (2006) 1–3
9. Guo, T., Kang, L., Li, Y.: A new algorithm for solving function optimization problems with inequality constraints. Natural Science Edition, Journal of Wuhan University. (1999)
10. Qin, Z.: Research on face detection and face recognition technology (2005)
11. Wiskott, L., Fellous, J., Kfuger, N., Malsburg, C.: Face recognition by elastic bunch graph matching. Pattern Analysis and Machine Intelligence, IEEE Transactions on. **19** (1997) 775–779
12. Pan, Z., Kang, L., Chen, Y.: Evolutionary Computation. Tsinghua University Press (2000)
13. Hu, Y., Yin, B., Cheng, S., Gu, C., Liu, W.: Research on key technology in construction of a chinese 3d face database. Journal of Computer Research and Development. (2005) 1–3

Automated Design Approach for Analog Circuit Using Genetic Algorithm

Xuewen Xia[1], Yuanxiang Li[2], Weiqin Ying[2], and Lei Chen[1]

[1] School of Computer Science, Wuhan University, Wuhan 430079, China
[2] State Key Lab. of Software Engineering, Wuhan University, Wuhan 430072, China
laughkid@163.com

Abstract. The technology of electronic design automation (EDA) has improved the efficiency of design process, however, designer is still required much special knowledge of circuit. During the past decade, using genetic algorithm (GA) to design circuit had attracted many experts and scholars. However, too much more attention was focus on a circuit's function and many other factors had been neglected which caused the circuit had little applicability. This paper proposes an automated design approach for analog circuit based on a multi-objective adaptive GA. The multi-objective fitness evaluation method, which can dynamic adjust parameter, is selected. And a parallel evolution strategy which separates evolution of circuit structure and element value is adopted but also organically combined them by weight vectors. The experimental results indicate that this approach obviously be able to improve the evolution efficiency and could generate numbers of suitable circuits.

Keywords: Evolutionary algorithms, Electronic Design Automation, Evolving hardware.

1 Introduction

In contrast to conventional circuit design where the designers were required much special knowledge of circuit, Evolvable Hardware (EHW) technology needs fewer designers' intervention and special knowledge of circuit during the design process.

Evolutionary Hardware uses techniques derived from evolutionary computation such as genetic algorithm and genetic programming to develop electronic circuits, which capable of solving real world problems. The researches of Koza [1]and his collaborators on analog circuit synthesis through genetic programming [2](GP) is likely the most successful evolutionary computation-based approach so far, but this approach, like many other experiments, indicated that design an applicable circuit, especially analog circuit, is very time-consuming [3], and even is infeasible. There-fore, various experiments on speeding up the GA computation have been undertaken [4], and other approaches to the problem have been undertaken by using variable length chromosome [5].

In this paper, a simple encoding scheme based on Spice Netlist file format is selected. And a dynamic adjusting method based on evolutionary strategy which could improve the evolution efficiency is selected. Two representative analog

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 1124–1130, 2007.
© Springer-Verlag Berlin Heidelberg 2007

circuits are adopted to test the approach because design analog circuit is more difficult than design digital circuit.

The paper is organized as follows. The approach of evolving analog circuits, includes the presentation of circuit, the evolutionary strategy and the evaluation method are introduced in section 2. In section 3, the results of experiment and analysis are presented. At last, a conclusion is provided.

2 Approach of Evolving Analog Circuits

2.1 Encoding and Decoding Scheme

It is an elementary problem in evolvable hardware that to chose a suitable method of encoding and decoding. In general, there are two methods: Global encoding, which refers to encoding with all information of a circuit, and local encoding, which refers to encoding with the part information of a circuit. The latter method is selected in this paper in which the circuit has been separated into two parts: structure part and element value part. And these two parts are respectively evolved with each local coding.

Instead of some detailed requirements to initialize the circuit, there are only the numbers of elements and the type of it should be defined advance according to the experiential knowledge of circuit's complexity.

For the structural evolution, the type and link-nodes of elements are encoded into binary code, but the value of it is not included. The number of an element's nodes is determined by the maximum nodes that the element has.

$$E_1 = [Type, Node_1, Node_2, ..., Node_n]. \tag{1}$$

In (1), $Type$ denotes the index of element type, while $Node_i$ (i[1, n]) denotes each link-node of element. The value of element is random generated at the beginning of evolution process. The chromosome of a circuit is composed of every element's binary code which linking with each other.

For the value evolution, the code is almost the same as structure evolution, but the element's value is, instead of link-node, encoded.

$$E_2 = [Type, Value_1, Value_2, ..., Value_m]. \tag{2}$$

In (2), the $Valu_i$ ($i \in [1, m]$) denotes many different parameters of an element, and the number of the element's value is determined by it's $Type$. For example, the resistivity and the linear temperature-coefficient are selected to be parameters of a resistance.

Before simulated by SPICE emluator, these two results (structure and elements) should be combined in order to get the whole information of a circuit. Firstly, the set of element from structural evolution should be decoded individually, so the type of each element could be acquired by table querying, and every link-nodes of element should be recorded in order to obtain the topology of circuit. Then the set of value of each element should be obtained by decoding the

set of element from value evolution. Finally the whole data of each element in the circuit are presented. The format could be described as fellow:

$$E = [Type, Node_i, Value_j]. \tag{3}$$

2.2 Simulate Circuit with Pretreatment Code

wing to the features of genetic algorithm's encoding model and the operators of it, the element's link condition might not be directed properly. It may cause some topology errors which could result in more than fifty percents of circuits be illegal, which would obviously decrease the efficiency and feasibility of evolution because majority of time is spent on the simulating during the process. To avoid this case, the circuit must be pretreated and be checked for these cases:

1. Less than 2 connections at a node;
2. No DC path to ground from a node;
3. 3. Inductor/voltage source loop found.

These errors could be obtained by scanning the circuit advance, and then corrected or deleted them, as a result, the circuit's validity could be guarantied to some extend.

At last, combining the whole valid circuit's information and the commands of output data, a regular SPICE Netlist file would be generated, hence the simulation could be proceed successfully by software SPICE.

2.3 Evolutionary Strategies Based on the Dynamic Adjusting Weight Vector

Since rates of mutation and crossover influence the efficiency of genetic algorithm, so they should be altered along with evolution process and the fluctuation of environment. Because of the parallel two layers evolutionary strategy, mutation rates (P_sm) and crossover rates (P_sc) of structural evolution have relations to those of value evolution. With the increasing generation and the rates of fitness, P_sm and P_sc of structure and value will became less.

$$P_{sm} = K_1(1 - exp(-\frac{dFitness}{dGeneration})) \tag{4}$$

$$P_{sc} = K_2 P_{sm} \tag{5}$$

In the equations above,$K_1 \in (0,1)$,$K_2 \in (0,1)$.

$$P_v = exp(-(P_{sc} + P_{sm})) \tag{6}$$

After the start of value evolution, the rates of mutation and crossover resume in order to preventing local convergence. Based on the traditional theory of evolutionary algorithm, the algorithm could not search effectively after the population had con-verged [6,7]. The approach in this paper could obtain optimal parameters from sorts of circuit, so it overcome local convergence during the early period, and also avoid some cases that one circuit been eliminated for its unsuitable element value.

2.4 Fitness Evaluation Method

The quality of evaluation function influences the result and the process of analog circuit's evolution greatly. A precisely comprehensive evaluation strategy would direct evolution to a more efficient process, which could obtain a more suitable result in a relative short period.

In this paper, an offline evolution model based on SPICE software has been selected, and a multi-objective fitness evaluation function is adopted in order to meet many design objectives. The fitness of circuit is described as:

$$Fitness = f(g, s, p) \tag{7}$$

The vector g denotes the optimization degree of a circuit performance, s denotes the number of type of elements, and p denotes the scale of a circuit. The experiences of design during past remind us that a good circuit should have less kinds of elements and small scale besides good performance. In this paper, function f is described as follow:

$$f(g, s, p) = c_1 g + c_2 s + c_3 p, \quad \sum_{i=1}^{3} c_i = 1 \tag{8}$$

In (8),c_i is weight vector which could be used to adjust the weight of three parameters according to designer's demands.

The performance of a circuit is determined by many aspects, so the vector g should be defined to be diversified function according to different design objectives. Owing to the separated parallel evolution mode, the fitness of circuit's structure and element value could be separated into two different functions. In the experiment of transistor amplifier design, a stable amplifier is the destination of structural evolution, while its amplification ratio is the mission of value evolution. When a low-pass filter is designed, the structural evolution will take charge of low-pass performance, and the feature of frequency and voltage are left to value evolution to be obtained. The evaluation strategy of transistor amplifier's structure is to analyze the curve from transient analysis.

The evaluation function is $g_1(x)$ that expressed by (9). The vector is a weight vector, $V_o(x)$ and $V_i(x)$ respectively denotes the voltages of output and input. Function $h_1(x)$ evaluates the stability of amplification of output voltage, and function $e(x)$ evaluates the AC work state of Bipolar Junction Transistor from individual x in order to guarantee a stable working condition to the transistor in its value domain. The value evaluation function is $g_2(x)$ that expressed by (10). The vector A is an expectant voltage gain, and function $h_2(x)$ will evaluate the difference between the expectant outcome and the experimenter's.

$$g_1(x) = \alpha h_1 \sum_m \left(\frac{v_{om(x)}}{v_{im(x)}} - \frac{v_{om-1(x)}}{v_{im-1(x)}}\right)^2 + (1-\alpha)e(x), \quad \alpha \in (0,1) \tag{9}$$

$$g_2(x) = \alpha h_2 \sum_m \left(A - \frac{v_{om(x)}}{v_{im(x)}}\right)^2 + (1-\alpha)e(x), \quad \alpha \in (0,1) \tag{10}$$

The evaluation strategy of a low-pass filter structural evolution is to analyze the curve from frequency analysis. The fitness function is $g_3(x)$ that expressed by equation (11). $v_{max}(x)$ and $v_{min}(x)$ respectively expresses the maximum and minimum of output curve, $p(x)$ and $q(x)$ are pass-cutoff point and reject-cutoff point that could be measured by extremum analysis method. Extremum analysis method is based on comparing every extremum point with $v_{max}(x)$ to insure the position of pass-cutoff point and reject-cutoff point, and analyzing the region between the two points combine with each point's slop or differential coefficient to insure pass-cutoff frequency f_p and reject-cutoff frequency f_r.

$$g_3(x) = \beta h_3(\sum_{i=1}^{p(x)}(v_{max}(x) - v_{oi}(x)) + \sum_{i=q(x)}^{n}(v_{oi}(x) - v_{min}(x))),\ \beta\in(0,1) \quad (11)$$

The evaluation strategy of low-pass filter value evolution is almost the same as that of the transistor amplifier. Evaluating the difference between experimental and expectant value of voltage gain and feature frequency, combining with evaluating work state of operational amplifier, the outcome of value evolution could be obtained.

3 Experimental Results and Analysis

In this paper, two kinds of circuits, passive filter and amplifier, are experimented because that with the following reasons:

1. Filter circuit includes resistance, capacitor, and inductance, therefore the circuit's configuration is very simple and it is easy to analyze the circuit. Furthermore, the analysis of input-output characteristics of passive filter is representative which involves many parameters of analog circuit.
2. Most analog circuit could be made up of resistance, capacitor, inductance, and dynatron. So an amplifier circuit which includes dynatron could represent majority structure of analog circuit.
3. When making an appraisal for a passive filter and amplifier, many design objectives should be considered, so the search of these circuits could testify the validity of multi-objective evaluation strategy proposed by this paper.

The parameters of evolution are set as follow that maximal number of element is 20, maximal number of node is 8, max generation is 2000, size of population is 60 (40 for structure and 20 for value), initial rate of mutation and crossover was P_m=0.1, P_c=0.03.

3.1 Experiment 1

In this experiment, the objectives of transistor amplifier are prearranged that voltage magnification is 350, the generate ratio of resistor, capacitance, power supply, and transistor is 6:3:1:2.

A relative good circuit has been obtained after 800 1000 generations, and a perfect circuit has selected to demonstrate in Figure.1:

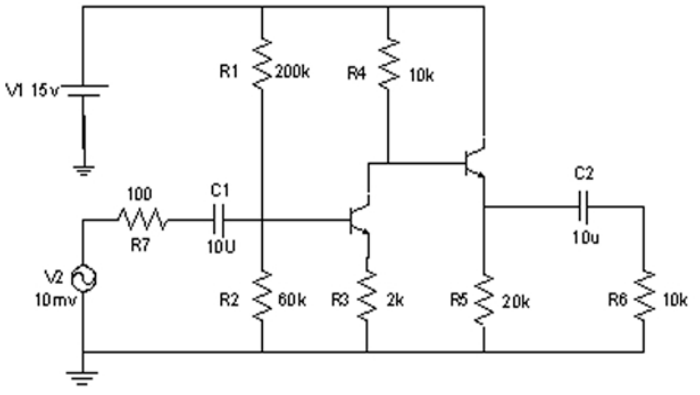


Fig. 1. The best voltage amplifier structure from generation 1000

3.2 Experiment 2

In this experiment, the objectives of low-pass filter (LPF) are prearranged as follow: voltage gain is 20 dB, transmission bands is 30KHz., the generate ratio of resistor, capacitance, power supply, and transistor is 6:3:1:2.

A relative good circuit has been obtained after 800 1000 generations, and a perfect circuit has selected to demonstrate in Figure.2:

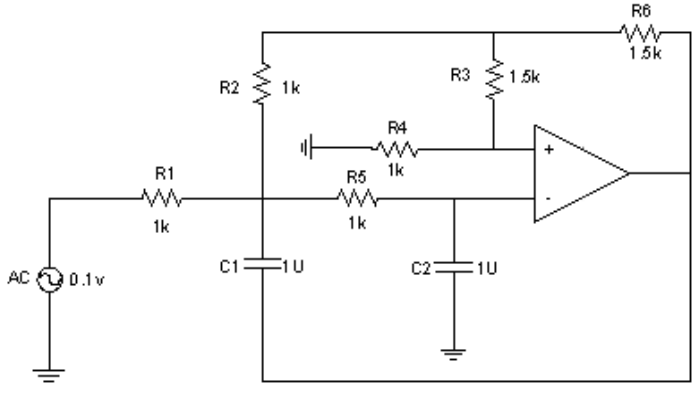


Fig. 2. The best low-pass filter structure from generation 1000

This filter allows frequency lower than 30 kHz to pass through it, but prevented higher frequencies from doing so.

3.3 Analysis

As we know that the structure and values of elements contribute to an analog circuit's function. Experiences indicate that congeneric circuits generally have the

same topology while some values of elements are different. In the experiments above, the proper topologies of a transistor amplifier and a low-pass filter were found during generation 500 600, despite that the magnification and the frequency were not consistent with the prearrangement circuits. The posterior evolutions were stress on how to adjust the values of elements in order to meet the scheduled requirements. Experiences in this paper indicated that, compare with traditional methods, this strategy shorten the period of EHW, especially in analog circuits.

4 Conclusions

In this paper, an evolution approach for designing analog circuit is proposed, which adopts a parallel evolutionary strategy that separated evolution of circuit into two parts, called structure and element value, which are organically combine by weight vectors. The experiments have indicated that parallel evolutionary strategy combining with separate fitness evaluate scheme can automatically design a circuit efficiently. Its excellence is saving time and multi-result.

Although little knowledge about circuit is required during EHW, however, the special characteristics of a circuit would consequentially improve the efficiency of EHW. So it is the main target of future researches that how to extract the characteristics of structure and element value from a preconcerted circuit to enhance the quality and to improve the efficiency of EHW. Since the strategies and operators of genetic are also important to EHW, so these strategies and operators should be explored and researched specially to ameliorate EHW.

Acknowledgement. This research is supported by the National Natural Science Foundation of China under Grant No.60473014.

References

1. Koza, J.R. Bennett, F.H. Andre, M.A. et al.: Automated synthesis of analog electrical circuits by means of genetic programming. IEEE Trans. on Evolutionary Computation **1**(2) (1997) 109–128
2. Koza, J.R.: Genetic Programming: On the Programming of Computers by Means of Natural Selection. MIT Press, Cambridge, MA (1992)
3. de Garis H.: Evolvable hardware: Genetic programming of a darwin machine. In: Proceeding of Artificial Neural Nets ad Genetic Algorithms, Austria: Springer-Verlag (1993) 441–449
4. Cantu-Paz E.: A survey of parallel genetic algorithms. Calculateurs parallels **10**(2) (1998) 141–171
5. Iwata, M.: A pattern recognition system using evolvable hardware. In: Parallel Problem Solving from Nature IV, Springer Verlag (1996) 761–770
6. de Garis H.: An artificial brain: Atr's cam-brain project aims to build/evolve an artificial brain with a million neural net modules inside a trillion cell cellular automata machine. New Generation Computing **12**(2) (1994) 215–221
7. Goldberg, D.E.: Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley Publishing Company, Reading,MA (1989)

Evolutionary Algorithm for Identifying Discontinuous Parameters of Inverse Problems

Zhijian Wu, Dazhi Jiang, and Lishan Kang

The State Key Laboratory of Software Engineering,
Wuhan University, Wuhan, China
{zjwu9551,jiangdazhi111007}@sina.com

Abstract. In this paper, we make an investigation into the discontinuous parameter identification in the case of elliptic problem. The discontinuous parameter is identified by evolutionary algorithm for the first time. For this kind of problem, we present a two-level evolutionary algorithm. The first level is the evolution for discontinuous point and the second level is the evolution for parameter. The numerical experiments suggest that the algorithm carries such features as good stability and adaptability and is not very sensitive to the noise in observation data as well.

Keywords: Parameter Identification, Inverse Problem, Evolutionary algorithms.

1 Introduction

The major purpose of this paper aims to propose an evolutionary algorithm in the identification of the unknown coefficient $q(x)$ in the case of elliptic problem. We will take into consideration the case when the coefficient $q(x)$ is discontinuous.

$$\begin{cases} -\dfrac{d}{dx}(q(x)\dfrac{d}{dx}u(x)) = f(x) \text{ in } \Omega \\ u(x) = 0 \text{ on } \Gamma \end{cases} \tag{1}$$

Here the Ω can be any bounded domain in R, with piece wise smooth boundary Γ and $f(x)$ being given.

So far, a lot of algorithms for solving the continuous parameter identification problem have been worked out, among which, three kinds of methods can be used. One of them is termed as the traditional mathematic and physical methods [1-5]. The second one is known as the evolutionary algorithms [6-10], and the numerical experiments indicate that these algorithms are good to solve inverse problem. The third one is to model parameter function by Genetic Programming [11-14]. But much less work has been done on the identification of discontinuous parameters [15]. So in this paper, we investigate the discontinuous parameter identification and the discontinuous parameter is identified by evolutionary algorithm for the first time. The paper is presented as follows. In the second section, the new algorithm is described. Some numerical experiments are described in the third section and in the fourth section some conclusions have been drawn.

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 1131–1138, 2007.
© Springer-Verlag Berlin Heidelberg 2007

2 Description of Algorithm

As an inverse problem in elliptic systems, parameter identifying is the process to find the potential solution $q^*(x)$ that makes $u_{q^*(x)}$ match the observation data of $u(x)$ as optimally as possible, where we obtain $u_{q^*(x)}$ by solving equation (1) with $q^*(x)$.

To put it simpler, the following elliptic problem can be considered.

$$\begin{cases} -\dfrac{d}{dx}(q(x)\dfrac{d}{dx}u(x)) = f(x), x \in (0,1) \\ u(0) = u(1) = 0 \end{cases} \tag{2}$$

We consider the case in which the $q(x)$ has one discontinuous point, which can be expanded to multi discontinuous point case. As for this kind of problem, we present a two-level evolutionary algorithm. The first level is the evolution for discontinuous point and the second level is the evolution for $q(x)$.

The interval [0,1] is divided equally into n parts, while the step size $h=1/n$, and mesh point $x_i=ih$ (i=0,1,…,n). Suppose we have the observation data of $u(x)$ at mesh point x_i (i=1,2,…,n-1), two boundary points values are given. Here the observation data of $u(x)$ can be denoted as $\tilde{u}_{ob} = (\tilde{u}_1, \tilde{u}_2, \ldots, \tilde{u}_{n-1})$.

In the first level evolution, we denote the individual as d_i (i=1,2,…,N_d). For individual d_i (for the sake of convenience, we suppose the d_i is just the same as the some node x_l), we conduct the second level evolution (i.e. the evolution of $q(x)$). Suppose $d_i=x_l=l*h$, we construct the following base functions:

$$\varphi_0(x) = \begin{cases} \dfrac{x_1 - x}{h}, & x \in [x_0, x_1], \\ 0, & otherwise. \end{cases} \qquad \varphi_i(x) = \begin{cases} \dfrac{x - x_{i-1}}{h}, & x \in [x_{i-1}, x_i], \\ \dfrac{x_{i+1} - x}{h}, & x \in [x_i, x_{i+1}], i = 1,2,\ldots,l-1 \\ 0, & otherwise. \end{cases}$$

$$\varphi_l^{(1)}(x) = \begin{cases} \dfrac{x - x_{l-1}}{h}, & x \in [x_{l-1}, x_l], \\ 0, & otherwise. \end{cases} \quad \varphi_l^{(2)}(x) = \begin{cases} \dfrac{x_{l+1} - x}{h}, & x \in [x_l, x_{l+1}], \\ 0, & otherwise. \end{cases}$$

$$\varphi_i(x) = \begin{cases} \dfrac{x - x_{i-1}}{h}, & x \in [x_{i-1}, x_i], \\ \dfrac{x_{i+1} - x}{h}, & x \in [x_i, x_{i+1}], \ i = l+1, l+2, \ldots, n-1 \\ 0, & otherwise. \end{cases}$$

$$\varphi_n(x) = \begin{cases} \dfrac{x - x_{n-1}}{h}, & x \in [x_{n-1}, x_n], \\ 0, & otherwise. \end{cases}$$

$$\text{Set } q(x) = \sum_{i=0}^{l-1} q_i \varphi_i(x) + q_l^{(1)} \varphi_l^{(1)}(x) + q_l^{(2)} \varphi_l^{(2)}(x) + \sum_{i=l+1}^{n} q_i \varphi_i(x)$$

where $q_l^{(1)} = \lim_{x \to x_l^-} q(x)$, $q_l^{(2)} = \lim_{x \to x_l^+} q(x)$.

Identification of $q(x)$ is converted to that of the discrete values $q = (q_0, q_1, ..., q_l^{(1)}, q_l^{(2)}, ..., q_n)$. So for the second level evolution, we denote the individual as $(q_0, q_1, ..., q_l^{(1)}, q_l^{(2)}, ..., q_n)$. For each individual d_i we evolve it to get the best $q_{d_i}(x)$. For this $q_{d_i}(x)$, we can get $u_{d_i}(x)$ by solving equation (2). The fitness of individual d_i is defined as:

$$fitness(d_i) = h\sum_{i=1}^{n-1}(\bar{u}_{d_i}(x_i) - \tilde{u}_i)^2 + \beta(\sum_{i=1}^{l-1}|q_i - q_{i-1}| + |q_l^{(1)} - q_{l-1}| + |q_{l+1} - q_l^{(2)}| + \sum_{i=1+2}^{n}|q_i - q_{i-1}|)$$

For two individuals d_1 and d_2, if $fitness(d_1) < fitness(d_2)$ it is said that d_1 is better than d_2.

The following operators will be adopted in our algorithm.

Smooth operator
Begin
For i=1 to l-1 do q_i=(q_{i-1}+ q_i + q_{i+1})/3;
For i=l+1 to n-1 do q_i =(q_{i-1} + q_i + q_{i+1})/3;
End

Multiple parent crossover operator
The multiple parent crossover operator has been used in our algorithm, it plays an essential role in keeping the diversity of population, and making the parents' merits inherited by their offspring as well. The crossover operator is described as follows (suppose there are N individuals in population P).

Procedure multiple parents crossover operator
Begin
Step 1: Select M individuals $\vec{q}^{(1)}, \vec{q}^{(2)}, \ldots, \vec{q}^{(M)}$ randomly from population P to form a sub-space $V = \left\{ \vec{q} \mid \vec{q} \in D^{n+1}, \vec{q} = \sum_{i=1}^{M} a_i \vec{q}^{(i)} \right\}$, where a_i satisfies the condition $\sum_{i=1}^{M} a_i = 1, a_i \in [-0.5, 1.5]$;

Step 2: Produce a new individual $\vec{q}_{new}$ randomly in sub-space V.

End.

The algorithm for solving the discontinuous parameter problem is described as follows:

The first level evolutionary algorithm:
Begin
Step 1: N_d individuals $d_1, d_2, ..., d_{N_d}$ are randomly and uniformly produced in the search space as to form the initial population $D^{(0)}$ and set t_d= 0;
Step 2: For each d_i conduct the second level evolutionary algorithm to get an individual q_{d_i} corresponding to d_i.
Step 3: When terminative condition is satisfied with go to Step 5;

Step 4: Genetic operations for $D^{(t_d)}$. Conduct multiple parent crossover operating, and produce a new individual d_{new}. Conduct the second level evolutionary algorithm for d_{new}. If the d_{new} is better (i.e. the fitness is smaller) than the worst one in $D^{(t_d)}$ then substitute it for the worst one and form a new population $D^{(t_d+1)}$ else $D^{(t_d+1)} = D^{(t_d)}$. t_d= t_d +1. go to Step 3;
Step 5: Output the best individual d and the corresponding q_d.
End

The second level evolutionary algorithm
Begin
Step 1: N_q individuals are randomly and uniformly produced in the search space as to form the initial population $P^{(0)}$ and set $t_q = 0$;
Step 2: When terminative condition is satisfied with go to end;
Step 3: Genetic operations (crossover and smooth operation). Conduct multiple parent crossover operation, and produce a new individual. Conduct the smooth operator for the new individual. If the new individual is better (i.e. the fitness is smaller) than the worst one in $P^{(t_q)}$ then substitute it for the worst one and form a new population $P^{(t_q+1)}$ else $P^{(t_q+1)} = P^{(t_q)}$. t_q= t_q + 1. go to Step 2;
Step 4: Output the best solution.
End

In the following experiments, in the case of the first level evolution (discontinuous point evolution), the Popsize N_d is 10, the dimension M of crossover subspace is 3. As for the second level evolution ($q(x)$ evolution), the Popsize N_q is 100, the dimension M of crossover subspace is 10.

As for the second level evolution, the terminate condition is

$$\left(\frac{\sqrt{\sum (q_i^{(best)} - q_i^{(worst)})^2}}{\sqrt{\sum (q_i^{(best)})^2}} < 10^{-2} \text{ and } \frac{\sqrt{\sum_{i=1}^{n-1} (u^{(best)}(x_i) - \tilde{u}(x_i))^2}}{\sqrt{\sum_{i=1}^{n-1} (u^{(best)}(x_i))^2}} < 5 \times 10^{-2} \right) \text{ or}$$

(the generation for the second level evolution 20000).
As for the first level evolution, the terminate condition is

$$\left(\frac{\sqrt{\sum_{i=1}^{n-1} (u^{(best)}(x_i) - \tilde{u}(x_i))^2}}{\sqrt{\sum_{i=1}^{n-1} (u^{(best)}(x_i))^2}} < 10^{-3} \right)$$ or (the generation for the first level evolution 8).

3 Numerical Experiments

Example 1:

$$\begin{cases} \frac{d}{dx}(q(x)\frac{d}{dx}u(x)) = f(x), x \in (0,1) \\ u(0) = 0, u(1) = 0 \end{cases}$$

Where $q(x)=\begin{cases}\pi, 0\le x\le 0.5\\ 1, 0.5<x\le 1\end{cases}$, $u(x)=\begin{cases}x-\frac{1}{2}+\sin(\pi x), 0\le x\le 0.5\\ 1-\cos(\pi x), 0.5<x\le 1\end{cases}$

$\tilde{u}_i=(1+\delta_i)u(x_i)$, where δ_i $(i=0,1,...n-1)$ is a random number in $(-\delta, \delta)$

Test 1-1: δ =0, β =10^{-4}. Run time is 90 seconds. The times for evaluating the discontinuous point are 14. The times for evaluating q are 522151. The discontinuous point is 0.500000, $\frac{\sqrt{\sum_{i=1}^{n-1}(u^{(best)}(x_i)-\tilde{u}(x_i))^2}}{\sqrt{\sum_{i=1}^{n-1}(u^{(best)}(x_i))^2}}=0.000070$.

In the following Figure, the line —— presents the original function $q(x)$, while the line + + + presents the $q^*(x)$ identified by our algorithm.

Test 1-2: $\delta=1\%$, β =10^{-4}. Run time is 190 seconds. The times for evaluating the discontinuous point are 26. The times for evaluating q are 1042700. The non-continuous point is 0.500000, $\frac{\sqrt{\sum_{i=1}^{n-1}(u^{(best)}(x_i)-\tilde{u}(x_i))^2}}{\sqrt{\sum_{i=1}^{n-1}(u^{(best)}(x_i))^2}}=0.006393$.

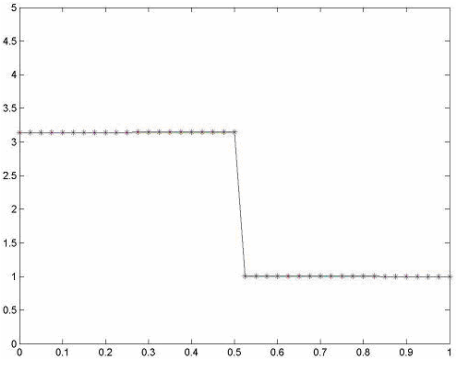


Fig. 1. Test 1-1(δ =0)

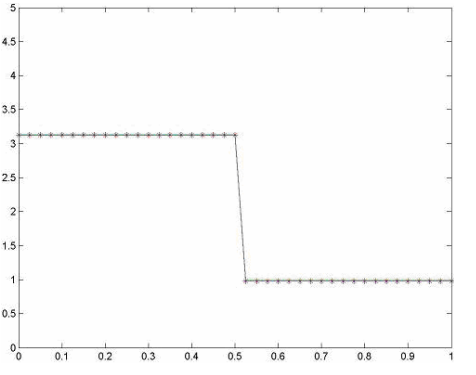


Fig. 2. Test 1-2($\delta=1\%$)

Test 1-3: $\delta=5\%$, β =10^{-3}. Run time is 197 seconds. The times for evaluating the discontinuous point are 26. The times for evaluating q are 1042700. The non-continuous point is 0.500000, $\frac{\sqrt{\sum_{i=1}^{n-1}(u^{(best)}(x_i)-\tilde{u}(x_i))^2}}{\sqrt{\sum_{i=1}^{n-1}(u^{(best)}(x_i))^2}}=0.027078$.

Test 1-4: $\delta=10\%$, β =10^{-3}. Run time is 192 seconds. The times for evaluating the discontinuous point are 26. The times for evaluating q are 1042700. The discontinuous point is 0.500000, $\frac{\sqrt{\sum_{i=1}^{n-1}(u^{(best)}(x_i)-\tilde{u}(x_i))^2}}{\sqrt{\sum_{i=1}^{n-1}(u^{(best)}(x_i))^2}}=0.045741$.

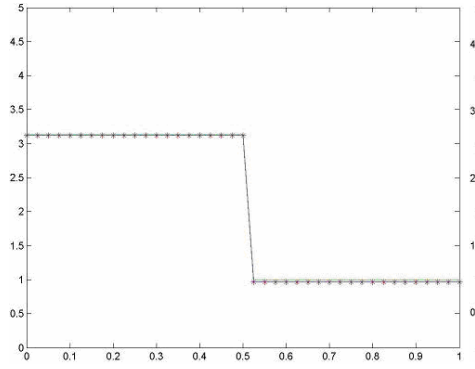


Fig. 3. Test 1-3($\delta = 5\%$)

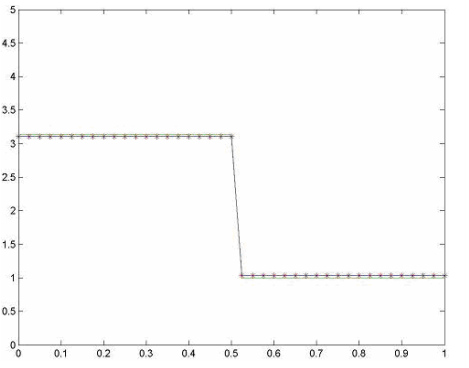


Fig. 4. Test 1-4($\delta = 10\%$)

Example 2:

$$\begin{cases} \dfrac{d}{dx}(q(x)\dfrac{d}{dx}u(x)) = f(x), x \in (0,1) \\ u(0) = 0, u(1) = 0 \end{cases}$$

Where $q(x) = \begin{cases} \sin(\pi x)+1, & 0 \le x \le 0.5 \\ 4x^2, & 0.5 < x \le 1 \end{cases}$, $u(x) = \begin{cases} \sin(\pi x)+2, & 0 \le x \le 0.5 \\ 2\sin(\pi x)+1, & 0.5 < x \le 1 \end{cases}$

$\tilde{u}_i = (1+\delta_i)u(x_i)$, where δ_i $(i = 0,1,...n-1)$ is a random number in $(-\delta, \delta)$

Test 2-1: $\delta = 0$, $\beta = 10^{-7}$. Run time is 116 seconds. The times for evaluating the discontinuous point are 12. The times for evaluating q are 481300. The discontinuous point is 0.500000, $\dfrac{\sqrt{\sum_{i=1}^{n-1}(u^{(best)}(x_i) - \tilde{u}(x_i))^2}}{\sqrt{\sum_{i=1}^{n-1}(u^{(best)}(x_i))^2}} = 0.000171$.

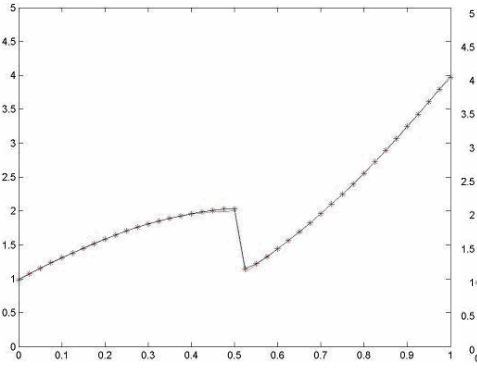


Fig. 5. Test 2-1($\delta = 0$)

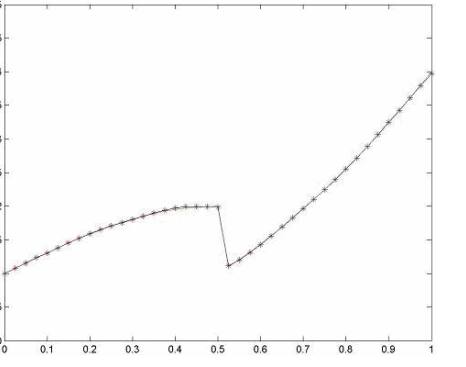


Fig. 6. Test 2-2($\delta = 1\%$)

Test 2-2: $\delta = 1\%$, $\beta = 10^{-7}$. Run time is 145 seconds. The times to evaluate the discontinuous point are 12. The times to evaluate q are 481300. The non-continuous point is 0.500000, $\frac{\sqrt{\sum_{i=1}^{n-1}(u^{(best)}(x_i)-\tilde{u}(x_i))^2}}{\sqrt{\sum_{i=1}^{n-1}(u^{(best)}(x_i))^2}} = 0.00145$.

Test 2-3: $\delta = 5\%$, $\beta = 10^{-7}$. Run time is 258 seconds. The times for evaluating the discontinuous point are 26. The times for evaluating q are 1042700. The non-continuous point is 0.525000, $\frac{\sqrt{\sum_{i=1}^{n-1}(u^{(best)}(x_i)-\tilde{u}(x_i))^2}}{\sqrt{\sum_{i=1}^{n-1}(u^{(best)}(x_i))^2}} = 0.006168$.

Test 2-4: $\delta = 10\%$, $\beta = 10^{-7}$. Run time is 621 seconds. The times for evaluating the discontinuous point are 38. The times for evaluating q are 2343200. The non-continuous is 0.475000, $\frac{\sqrt{\sum_{i=1}^{n-1}(u^{(best)}(x_i)-\tilde{u}(x_i))^2}}{\sqrt{\sum_{i=1}^{n-1}(u^{(best)}(x_i))^2}} = 0.009756$.

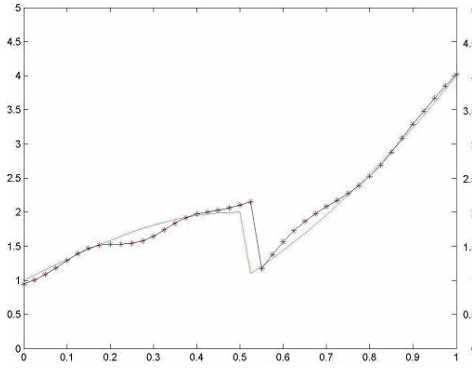


Fig. 7. Test 2-3($\delta = 5\%$)

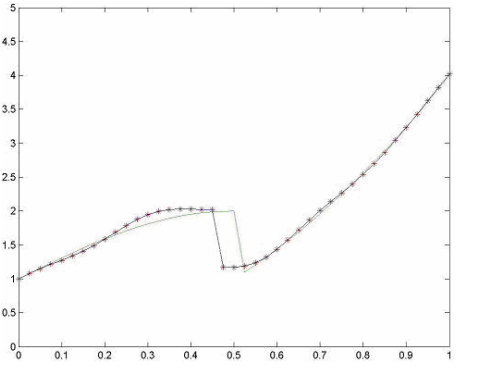


Fig. 8. Test 2-4($\delta = 10\%$)

From the above described numerical experiments It can be seen that the parameter function found by our algorithm is very approximate to the original parameter. All numerical experiments have been conducted on PC

4 Conclusions

The results of our experiments suggest that the algorithm presented in this paper has a good appearance for inverse problems in which the coefficient $q(x)$ has discontinuous point. And what's more, the numerical experiments demonstrate that our algorithm carries the feature of prefect stability and adaptability and that it is not very sensitive to the noise, which tends to be the most important factors to parameter identification in the inverse problems.

In is a new challenge to identify the discontinuous parameter by evolutionary algorithm. Only a simple case has been discussed in this paper, so in the further research, we will investigate some other kinds of inverse problems and model the discontinuous parameters by GP as well.

Acknowledgments. This research work was supported by Natural Science Foundation of Hubei Province (No. 2005ABA239)

References

1. Guo, B., and Zou, J. An augmented Lagrangian method for parameter identifications in parabolic systems. Journal of Mathematical Analaysis and Applications, 263 (2001) 49-68
2. Ito, K., and Kunisch, K. The augmented Lagrangian method for parameter estimation in elliptic system. SIAM J. Control Optim., 28(1990) 113-136
3. Keung, Y. L., and Zou, J. Numerical identifications of parameters in parabolic systems. Inverse Problems, 14 (1998) 83-100
4. Keung, Y. L., and Zou, J. An efficient linear solver for nonlinear parameter identification problems. SIAM J. Sci. Comput., 22, 5 (2000) 1511-1526
5. Xie, J., and Zou, J. Numerical reconstruction of heat fluxes. SIAM J. Numer. Anal. 43 (2005) 1504-1535
6. Burczynski, T., Beluch, W., Dlugosz, A., Orantek, P., and Nowakowski, M. Evolutionary methods in inverse problems of engineering mechanics. ISIP 2000 International Symposium on Inverse Problems in Engineering Mechanics (2000)
7. Collet, P., Lution, E., Raynal, F., and Schoenauer, M. Polar IFS+Parisian genetic programming=efficient IFS inverse problems. Genetic Programming and Evolvable Machines, 1 (2000) 339-361
8. Wu, Z., Tang, Z., Zou, J., Kang, L., and Li, M. Evolutionary algorithm for solving parameter identification problems in elliptic systems. In Proceedings of 2004 Congress on Evolutionary Computation, USA (2004) 803-808
9. Wu, Z., Tang, Z., Zou, J., Kang, L., and Li, M. An evolutionary algorithm for parameters identification in parabolic systems. In Proceedings of 2004 Genetic and Evolutionary Computation Conference, USA (2004) 1336-1337
10. Wu, Z., Kang L., Zou J., Tang Z., and Li, M. An evolutionary algorithm for identifying parameters in parabolic systems. In Progress in Intelligence Computation & Applications, China, Wuhan (2005) 92-96
11. Cao, H., Kang, L. and Chen, Y. Evolutionary modeling of system of ordinary differential equations with genetic programming. Genetic Programming and Evolvable Machines, 1, 4(2000) 309-337
12. Koza, J. GeneticPprogramming: On the Programming of Computers by Means of Natural Selection. Cambridge, MA: MIT Press (1992)
13. Xiong, S., and Li Y. An evolutionary modeling approach of partial differential equations. Wuhan University Journal of Natural Sciences, 10, 5 (1999) 767-770
14. Xiong, S., and Lu X. A genetic programming approach to partial differential equation inverse problems. Journal of Wuhan University of Technology (Information & Management Engineering), 25, 3 (2003) 11-15
15. Chen, Z., and Zou, J. An augmented Lagrangian method for identifying discontinuous parameters in elliptic problems. SIAM J. Control Optim., 37, 3(1999) 892-910

A WSN Coalition Formation Algorithm Based on Ant Colony with Dual-Negative Feedback

Na Xia, Jianguo Jiang, Meibin Qi, Chunhua Yu, Yue Huang, and Qi Zhang

School of Computer and Information,
Hefei University of Technology,
Hefei, 230009 PR China
jgjiang@hfut.edu.cn

Abstract. In large-scale, complicated Wireless Sensor Networks, the cooperation among sensor nodes is a key topic, and has been receiving more and more attention. The dynamic coalition mechanism in Multi-Agent System is an important method for this topic, and then an energy-efficient coalition formation algorithm is needed since the energy resource of WSN is restricted. This paper proposes a WSN coalition formation algorithm based on Ant Colony System with dual-negative feedback characteristic. Coalition sponsor inclines to wake those sensor nodes which seldom or never joined coalitions before to form coalition, which can balance the energy consumption among sensor nodes, so as to extend the network lifetime for more tasks. The results of simulation experiment show the validity of this algorithm.

Keywords: Wireless Sensor Networks (WSN), Cooperation, Coalition, Ant Colony System.

1 Introduction

Wireless sensor networks (WSN) is garnering a lot of research interets due to its important applications such as battlefield surveillance, environment monitoring, home security and target tracking, etc [1,2]. Due to the severe constraints of energy, computation and communication, sensor nodes have to cooperate with each other to perform tasks. So the cooperation among sensor nodes is a key topic. As the similar characeristics of distribution, autonomy and self-organization between Muti-Agent System (MAS) and WSN, the MAS theory has become an important method for this topic [2,3,4]. L. K. Soh and C. Tsatsoulis [5,6,7] investigated the WSN cooperation problem adopting dynamic coalition method in MAS. A dynamic sensor node coalition is formed for a task, and the coalition will dismiss after the task is completed. The coalition formation algorithm is based on case-based reasoning (CBR). In the course of coalition formation, the coalition sponsor describes the current task in parameters, searchs in the case database and finds the most similar case according to the task description. Then the corresponding strategy is taken out and modified to suit the current task.

CBR based coalition formation algorithm reduces computation resource spending to a certain extent, but it brings serious disadvantage to the energy

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 1139–1146, 2007.
© Springer-Verlag Berlin Heidelberg 2007

usage of WSN in other ways: those sensor nodes in successful case would be chosen repeatedly to form coalition to perform tasks, so their limited battery energy will be consumed up too early. The death of these sensor nodes will affect the function of WSN and shorten the network lifetime. This paper proposes a WSN coalition formation algorithm with dual-negative feedback characteristic. It can balance the energy consumption among sensor nodes, so as to extend the every sensor node and network lifetime for more tasks.

2 The DN-AC Algorithm

Ant Colony Optimization (ACO) [8,9,10] is inspired by social ants' collective behavior. In Ant Colony System, artificial ants explore solution space, and the transition probability depends on pheromone intensity and heuristic information. The pheromone accumulation (positive feedback) leads ants to consistent good solution, whereas the pheromone evaporation (negative feedback) is helpful to avoid early convergence to suboptimal solution. Based on the feedback mechanism in Ant Colony System, we propose a dual-negative feedback for WSN coalition formation, and design the algorithm for target tracking application. According the movement of target, the algorithm can wake certain sensor nodes dynamically to form coalition to track the target. The tracking accuracy and energy-load balancing are the primary performance metrics for the algorithm.

2.1 Algorithm

Define the surveillance zone as F, Consider n sensor nodes of WSN randomly scattered in F. The sensor node set can be expressed as $A = \{A_1, \cdots, A_n\}$, and the position set of sensor nodes can be expressed as $L = \{L_1, \cdots, L_n\}$, where L_k represents the location of A_k.

Define the current energy status of A_k as pheromone τ_k, $\tau_k \in [0, 1]$, which can be expressed by the residual energy percentage of A_k.

When a target O enters F, the sensor node A_i which first detects the target will become coalition sponsor $A_{sponsor}$. It will wake certain sensor nodes to form coalition for tracking mainly considering their pheromone value (energy status) and the distance from these sensor nodes to the target. The waking probability of A_k at time t is calculated as follows:

$$p_k(t) = \begin{cases} [\tau_k]^{\alpha} \left[\frac{R_c - \|L_k - l_o(t)\|}{R_c}\right]^{\beta} & \|L_k - l_o(t)\| \le R_c \\ 0 & otherwise \end{cases} \quad k = 1, \cdots, n \tag{1}$$

Here, $l_o(t)$ is the located position of target O at time t. R_c is the detection radius of sensor node. It is evident that larger τ_k and smaller distance means that the sensor node has more chance to be wakened.

Wake sensor nodes stochastically according to $p_k(t)$. The wakened sensor nodes form coalition to track the target for Δt period, and then update the waking probability $p_k(t)$ using the newly located position of target O, so as to

execute the next waking operation. With the target moving, the sensor node coalition is dynamic in the surveillance zone. While target moves out of F, a round of target tracking is over.

The wakened/sleeping state of sensor node A_k at time t can be represented by the indicator $\theta_k(t)$ as follows:

$$\theta_k(t) = \begin{cases} 1 \ A_k \ is \ wakened \ at \ time \ t \\ 0 \ A_k \ is \ sleeping \ at \ time \ t \end{cases} \quad k = 1, \cdots, n \tag{2}$$

The integral working time of sensor node A_k in dynamic coalition during a round of target tracking can be represented as follows:

$$T_k = \sum_t \theta_k(t) \ \ k = 1, \cdots, n \tag{3}$$

After a round of target tracking, the pheromone value on every sensor node will be updated. For those sensors which participated in dynamic coalition, besides pheromone ***evaporation***, their pheromone will be ***decreased further***, which is we termed ***Dual-Negative feedback***. The updated pheromone value for A_k is calculated as follows:

$$\tau_k \leftarrow \rho \cdot \tau_k - \Delta\tau_k \ \ k = 1, \cdots, n \tag{4}$$

Here $1 - \rho \in (0, 1)$ is the decay coefficient which represents the energy consumption under sleeping model. The pheromone decrease $\Delta\tau_k$ represents the energy consumption of sensor node A_k during this round of target tracking, which is as follows:

$$\Delta\tau_k = \begin{cases} \xi \cdot \frac{T_k}{\sum_k T_k} \ if \ A_k \ participate \ in \ dynamic \ coalition \\ 0 \ otherwise \end{cases} \quad k = 1, \cdots, n \tag{5}$$

$\Delta\tau_k$ is direct proportion to T_k. Define the pheromone decrease coefficient as ξ, which can be adjusted.

Consequently, when next target appears, it is less possible for $A_{sponsor}$ to wake the sensor nodes which participated in coalition before, whereas $A_{sponsor}$ inclines to wake those sensor nodes which seldom or never joined coalition to form coalition. This behavior balances energy-load among sensor nodes, and network lifetime extending is achieved.

To guarantee the tracking reliability and avoid waste of resource, we define the minimal node number $w_{\min}$ ($w_{\min} \geq 3$) and maximal node number $w_{\max}$ for a dynamic coalition. After $NC_{\max}$ rounds of target tracking, query network, and update $\tau_k(t)$ using the actual energy status of each sensor. This operation is good to the correctness of latter decision for algorithm.

We term this dynamic coalition formation algorithm based on Ant Colony with Dual-Negative feedback as **DN-AC**, and its pseudocode is as follows:

- Step1. Initialize: Set initial value $\tau_k = 1$ (100%).
- Step2. Set $NC = 0$.

- Step3. Set $t = 0, \theta_k(t) = 0, T_k = 0, \Delta\tau_k = 0$.
- Step4. If (O is detected by A_i) then
  $\{A_i$ becomes $A_{sponsor}, l_o(t) = L_i$;
  $A_{sponsor}$ wakes certain sensor nodes to form coalition according to $p_k(t)$ by (1), update $\theta_k(t)$, the coalition tracking O for Δt period, Set $t = t + \Delta t$, update $l_o(t)\}$
  Else wait.
- Step5. While ($l_o(t) \in F$)
  $\{A_{sponsor}$ wakes certain sensor nodes to form coalition according to $p_k(t)$ by (1), and selects the sensor node with maximal $p_k(t)$ value as new $A_{sponsor}$, update $\theta_k(t)$, the coalition tracking O for Δt period, $t = t + \Delta t$, update $l_o(t)\}$.
- Step6. Output $l_o(t), t = 0, \Delta t, 2\Delta t, 3\Delta t, \cdots$.
- Step7. For $k = 1$ to n do
  Update the pheromone τ_k by (3), (4), (5).
- Step8. Set $NC = NC + 1$.
- Step9. If ($NC < NC_{\max}$) then Goto step 3
  Else query network, update the τ_k using the actual energy status of each sensor node, Goto step 2.

2.2 Algorithm Evaluation

The performance metrics being investigated are:

(1) *Tracking error of target O*

Define the tracking error of target O as follows:

$$\varphi(O) = \overline{|l_o(t) - l_o^*(t)|} \tag{6}$$

Where $l_o(t)$ is the target position located by the sensor node coalition at time t, while $l_o^*(t)$ is the actual position of target at time t. So, $\varphi(O)$ represents the accuracy of target tracking.

(2) *Mean square error of network energy*

Define the mean square error of network energy at time t as follows:

$$\sigma(t) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\tau_i - \overline{\tau})^2} \tag{7}$$

Here $\tau_1, \cdots, \tau_n$ are the pheromone values of n sensor nodes at time t. This metric represents the status of network energy. The smaller the value is the more energy load-balanced the network achieves.

(3) *Health degree of network*

Define the total number of living sensor nodes (i.e. not run out of energy) at time t as $n_{living}(t)$. The health degree of network at time t can be defined:

$$H(t) = \frac{n_{living}(t)}{n} \tag{8}$$

This metric represents the sensor protection performance of the algorithm. Otherwise, the network lifetime T can be defined as the time period from the instant the network is deployed to the moment when 30% of all sensor nodes run out of energy.

3 Simulation Experiment

Fig.1 depicts the surveillance zone, which is a rectangle from (0m, 0m) to (200m,200m). There are 500 sensor nodes shown by dots randomly scattered in the zone. In turn, the targets O_1, O_2, O_3, O_4, O_5 and O_6 move from the same start point $(0, 0)$ with constant velocity but different angles $35°, 38.7°, 42°, 45°, 47.7°$ and $50.2°$ to cross the surveillance zone within $40s$. These targets' trajectory is represented by dashed line.

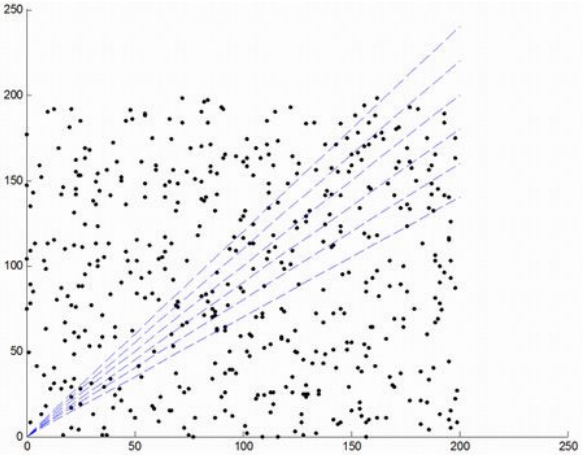


Fig. 1. Surveillance zone

We adopt CBR method and DN-AC algorithm to form dynamic coalition for target tracking respectively, and compare 50 times Monte Carlo simulation results. The parameters of DN-AC are shown in Table 1.

For DN-AC, it forms dynamic coalition to track O_1, and the tracking result is depicted in Fig.2. The red circles represent the sensor nodes wakened

Table 1. Parameters setting

R_c	Δt	τ_k	α	β	ρ	ξ	$w_{\min}$	$w_{\max}$
$40m$	$2s$	1	1	1	0.99	20	3	5

to form dynamic coalition during this round of target tracking, moreover the continuous green line represents the trail gained by these sensor nodes locating O_1. The tracking error of this round is $1.19m$. In Fig.3, the tracking errors of O_1, O_2, O_3, O_4, O_5 and O_6 for CBR and DN-AC are compared. The result shows the tracking accuracy of the two algorithms is similar.

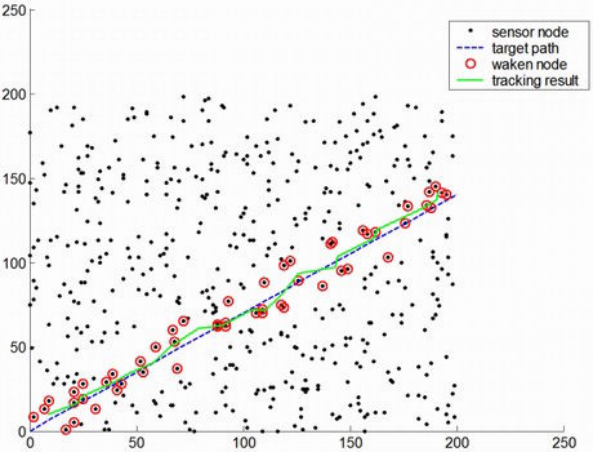


Fig. 2. Tracking result to O_1 using DN-AC, $\varphi(O_1) = 1.19m$

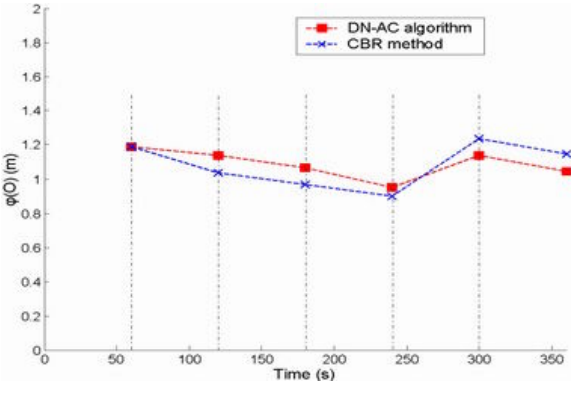


Fig. 3. The comparison of tracking errors of 6 targets between two algorithms

During the 6 rounds of target tracking, the variance of the mean square error of network energy $\sigma(t)$ and the health degree of network $H(t)$ for two algorithms are depicted in Fig.4 and 5. In Fig.4, for CBR, the $\sigma(t)$ value rises rapidly because CBR method results in some sensor nodes running out of energy too early and some other nodes consuming energy badly, so that the energy level of

network is rather separate. Since DN-AC can balance the energy consumption among sensor nodes, the $\sigma(t)$ value rises more slowly than CBR. In Fig.5, It is evident that DN-AC exhibits better network health degree than CBR. It is found that there are 30 sensor nodes dead after the 6th round of target tracking for CBR (shown in Fig.6), whereas the number is 15 for DN-AC (shown in Fig.7). Therefore DN-AC achieves significantly better energy-load balancing performance than CBR, so as to extend the network lifetime effectively.

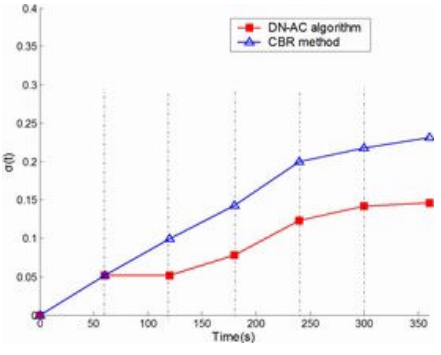


Fig. 4. The variance of mean square error of network energy for two algorithms

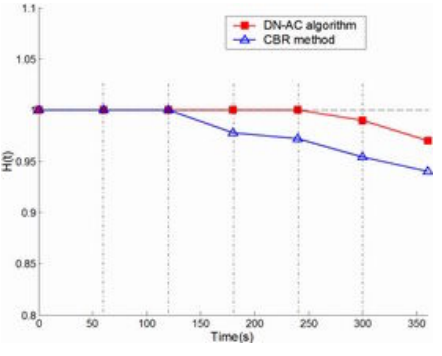


Fig. 5. The variance of health degree of network for two algorithms

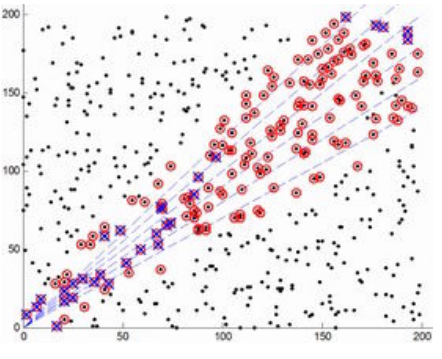


Fig. 6. The status of dead sensor nodes after the 6th round of target tracking for CBR, the crosses representing the dead sensor nodes

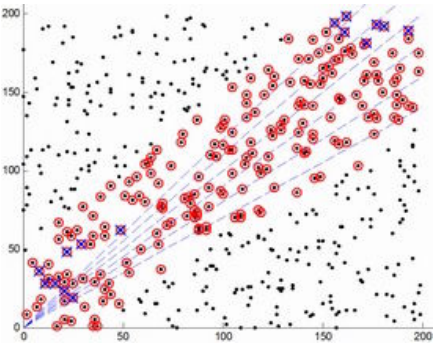


Fig. 7. The status of dead sensor nodes after the 6th round of target tracking for DN-AC, the crosses representing the dead sensor nodes

4 Conclusions

This paper investigates the energy-efficient dynamic coalition formation algorithm in WSN, and proposes a coalition formation algorithm based on Ant

Colony System with dual-negative feedback characteristic. With coalition quality guaranteed, it can balance the energy consumption among sensor nodes, so as to extend the network lifetime for more tasks. The results of simulation experiment show the validity of this algorithm for target tracking application.

Acknowledgments. This work is supported by NSFC, National Natural Science Foundation of China, No. 60474035, and Natural Science Foundation of Anhui Province, No.070412035.

References

1. L. F. Akyildiz, W. Su, Y. Survey on Sensor Networks. IEEE Communication Magazine, 2002, 40(8): 102–114.
2. H. B. Yu, P. Zeng, W. Liang. Intelligent Wireless Sensor Networks. Beijing: Science Press, 2006. 215–278.
3. C. Che, W. Liang, Y. Zhou, et al. Cooperation Problem of Wireless Sensor Network Based on Multi-agent. Chinese Journal of Scientific Instrument, 2005, 26(8A): 229–232.
4. V. Lesser, C. L. Ortiz, M. Tambe. Distributed Sensor Networks: A Multiagent Perspective. Kluwer Academic Publishers, 2003.
5. L. K. Soh, C. Tsatsoulis. Real-time Satisfying Multiagent Coalition Formation. In Working Notes of AAAI Workshop on Coalition Formation in Dynamic Multiagent Environment. 2002. 7–15.
6. L. K. Soh, X. Li. An Integrated Multi-Level Learning Approach to Multiagent Coalition Formation. In Proceedings of IJCAI'03. Acapulco, Mexico, 2003. 619–624.
7. L. K. Soh, C. Tsatsoulis. Satisfying Coalition Formation among Agents. In Proceedings of AAMAAS'02. Bologna, Italy, 2001. 15–19.
8. M. Dorigo, T. Stutzle. Ant colony optimization. MIT Press, U. S., 2004.
9. R. Wang, Y. Liang, G. Q. Ye, et al. Swarm Intelligence for the Self-Organization of Wireless Sensor Network. In Proceedings of the IEEE Congress on Evolutionary Computation. Vancouver, BC, Canada, 2006. 838–842.
10. S. Selvakennedy, S. Sinnappan and Y. Shang. T-ANT: A Nature-Inspired Data Gathering Protocol for Wireless Sensor Networks. Journal of Communications, 2006, 1(2): 22–29.

An Improved Evolutionary Algorithm for Dynamic Vehicle Routing Problem with Time Windows

Jiang-qing Wang, Xiao-nian Tong, and Zi-mao Li

College of Computer Science, South-central University for Nationalities, Wuhan, 430074, China

Abstract. The dynamic vehicle routing problem is one of the most challenging combinatorial optimization tasks. The interest in this problem is motivated by its practical relevance as well as by its considerable difficulty. We present an approach to search for best routes in dynamic network. We propose a dynamic route evaluation model for modeling the responses of vehicles to changing traffic information, a modified Dijkstra's double bucket algorithm for finding the real-time shortest paths, and an improved evolutionary algorithm for searching the best vehicle routes in dynamic network. The proposed approach has been evaluated by simulation experiment using DVRPSIM. It has been found that the proposed approach quite efficient in finding real-time best vehicle routes where the customer nodes and network information changes dynamically.

Keywords: Combinatorial Optimization, Dynamic Vehicle Routing Problem, Dijkstra Algorithm, Evolutionary Algorithm.

1 Introduction

The Vehicle Routing Problem (VRP) has been largely studied because of the interest in its applications in logistic and supply-chains management[1,2,3,4,5]. The VRP can be classified into two categories: static and dynamic [6]. The Dynamic Vehicle Routing Problem (DVRP) [7,8] is a richer problem compared to the static ones [9]. It not only involves increasing the problem size as new customer nodes enter the network but also changing in dynamic network, and highly sensitive to real-time traffic information. Most of current heuristic algorithms developed for the DVRP consider static traffic information, the travel time between customer nodes depend on distances only [10,11,12,13]. However, the best path from any given node to the final destination depends not only on the node but also on the arrival time at that node [14]. Some of the researches have been concentrated on the development of real-time travel time functions [15,16,17]. In these approaches, traffic congestion and random fluctuation of traffic flow are not reflected.

The problem considered in this paper is the Dynamic Vehicle Routing Problem with Time Window (DVRPTW). We present an approach to search for best routes in dynamic network. The approach considers route attributes, real-time

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 1147–1154, 2007.
© Springer-Verlag Berlin Heidelberg 2007

traffic information and dynamic demand information simultaneously, and can find the best vehicle routes for the DVRPTW.

The rest of the paper is organized as follows. Section 2 proposes the mathematical model of the DVRPTW. Section 3 develops a new route evaluation model. Using this model, a modified Dijkstra's double bucket algorithm is presented. Section 4 designs an improved evolutionary algorithm for the DVRPTW. Section 5 estimates the value of the developed approach. In Section 6, we discuss conclusions.

2 Mathematical Model of the DVRPTW

The DVRPTW is given by a set of vehicles K, a special node called the depot, a set of customer nodes V, and a network connecting the depot and customers. For simplicity, we denote depot as customer 0. Since each vehicle has a limited capacity Q_k, and each customer has a demand q_i, Q_k must be greater than or equal to the summation of all demands on any route. Any customer i must be serviced within a pre-defined time window $[T_{start_i}, T_{end_i}]$. Let us assume that T_i is the arrival time of customer i, $T_{i,j}$ is the travel time between customer i and j, $\alpha_i (i = 1, 2, 3, 4)$ is the penalty coefficient, T is the end time period. We trade off the vehicle number, travel time, wait cost of vehicles and wait cost of customers in the objective function of the DVRPTW, which can be formulated as the following.

$$\min(\alpha_1 K + \sum_{k \in K} (\alpha_2 \sum_{p=1}^{m_k} T_{i^t_{p-1}, i^t_p} + \alpha_3 \sum_{p=0}^{m_k} (T_{start_{i^t_p}} - T_{i^t_p})^+ + \alpha_4 \sum_{p=0}^{m_k} (T_{i^t_p} - T_{end_{i^t_p}})^+)) \tag{1}$$

Subject to

$$i^t_{m_k} = 0, \forall k \in K \tag{2}$$

$$\sum_{j \in V} \sum_{t=0}^{T} x^t_{0j} = \sum_{j \in V} \sum_{t=0}^{T} x^t_{jo} = K \tag{3}$$

$$\sum_{j \in V} \sum_{t=0}^{T} x^t_{ij} = \sum_{j \in V} \sum_{t=0}^{T} x^t_{ji} = 1, i \in \{V - \{0\}\} \tag{4}$$

$$\sum_{p=0}^{m_k} q_{i_p} \le Q_k, \forall k \in K \tag{5}$$

Where:

$\{i^t_0, i^t_1, \ldots, i^t_{m_k}\}$: the route of vehicle k at time t,

$$x^t_{ij} = \begin{cases} 1, \text{ if any vehicle departures from customer i to j at time t} \\ 0, \text{ otherwise} \end{cases}$$

$$(x - y)^+ = \max\{0, x - y\}.$$

3 Real-Time Shortest Path in Dynamic Network

3.1 Dynamic Route Evaluation Model

A number of practical traffic information is selected, including real-time traffic information and route attributes, as multiple criteria for the developed dynamic route evaluation model.

1. Route length: the physical distance of the route.
2. Route width: the number of lanes in the route. The optimal speed of vehicles, Speed $_{optimal}$, is based on lanes in the route when they travel along this route.
3. Route difficulty: turning movement is selected to measure the route difficulty.
4. Actual speed of vehicles on one route:

$$Speed_{\ actual} = \alpha Speed_{optimal} \quad \alpha \in [0,1]$$

Where α is based on route difficulty, accident, traffic congestion, weather conditions, etc.

We use these multiple criteria to evaluate real-time travel times for each route in dynamic network, which can be categorized by several criteria. Firstly, it may be classified into two types: static and dynamic, according to how it is defined with respect to time. Secondly, it may be classified into two types: stochastic and deterministic, according to whether it is random variables or not. Lastly, it may also be classified into two value types: crisp and fuzzy (Table 1).

Table 1. Multiple Criteria for Route Evaluation

Criteria	Unit	Certainty	Variability	Measurement
Route length	*km*	DT	ST	Crisp
Route width	NM	DT	ST	Crisp
Route difficulty	NM	DT	ST	Crisp
Accidents	Level	SC	DN	Fuzzy
Traffic congestion	Level	SC	DN	Fuzzy
Weather conditions	Level	SC	DN	Fuzzy

NM: normalized, DT: deterministic, SC: stochastic, ST: static, DN: dynamic
Real-time travel time on this route:

$$Travel\ time = Route\ length\ /\ Speed_{\ actual}$$

3.2 Real-Time Shortest Path Algorithm

Here, we develop a modification of Dijkstra's double bucket algorithm for path finding in dynamic network. In the developed algorithm, the length of a path is defined to be the travel time on this path. In the problem we considered, a

directed graph is given, G=(V, E), where V represents nodes and E represents the route between two nodes. For each node $v \in V$, it is assigned a potential $d(v) \geq 0$, where in this case representing the current shortest time to the source node. For each edge $e \in E$, it is assigned a cost function $c(e) \geq 0$, representing current travel time between two nodes connected by the edge. The length of a path is now defined to be the summation of the cost of the edges on that path. There may be impedance on an edge corresponding to some traffic limitations, such as accidents, traffic congestion, and weather conditions. So $c(e)$ and $d(v)$ are time-varying according to real-time traffic information in a day. Other variables used by the algorithm are number of nodes, edges, low level buckets and largest normal travel time along a single edge which are denoted as n, m, B, C respectively.

The developed algorithm calculates the travel time between nodes when rerouting request is accepted. The output of the algorithm is the real-time shortest paths between nodes at that time. The time taken by the algorithm on a graph with n vertices is $O(m + n(B + C/B))$, and by the standard Dijkstra's algorithm, it is $O(n^2)$. This characteristic is very important for the DVRP.

4 Improved Evolutionary Algorithm (IEA) for the DVRPTW

4.1 Representation

We use the modified form of random keys representation [18]. In our representation, a chromosome consists of genes and each gene represents a customer node. The customer nodes have fixed gene positions in the chromosomes and the order in which they are visited is determined by sorting on the gene values. The random keys have information about the vehicle number used for a service and the value for sorting, where the digit before the point represents the vehicle number and the digit after the point are used as sort keys to decode visiting sequence. For example, a chromosome to 8 customers problem may be:

1	2	3	4	5	6	7	8
1.324	2.315	2.761	2.189	1.436	1.847	1.875	1.104

The route for the previous chromosome can be represented as follows.

Vehicle$_1$: 8→1→5→6→7, Vehicle$_2$: 4→2→3

The steps for generating chromosome are as follows.

1. Generate the vehicle number.
2. Generate sorting number.
3. Combine the vehicle number and sorting number.

4.2 Handling of the Constraints

The initial population of chromosomes is generated randomly and may not satisfy the constraints of the proposed DVRPTW. And some new chromosomes generated after genetic operators (crossover, mutation) may not satisfy the constraints. So, the constraints-checking steps are executed after new chromosomes

being generated. We consider soft time windows in this paper, so we use the penalty method. The infeasible chromosomes have fewer opportunities than the feasible chromosomes, and have the chances to be turned to the feasible chromosomes by the genetic operator.

4.3 Crossover Operator

We use the two-points crossover operator. It is assumed that there are two chromosomes p_1, p_2 as follows and two generated crossover points are 2 and 5.

p_1: 1.324 2.315 2.761 2.189 1.436 1.847 1.875 1.104

p_2: 2.134 1.516 2.385 2.034 1.891 1.625 2.329 1.618

After the crossover operation, two children c_1 and c_2 are generated as follows.

c_1: 1.324 2.315 **2.385 2.034 1.891** 1.847 1.875 1.104

c_2: 2.134 1.516 **2.761 2.189 1.436** 1.625 2.329 1.618

4.4 Mutation Operator

The mutation operator changes the vehicle number with a newly generated vehicle number and does not change the information that is used as sorting key, because the sequence of the customer nodes can be changed by just changing the assigned vehicle number of a customer. Let us assume that there are one chromosomes p_1 as follows and the mutation point is 6.

p_1: 1.324 2.315 2.761 2.189 1.436 1.847 1.875 1.104

After the mutation operation, the children c_3 is generated as follows.

c_3: 1.324 2.315 2.761 2.189 1.436 **2**.847 1.875 1.104

5 Experimental Results and Analysis

In the literature, there is no commonly used benchmark for the DVRPTW, so the authors have generated their own Dynamic Vehicle Routing Problem SIMulator(DVRPSIM) to evaluate the benefits of the developed approach. The simulated system of the DVRPTW is made up of three modules, route evaluation module, shortest path module and routing plan module, as shown in Fig.1.

The main function of route evaluation module is to evaluate the actual travel times of vehicles on each route using the real-time traffic information and the route attributes, to transmit the result to the shortest path module to determine the real-time shortest path between customers. According to the result obtained from the shortest path module, routing plan module based on IEA can determine the best routes for vehicles whenever requested. Vehicles will use these routes and drive on them in the dynamic network. When the rerouting requests are accepted, the system will determines the real-time best routes for vehicles.

In order to compare IEA with other algorithms, we choose two famous algorithms, Branch-Bound algorithm and Clarke-Wright algorithm, as benchmarks. We listed the comparison results in Table 2.

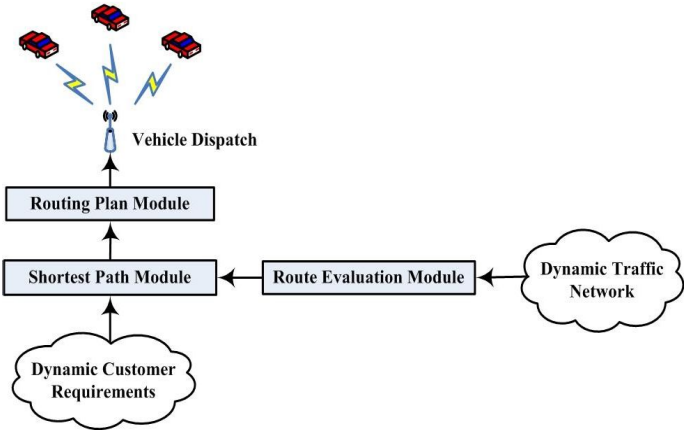


Fig. 1. Architecture of the Simulated System

Table 2. Results of Different Algorithms

	IEA	B-B	C-W
Number of vehicle	4	4	4
Route Cost	2808	2737	2907
Wait Cost of Vehicles	49	32	58
Wait Cost of Customers	27	33	53
Total Cost	5924	5749	6227
Calculation Time	63.117	512.096	50.673

Table 2 shows that, compared with B-B, the calculation time of IEA is much lower and the route cost is a little bigger. And compared with C-W, its route cost is better and the calculation time is similar. Fig. 2 shows the cost comparison of these algorithms, and Fig. 3 shows the time comparison of them.

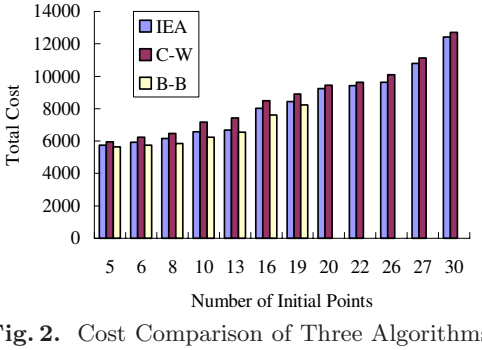


Fig. 2. Cost Comparison of Three Algorithms

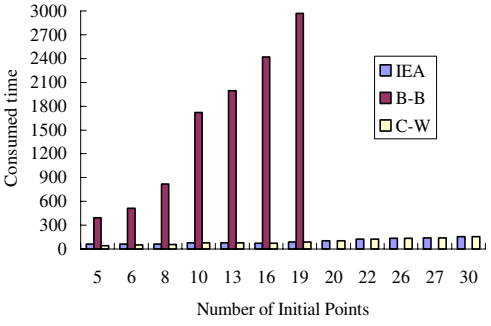


Fig. 3. Time Comparison of Three Algorithms

6 Conclusions

This paper presents an approach for the DVRPTW. We have proposed a dynamic route evaluation model to evaluate routes using route attributes and real-time traffic information. We have developed a modified Dijkstra's algorithm for finding real-time shortest paths in dynamic network. We have designed an improved evolutionary algorithm for searching the best vehicle routes of the DVRPTW. We have performed a simulation test using DVRPSIM. In the simulation test, we have compared three algorithms: IEA, B-B, and C-W. Our primary conclusion is that the developed approach based on IEA can find the best vehicle routes for the DVRPTW efficiently.

Acknowledgement. The authors gratefully acknowledge the financial support of the National Natural Science Foundation of China under Grant No.60603008.

References

1. Tighe, A., Smith, F.S., Lyons, G.: Priority based solver for a real-time dynamic vehicle routing. In: Systems, Man and Cybernetics, 2004 IEEE International Conference on. Volume 7. (2004) 6237–6242
2. Donati, A.V., Montemanni, R., Gambardella, L.M., Rizzoli, A.E.: Integration of a robust shortest path algorithm with a time dependent vehicle routing model and applications. In: Computational Intelligence for Measurement Systems and Applications, 2003. CIMSA '03. 2003 IEEE International Symposium on. (2003) 26–31
3. Tan, K.C., Lee, T.H., Chew, Y.H., Lee, L.H.: A multiobjective evolutionary algorithm for solving vehicle routing problem with time windows. In: Systems, Man and Cybernetics, 2003. IEEE International Conference on. Volume 1. (2003) 361–366
4. Tan, K.C., Lee, T.H., Ou, K., Lee, L.H.: A messy genetic algorithm for the vehicle routing problem with time window constraints. In: Evolutionary Computation, 2001. Proceedings of the 2001 Congress on. Volume 1. (2001) 679–686 vol. 1

5. Alvarenga, G.B., Mateus, G.R.: A two-phase genetic and set partitioning approach for the vehicle routing problem with time windows. In: Hybrid Intelligent Systems, 2004. HIS '04. Fourth International Conference on. (2004) 428–433
6. Psaraftis, H.N.: Dynamic vehicle routing: Status and prospects. Annals of Operations Research **61**(1995) (1995) 143–164
7. Alvarenga, G.B., de Abreu Silva, R.M., Mateus, G.R.: A hybrid approach for the dynamic vehicle routing problem with time windows. In: 5th International Conference on Hybrid Intelligent Systems (HIS 2005). (2005) 61–67
8. Song, J., Hu, J., Tian, Y., Xu, Y.: Re-optimization in dynamic vehicle routing problem based on wasp-like agent strategy. In: Intelligent Transportation Systems, 2005. Proceedings. 2005 IEEE. (2005) 231–236
9. Qiang, L.: Integration of dynamic vehicle routing and microscopic traffic simulation. In: Intelligent Transportation Systems, 2004. Proceedings. The 7th International IEEE Conference on. (2004) 1023–1027
10. Lou, S.Z., Shi, Z.K.: An effective tabu search algorithm for large- scale and real-time vehicle dispatching problems. In: Machine Learning and Cybernetics, 2005. Proceedings of 2005 International Conference on. Volume 6. (2005) 3579–3584
11. Del Bimbo, A., Pernici, F.: Distant targets identification as an on-line dynamic vehicle routing problem using an active-zooming camera. In: Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on. (2005) 97–104
12. Tian, Y., Song, J., Yao, D., Hu, J.: Dynamic vehicle routing problem using hybrid ant system. In: Intelligent Transportation Systems, 2003. Proceedings. 2003 IEEE. Volume 2. (2003) 970–974
13. Ce, F., Hui, W., Ying, Z.: Solving the vehicle routing problem with stochastic demands and customers. In: Parallel and Distributed Computing, Applications and Technologies, 2005. PDCAT 2005. Sixth International Conference on. (2005) 736–739
14. Seongmoon, K., Lewis, M.E., White, C. C., I.: Optimal vehicle routing with real-time traffic information. Intelligent Transportation Systems, IEEE Transactions on **6**(2) (2005) 178–188
15. Jung, S.J.: A Genetic Algorithm for the Vehicle Routing Problem with Time Dependent Travel Times. PhD thesis, University of Maryland, USA (2000)
16. Fischetti, M., Laporte, G., Mattello, S.: The delivery man problem and cumulative matroids. Operation Research **41** (1993) 1055–1076
17. Malandraki, C., Daskin, M.S.: Time dependent vehicle routing problems: Formulations, properties and heuristic algorithms. Transportation Science **26**(3) (1992) 185–200
18. Bean, J.: Genetic algorithms and random keys for sequencing and optimization. ORSA Journal on Computing **6**(2) (1994) 154–160

The Geometry Optimization of Argon Atom Clusters Using Differential Evolution Algorithm

Yongxiang Zhao, Shengwu Xiong, and Ning Xu

School of Computer Science and Technology, Wuhan University of Technology, Wuhan 430070, China
zhaosanhe@263.net, xiongsw@mail.whut.edu.cn

Abstract. Recently atomic cluster structures have been intensively studied because of their importance in physics, chemistry and material science. However, finding the lowest energy structure, which is the most stable configuration, is NP-hard. Differential Evolution (DE) algorithm is a new heuristic approach which mainly has three advantages: finding the true global minimum regardless of the initial parameter values, fast convergence, and using few control parameters. In this paper we describe a new search method that uses differential evolution (DE) algorithm to optimize the geometry of small argon atom clusters. Experimental results show that the exact global optimal configuration of argon clusters with atom number N≤ 16 can be found in a reasonable computing time, and approximate optimization can also be obtained for clusters with N=30. From their 3-D geometry structures, we can see that their optimal energy structures are highly symmetrical.

Keywords: Argon Atom Cluster, Structure Optimization, Genetic Algorithm, Differential Evolution.

1 Introduction

Recently atomic cluster structures have been intensively studied because of their importance in physics, chemistry and material science. One method of solving this optimization problem is to explore the potential energy surface (PES) composed of all possible cluster conformations. Unfortunately, as the cluster size increases, so does the number of degrees of freedom in the placement of the atoms. This characteristic produces a PES where the number of local optima grows exponentially with the cluster size [1]. Determining the ground-state energy level, which is the most energetically stable level, is extremely difficult. Wille and Vennik [2] proved this problem is NP-hard for homo-nuclear clusters (i.e., clusters with only one type of atom) and Greenwood [3] later proved the same thing for hetero-nuclear clusters.

Hence, heuristic search techniques are widely used in this area, such as Genetic algorithms [4,5], Simulated Annealing [6].

Differential Evolution (DE) algorithm [7,8] is a new heuristic approach which mainly has three advantages: finding the true global minimum regardless of the initial parameter values, fast convergence, and using few control parameters. In this paper

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 1155–1158, 2007.
© Springer-Verlag Berlin Heidelberg 2007

we describe a new search method that uses differential evolution (DE) algorithm to optimize the geometry of small argon atom clusters. Experimental results show that the exact global optimal configuration of argon clusters with atom number N≤ 16 can be found in a reasonable computing time, and approximate optimization can also be obtained for clusters with N=30.

The paper is organized as follows. An overview of atomic clusters is provided in Section 2. Section 3 reviews the DE approach and shows how it is used to search for low energy conformations. Section 4 presents some experiments conducted with small clusters of argon atoms. Finally, Section5 concludes the paper with summary and future research directions.

2 Homo-nuclear Clusters and Potential Energy Functions

The objective is to find the lowest energy conformation because it is the most stable conformation. In this work, we approximate the total energy as the sum of all the pairwise interactions between atoms:

$$E = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} v(r_{ij}) \tag{1}$$

Where r_{ij} is the Euclidean distance between atoms i and j, and $v(r_{ij})$ is the pairwise potential energy function.

A commonly used pairwise potential energy function for (1) is the Lennard-Jones potential function [4]:

$$v(r_{ij}) = 4.0\varepsilon \left[\left(\frac{\sigma}{r_{ij}} \right)^{12} - \left(\frac{\sigma}{r_{ij}} \right)^{6} \right] \tag{2}$$

$$r_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2} \tag{3}$$

where $\varepsilon = 1$, $\sigma = 3.36$ Å, $x_i, y_i, z_i \in [-1/2 \cdot \sqrt[3]{6N} \cdot \sigma, 1/2 \cdot \sqrt[3]{6N} \cdot \sigma]$.

3 Using DE in Cluster Searches

The DE algorithm is a population based algorithm like genetic algorithms using the similar operators: crossover, mutation and selection [7]. The main difference in constructing better solutions is that genetic algorithms rely on crossover while DE relies on mutation operation. This main operation is based on the differences of randomly sampled pairs of solutions in the population.

The algorithm uses mutation operation as a search mechanism and selection operation to direct the search toward the prospective regions in the search space. The DE algorithm also uses a non-uniform crossover that can take child vector parameters from one parent more often than it does from others. By using the components of the existing population members to construct trial vectors, the recombination (crossover) operator efficiently uses information about successful combinations, enabling the search for a better solution space.

An optimization task consisting of D parameters can be represented by a D-dimensional vector. In DE, a population of NP solution vectors is randomly created at the start. This population is successfully improved by applying mutation, crossover and selection operators. The details of the mutation, crossover and selection operators were described in literature [7].

4 Experimental Results

These experiments were all executed under Microsoft Windows Server 2003 with 3.00GHz of Intel(R) Pentium(R) 4 CPU and 512 MB of RAM. Each experiment was executed 10 times, and the parameters of the DE algorithm [7] in the experiments were as follows:

Dimensions D = 3N, where N is argon atom number, the number of population NP = 30, the scale factor F = 0.35, the crossover rate CR = 0.2, the maximum generation is 100000, and we use the strategy *DE/rand-to-best/1/exp*.

The optimized energy results for argon atom clusters are shown in Table 1. Experimental results show that the exact global optimal configuration of argon clusters with atom number N≤ 16 can be found in a reasonable computing time, and approximate optimization can also be obtained for clusters with N=30.

The optimized geometry structures for argon atom clusters (N= 12, 14, 16 and 30) are shown in Fig. 1. From the pictures, we can see that their geometry structures are highly symmetrical.

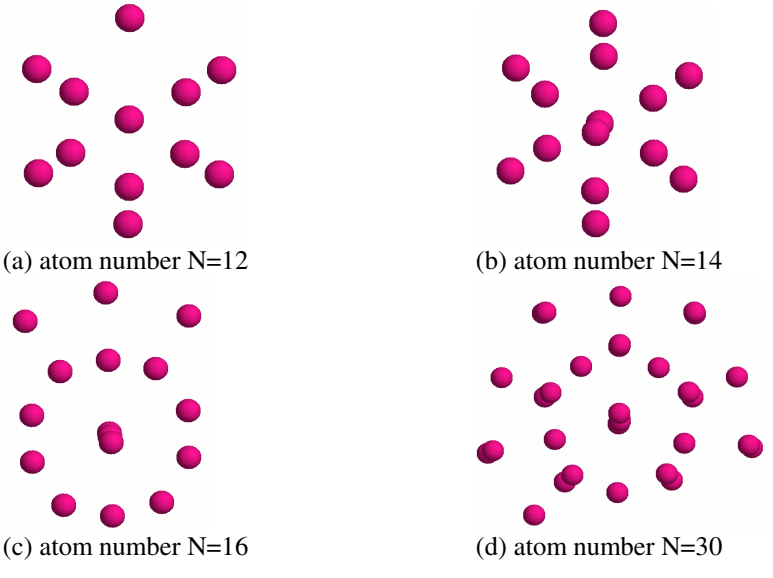


Fig. 1. The structure of argon atom clusters optimized by DE algorithm. Drawn with RasMol [9].

Table 1. The optimized energy by DE and its comparison with literature value

Atom number N	Energy in Literature[4]	Energy in DE algorithm
12	-37.97	-37.968
13	-44.33	-44.330
14	-47.85	-47.845
15	-52.32	-52.320
16	-56.82	-56.816
30	-128.29	-127.286

5 Summaries and Future Work

In this paper we use differential evolution algorithm to find global minimum energy structure of small argon clusters. Experimental results show that the exact global optimal configuration of argon clusters with atom number $N \leq 16$ can be found in a reasonable computing time, and approximate optimization can also be obtained for clusters with N=30. From their 3-D geometry structures, we can see that their optimal energy structures are highly symmetrical.

For future work, we will use the proposed DE method in other optimization areas, such as Si atomic clusters and protein 3-D structure prediction.

Acknowledgments. This work was in part supported by NSFC (Grant No. 60572015) and 973 Pre-research Project (Grant No. 2004CCA02500).

References

1. Berry, R.: Potential surfaces and dynamics: what clusters tell us. Chem. Rev., Vol. 93, (1993) 2379–2394
2. Wille, L., Vennik, J.: Computational complexity of the ground-state determination of atomic clusters. J. Phys. A, Vol. 18, (1985) L419–L422
3. Greenwood, G.: Revisiting the complexity of finding globally minimum energy configurations in atomic clusters. Z Phys. Chem., Vol. 211, (1999) 105–114
4. JIANG, H.Y.: The Geometry Optimization of Argon Atom Clusters Using a Parallel Ge-netic Algorithm, Computers and Applied Chemistry, Vol. 19, (2002) 9–12 (In Chinese)
5. XIA, B.Y.: The Optimization of the Argon Atom Cluster Structure Using a Modified Genetic Algorithm, Computers and Applied Chemistry, Vol. 18, (2001) 139–142 (In Chinese)
6. Stillinger, F., Weber, T.: Computer simulation of local order in condensed phases of argon. Phys. Rev.B, Vol. 31, (1985) 5262–5268
7. Storn, R., Price, K.: Differential evolution: a simple and efficient adaptive scheme for global optimization over continuous spaces. Technical Report TR-95-012, International Computer Science Institute, Vol. 8, (1995) 22–23
8. Abbass, H.A., Sarker, R., Newton, C.: PDE: A pareto-frontier differential evolution approach for multi-objective optimization problems. CEC2001 Proceedings, (2001) 971–978
9. Sayle, R., Milner-White, E.J.: RasMol: Biomolecular Graphics for All, Trends Biochem. Sci. Vol. 20, (1995) 374–376

A Genetic Algorithm for Solving a Special Class of Nonlinear Bilevel Programming Problems*

Hecheng Li[1,2] and Yuping Wang[1]

[1] School of Computer Science and Technology,
Xidian University, Xi'an, 710071, China
[2] School of Science, Xidian University, Xi'an, 710071, China
lihecheng@qhnu.edu.cn, ywang@xidian.edu.cn

Abstract. A special nonlinear bilevel programming problem (BLPP), whose follower-level problem is a convex programming with a linear objective function in y, is transformed into an equivalent single-level programming by using Karush-Kuhn-Tucker (K-K-T) conditions. To solve the equivalent problem effectively, a new genetic algorithm is proposed. First, a linear programming (LP) is constructed to decrease the dimensions of the transformed problem. Then based on a constraint-handling scheme, a second-phase evolving process is designed for some offspring of crossover and mutation, in which the linear property of follower's function is used to generate high quality potential offspring.

Keywords: Bilevel programming problems, genetic algorithm, linear programming problem, constraint handling, optimal solutions.

1 Introduction

The bilevel programming problem (BLPP) is a mathematical model of the leader-follower game. As an optimization problem with a hierarchical structure, BLPP has a wide variety of applications[1]. However, owing to the complex structure, the vast majority of research on BLPP is concentrated on the linear version of the problem, and a few works on the nonlinear BLPP[2]. Moreover, most of existing algorithms for nonlinear BLPP are usually based on the assumption that all of functions are convex and twice differentiable[3]. In recent years, genetic algorithms (GAs) have been used for solving BLPP[2,4,6]. [2] proposed an evolutionary algorithm for solving the BLPP in which the follower's problems are convex. In this paper, we further discuss the simplified model of the BLPP given in [2], in which the follower's objective function is linear in y. We construct an LP to avoid increasing the dimensions of the search space, and design a second-phase evolving process after both crossover and mutation.

Since a convex programming can be transformed into another convex programming with a linear objective function, as a result, the proposed algorithm can also be used for solving the BLPP given in [2].

* This work is supported by the National Natural Science Foundation of China (No. 60374063).

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 1159–1162, 2007.
© Springer-Verlag Berlin Heidelberg 2007

2 Transformation of the Problem

We consider the following nonlinear bilevel programming problem (BLPP):

$$\begin{cases} \min\limits_{x\in X} F(x,y) \\ s.t.\ G(x,y)\le 0 \\ \min\limits_{y\in Y} f(x,y)=c(x)^T y \\ s.t.\ g(x,y)\le 0 \end{cases} \tag{1}$$

where $F: R^n \times R^m \to R$, $G: R^n \times R^m \to R^p$, $g: R^n \times R^m \to R^q$, and $c: R^n \to R^m$. For x fixed, each component of g is convex and differentiable in y. Let the search space $\Omega = \{(x,y)|x \in X, y \in Y\}$, and the constraint region $S = \{(x,y) \in \Omega | G(x,y) \le 0, g(x,y) \le 0\}$. For other related definitions, refer to [2,3].

We assume that $intS \neq \phi$. Replace the follower's programming problem by K-K-T conditions, we can transform the nonlinear BLPP (1) as follows:

$$\begin{cases} \min\limits_{x,y,\lambda} F(x,y) \\ s.t.\ G(x,y)\le 0, g(x,y)\le 0, \lambda \ge 0 \\ c(x)+(\bigtriangledown_y g(x,y))^T\lambda = 0, \lambda^T g(x,y)=0 \end{cases} \tag{2}$$

where $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_q)^T$ is Lagrangian multipliers.

3 Constraint-Handling and Decreasing the Dimensions of the Transformed Problem

For any infeasible individual B, a new constraint-handling scheme is designed to generate an approximate feasible individual D. Firstly, randomly choose an individual $A \in S$. Let $\hat{S} = \{(x,y) \in \Omega | G(x,y) \le \epsilon_1, g(x,y) \le \epsilon_2\}$, and $\epsilon = (\epsilon_1^T, \epsilon_2^T)^T$, where ϵ_i are small positive vectors and tend to zero with the increasing of the generations. Let $D = rB + (1-r)A$, where $r \in (0,1)$ is random number, if $D \in \hat{S}$, then stop. Otherwise, let $B = D$, and re-compute D. The process is repeated until $D \in \hat{S}$.

For fixed $\bar{x}$ and $\bar{y}$, one can get λ by solving the following LP:

$$\begin{cases} \bar{u}(\bar{x},\bar{y}) = \min\limits_{\lambda,U}(1,1,\ldots,1)U \\ s.t. h(\bar{x},\bar{y},\lambda)+U=0 \\ \lambda \ge 0, U \ge 0 \end{cases} \tag{3}$$

where $h(\bar{x},\bar{y},\lambda) = ((c(\bar{x})+(\bigtriangledown_y g(\bar{x},\bar{y}))^T\lambda)^T, \lambda^T g(\bar{x},\bar{y}))^T$, U is an artificial vector. Thus we only need to evolve $(\bar{x},\bar{y})$ in the algorithm, this is equivalent to the reduction of the problem dimensions.

4 Proposed Algorithm(Algorithm 1)

Step 1 (Initialization). Randomly generate the initial population $pop(0)$ of N_p points in Ω such that there is at least one point in S. Apply the constraint-handling scheme to the points which don't belong to $\hat{S}$ such that these N_p points are in $\hat{S}$. Denote $N = \{(x, y) \in pop(0) \cap S\}$ and let $k = 0$.

Step 2. Evaluate the fitness $\bar{F}(x, y) = \begin{cases} F(x, y), & \bar{u}(x, y) = 0; \\ K + \mu\bar{u}(x, y), & \bar{u}(x, y) \neq 0. \end{cases}$ where K is an upper-bound of $F(x, y)$ on the set $\{(x, y)|\bar{u}(x, y) = 0\}$, $\mu \geq 0$.

Step 3 (*Crossover*). For each pair of randomly matched parents p_1 and p_2, the crossover generates offspring: $o1 = rp_1 + (1 - r)p_2, o2 = (1 - r)p_1 + rp_2$, where $r \in [0, 1]$ is random number. Let $O1$ stands for the set of all these offspring.

Sept 4 (*Mutation*). Gaussian mutation is executed, and the offspring set is denoted by $O2$.

Step 5 (*Constraint-handling*). Let $O = O1 \cup O2$. If any point in O is not in S, then arbitrarily choose $\eta \in N$ to replace a point in O. Apply the proposed constraint-handling method to modify each point $\tau \in O$ which is not in $\hat{S}$, such that all points in O are in $\hat{S}$. Let $N = \{(x, y) \in S \cap O\}$ and $\epsilon = \theta\epsilon, \theta \in [0, 1]$.

Step 6 (*Improving offspring by the second-phase evolving*). For each point $(x, y) \in N \subset O$, let d be a descent direction of $f(x, y)$ in y for x fixed. Take $\rho > 0$ such that $\bar{y} = y + \rho d$ reaches the boundary of the feasible region of the follower's problem for x fixed. Replace (x, y) by $(x, \bar{y})$ in O.

Step 7 (*Selection*). Evaluate the fitness values of all points in O. Select the best n_1 points from the set $pop(k) \cup O$ and randomly select $N_p - n_1$ points from the remaining points of the set. All these selected points form the next population $pop(k + 1)$.

Step 8. If the termination condition is satisfied, then stop; Otherwise, let $k = k + 1$, go to *Step* 3.

5 Simulation Results

In this section, 10 benchmark problems $F1 - F10$ are selected from the references [3,4,5,6,7,8] for simulation. In order to demonstrate the effectiveness of the proposed algorithm on the BLPPs with nondifferentiable leader level functions, we construct two benchmark problems $F11$ and $F12$ only by replacing the leader's objective functions in $F8$ and $F10$ by $F(x, y) = |sin(2x_1 + 2x_2 - 3y_1 - 3y_2 - 60)|$ and $\min_x F(x, y) = |sin((x_1 - 30)^2 + (x_2 - 20)^2 - 20y_1 + 20y_2 - 225)|$, respectively.

The parameters are chosen as follows: $N_p = 30$, the crossover probability $p_c = 0.8$, the mutation probability $p_m = 0.3$, $n_1 = 10$, $\mu = 1$, the initial $\epsilon = (1, \cdots, 1) \in R^{p+q}$, $\theta = 0.7$ for $k \leq k_{max}/2$, while $\theta = 0$ for $k > k_{max}/2$, where k represents generation number, while k_{max} the maximum generation number. For $F1 - F3$, $F5$, $F8$, $F9$ and $F11$, $k_{max} = 50$, while for other problems, $k_{max} = 100$. We execute *Algorithm* 1 in 30 independent runs on each problem, and record the following data: (1) leader's(follower's) objective values $F(x^*, y^*)(f(x^*, y^*))$ at the best solution; (2) the leader's objective function value $F(\bar{x}, \bar{y})$ at the worst point $(\bar{x}, \bar{y})$; (3) mean value of $F(x, y)$ in all 30 runs(denoted by F_{mean} in short).

Table 1. Comparison of the results found by Algorithm 1 and the related algorithms

	$F(x^*, y^*)$		$f(x^*, y^*)$			
No.	Algorithm 1	Ref.	Algorithm 1	Ref.	F_{mean}	$F(\bar{x}, \bar{y})$
$F1$[5]	-9	NA	-54	NA	-9	-9
$F2$[3]	2	2	12	12	2	2
$F3$[4]	1000	1000	1	1	1000	1000
$F4$[4]	-1.2098	3.57	7.617	2.4	-1.2096	-1.2090
$F5$[4]	100.003	100.58	0	0.001	100.012	100.039
$F6$[4]	81.3262	82.44	-0.3198	0.271	81.3263	81.3266
$F7$[6]	0	0	5	5	0	0
$F8$[7]	0	5	200	0	0	0
$F9$[6]	469.1429	469.1429	8.8571	8.8571	469.1429	469.1429
$F10$[8]	225	225	100	100	225	225
$F11$	0	NA	200	NA	0	0
$F12$	0	NA	100	NA	0	0

All results are presented in Table 1, where NA means that the result is not available for the algorithms and Ref. stands for the related algorithms in references. It can be seen from Table 1 that for $F4, F5, F6$ and $F8$, the best results found by *Algorithm* 1 are better than those by the compared algorithms. For $F11$ and $F12$, *Algorithm*1 found the optimal solutions. For other problems, the best results found by *Algorithm*1 are almost as good as those by the compared algorithms.

References

1. Colson, B., Marcotte, P., Savard, G: Bilevel programming: A survey. A Quarterly Journal of Operations Research(4OR) 3(2005) 87–107
2. Wang, Yuping, Jiao, Yong-Chang, Li, Hong : An Evolutionary Algorithm for Solving Nonlinear Bilevel Programming Based on a New Constraint-Handling Scheme. IEEE Transactions on Systems, Man, and Cybernetics(C) 35(2)(2005) 221–232
3. Bard, J. F.: Practical Bilevel Optimization. Norwell, Kluwer, MA (1998)
4. Oduguwa, V., Roy, R.: Bi-level optimization using genetic algorithm. Proceeds of the 2002 IEEE International Conference on Artificial Intelligent Systems (ICAIS02) (2002) 123–128
5. Zheng, P.-E.: Hierarchical optimization algorithm-based solutions to a class of bilevel programming problems. Systems Engineering and Electronics 27(4) (2005) 663–665
6. Li, H., Wang, Yuping: A hybrid genetic algorithm for nonlinear bilevel programmings. J. Xidian University 29(6) (2002) 840–843
7. Aiyoshi, E., Shimuzu, K.: A solution method for the static con-strained Stackelberg problem via penalty method. IEEE Trans. Autom. Control AC-29(12) (1984) 1112-1114
8. Shimizu, K., Aiyoshi, E.: A new computational method for Syackelberg and min-max problems by use of a penalty method. IEEE Trans. Autom. Control AC-26(2) (1981) 460–466

Evolutionary Strategy for Political Districting Problem Using Genetic Algorithm

Chung-I Chou[1], You-ling Chu[1], and Sai-Ping Li[2]

[1] Department of Physics, Chinese Culture University, Taipei, Taiwan 111, R.O.C.
[2] Institute of Physics, Academia Sinica, Taipei, Taiwan 115, R.O.C.

Abstract. The aim of the Political Districting Problem is to partition a zone into electoral districts with constraints such as contiguity, population equality, etc. By using statistical physics methods, the problem can be mapped onto a q-state Potts model system, and the political constraints are written as an energy function with interactions between sites or external fields acting on the system. This problem is then transformed into an optimization problem. In this paper, we apply the genetic algorithm to Political Districting Problem. We will illustrate the evolutionary strategy for GA and compare with results from other optimization algorithms.

Keywords: Genetic Algorithm, q-state Potts model, Political Districting Problem.

1 Introduction

The aim of the Political Districting Problem is to partition a territory into electoral districts subject to constraints such as contiguity, population equality, etc. In our previous work [1], we have mapped the political districting problem onto a q-state Potts model In this model, we use "precinct" as the smallest unit and identify it as a site. The constraints can then be written as interactions between sites or external fields acting on the system. Districting into q voter districts is thus equivalent to finding the ground state of this q-state Potts model which then becomes an optimization problem. In this earlier work, we used simulated annealing method to study both computer generated and real world districting cases.

Since genetic algorithm (GA) is known to be very useful in studying optimization problems, we will here apply this method to the Political Districting Problem. There are indeed difficulties in applying GA to this problem. The usual GA maps variables to some 1D gene code, and to generate these gene codes (or chromosomes) in a gene pool as the parent generation. Evolutionary operators such as crossover, mutation, selection, will help to generate the child generation. As the evolution goes on, chromosomes with better fit will appear. For a spin system like we have here, problems will appear in the crossover stage. The usual crossover is to swap some part of the sequences among 1D parent gene codes. Since the Political Districting Problem is a 2D problem, only randomly

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 1163–1166, 2007.
© Springer-Verlag Berlin Heidelberg 2007

swapping the 1D parent gene codes will lose the important 2D structural information. Most of the newly generated chromosomes will then behave worse than their parents. The crossover strategy for the Political Districting Problem thus needs to be modified. We here design a process to mix the boundaries of the districting zone of a pair of parents, and to generate a child from these mixed boundaries. In this way, each child will in some way look similar to its parents. In the following, we will illustrate our evolutionary strategy for GA and compare the results with other optimization algorithms.

2 Model for Political Districting Problem

In our model [1], the total number of sites (precincts) is equal to N and each spin can have q-states (voter districts). Each spin can have its state function as $S_i = 1 \ldots q$. The goal is to find the ground state of this q-state Potts model with the interaction given by the constraints. We consider population equality, contiguity and compactness here, which are the most common constraints people consider.

For population equality, we associate a random field to the site (precinct) p_i. It is easy to see that when p_i is a constant for every i, the population for each site (precinct) is equal. Therefore, the magnetization (the total voter population) of a voter district P_l can be written as $P_l = \sum_{i=1}^{N} p_i \cdot \delta_{S_i,l}$, where $\delta_{i,j}$ equals to 1 when i and j are equal and zero otherwise, the total population is $P_0 = \sum_{l=1}^{q} P_l$, and the average population for each voter district is $< P >= P_0/q$. Hence we can write the average population for each voter district as $E_P = \sum_{l=1}^{q} \left|1 - \frac{P_l}{<P>}\right|$. The smaller this energy is, the closer it is to the average value.

We next consider the constraint of contiguity and compactness. Define a connection table for each spin as $C_{i,j}$ which equals to 1 when spins i and j are connected to each other and zero otherwise. We here define the boundary of the domain of spins as $E_D = \sum_{i,j} \left(1 - \delta_{S_i,S_j}\right) \cdot C_{i,j}$, When this function is in its minimum, the district will have the smallest number of precincts on its boundary. The total energy is therefore given by $E = \lambda_P E_P + \lambda_D E_D$. Varying λ will affect the contribution of each constraint to the total energy function.

With this energy function, the problem now becomes an optimization problem and can be solved by using optimization algorithms.

3 Strategy for Genetic Algorithm

As mentioned above, genetic algorithms try to simulate the evolutional process of biological system. For the spin system here, the crossover procedure of the usual GA will lose important parent's 2D structure information. We will therefore introduce a modified crossover strategy for the Political Districting Problem.

Recall that the Key idea of the crossover strategy is to find the main information of the parent generation, and to help the child inherit this information. Since the aim of the political districting problem is the find some suitable districting zones, we believe the main information should be the boundary of each

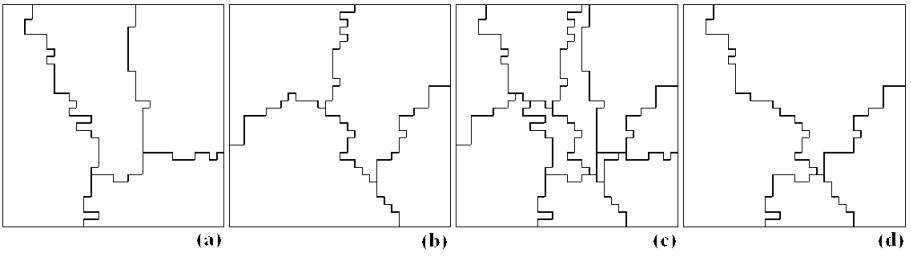


Fig. 1. A sketch of cross-over procedure. (a) The boundaries of districts from a parental chromosome; (b) Boundaries from another parental chromosome; (c) Mixed boundaries; (d) The filial chromosome.

districting zone. We thus here design a crossover process to mix the boundary of districting zone of each of the two parents, and generate a child districting zone from these mixed boundaries. Each child in some way inherit its parents' characteristic and would look similar to its parents. We illustrate the detail of our GA process in the following.

First, we should prepare a gene pool with many genes (or filial generations). To create a chromosome, we randomly put q seeds with different states into an N spin system and let these seeds grow until they fill all spins of system. This process will create a pool of chromosomes but only chromosomes with lower energy can be survived. These chromosomes compose of a gene pool of first generation. Once gene pool is constructed, the GA process can go on. A lot of children will be created by Cross-over and mutation procedure, and newer gene pool of children's generation will be constructed by using selection procedure.

The aim of our Crossover Strategy is to help the child generation inherit their parent's information of boundary of each districting zone. To achieve this aim, we randomly pick two chromosomes, and mix their boundaries into a new spin system and form a new chromosome. Since this chromosome with too many boundaries of districting zone, we randomly remove some boundary lines to make sure there are only q different states in this chromosome. By using this procedure, the child chromosome will inherit its parents' most important information. Fig. 1 is a sketch of the cross-over procedure.

The mutation procedure of GA is to increase the diversity of the gene pool, and to avoid being trapped in a local minimum. We use a simple method to create a mutated chromosome. We pick a chromosome from gene pool , and randomly cut a small domain in the spin system and let spins to start growing diffusively from the edge of domain, till the domain is filled . By controlling the size of the cut domain, we can control the verity between mutated chromosome and original one.

From both the crossover and mutation procedures, many children chromosomes will be created. These chromosomes will go through the selection procedure to determine which chromosomes will survive. Here we use a simple rule of selection–only chromosomes with lower energy will survive.

Table 1. A comparison of algorithms.(SR: Success Rate(%). Total number of trials is equal to 100).

No. of spins	100		400		900	
	E_{min}	SR	E_{min}	SR	E_{min}	SR
Method(i)	20	100	40	61	60	72
Method(ii)	20	100	40	49	60	70
Method(iii)	20	86	40	13	60	22
Method(iv)	20	85	40	2	60	4
Method(v)	20	—	40	—	60	—

4 Result and Conclusion

We first choose $N \times N$ square-lattice spin systems as a test of our idea. We use five methods to find the minimum of the political districting problem, namely, (i) GA with modified Crossover strategy and modified Mutation strategy, (ii) GA with modified Crossover strategy and usual single Spin-swapping Mutation strategy, (iii) GA with usual 1D Spin-swapping Crossover strategy and modified Mutation strategy, (iv) GA with usual 1D Spin-swapping Crossover strategy and single Spin-swapping Mutation strategy, and (v) simulated annealing method(SA). Table 1 is a summary of results using these five methods. In these test cases, the size of each gene-pool is equal to 400, the numbers of Crossover and Mutation in each generation are 4000 and 1000, and the total number of generation of each method is 300. The districting parameters λ_P and λ_D equal to 40 and 1. The results show that the GA with both modified Cross-over and Mutation strategy is more effective than any other GA methods.

In this paper, we mapped the Political Districting Problem onto a q-state Potts model in which the constraints can be written as interactions between sites or external fields acting on the system. We then show how to modify the genetic algorithm and apply it to the Political Districting Problem. Since the usual GA hardly communicate the structural information from the parent generation to their children in spin system, we design a new crossover strategy to help the child generation chromosomes inherit their parents' structural information. Our results show that this strategy work well in the Political Districting Problem.

Acknowledgments

This work was supported in part by the National Science Council, Taiwan, R.O.C. (grant no. NSC-94-2112-M-001-019 and NSC-94-2112-M-034-001).

Reference

1. Chou, C.I., Li, S.P.: Taming the Gerrymander–Statistical Physics Approach to Political Districting Problem. Physica A **369** (2006) 799–808

An ACO Algorithm with Adaptive Volatility Rate of Pheromone Trail

Zhifeng Hao[1,2], Han Huang[1], Yong Qin[3], Ruichu Cai[1]

[1] College of Computer Science and Engineering, South China University of Technology, Guangzhou 510640, P.R. China
[2] National Mobile Communications Research Laboratory, Southeast University, Nanjing 210096, P.R. China
[3] Center of Information and Network, Maoming University, Maoming, Guangdong, 525000, P.R. China
bssthh@163.com, hanhuang_scut@hotmail.com

Abstract. Ant colony optimization (ACO) has been proved to be one of the best performing algorithms for NP-hard problems as TSP. The volatility rate of pheromone trail is one of the main parameters in ACO algorithms. It is usually set experimentally in the literatures for the application of ACO. The present paper proposes an adaptive strategy for the volatility rate of pheromone trail according to the quality of the solutions found by artificial ants. The strategy is combined with the setting of other parameters to form a new ACO algorithm. Finally, the experimental results of computing traveling salesman problems indicate that the proposed algorithm is more effective than other ant methods.

Keywords. Ant colony optimization, pheromone trail, adaptive volatility rate

1 Introduction

ACO was first proposed by M. Dorigo and his colleagues as a multi-agent approach to deal with difficult combinatorial optimization problems such as TSP [1-2]. Since then, many applications to the NP-hard problems have shown the effectiveness of ACO [3].

The main parameters of ACO may conclude: k is the number of artificial ants, ρ is the parameter for volatility of pheromone trail and α, β determine the relative importance of pheromone value and heuristic information [2]. All of the parameters are usually set with experimental methods in the application of ACO [3]. The have been also several works [4-7] for the adaptive parameter setting of ACO algorithms.

This paper presents a tuning rule for ρ based on the quality of the solution constructed by the artificial ant. Then, the adaptive rule is used to form a new ACO algorithm, which is tested to compute several benchmark TSP problems. The experimental result indicates that the proposed ACO algorithm with adaptive ρ performs better than other ACO algorithms [2, 8].

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 1167–1170, 2007.
© Springer-Verlag Berlin Heidelberg 2007

2 Adaptive Volatility Rate of Pheromone Trail

In ACO algorithms [3], each ant builds a tour (i.e. a feasible solution to the TSP problem) by repeatedly applying the state transition rule as Equation 1 shows.

$$P_{gs}^{(m)}(t)=\begin{cases}\dfrac{[\tau_{gs}(t)]^{\alpha}[\eta_{gs}]^{\beta(g,t)}}{\sum_{r\in J_m(g)}[\tau_{gr}(t)]^{\alpha}[\eta_{gr}]^{\beta(g,t)}} & if \quad s\in J_k(g)\\ 0 & otherwise\end{cases} \tag{1}$$

where P^{m}_{gs} is the probability with which the m-th ant moves from city g to city s in the t-th iteration, τ is the pheromone, $\eta=1/d$ is the reciprocal of distance d_{gs}, and $J_m(g)$ is the set of cities not visited when ant m is at city g.

After constructing its tour, an artificial ant also modifies the amount of pheromone on the visited edges by applying the pheromone updating rule:

$$\tau_{gs}(t+1)=(1-\rho)\tau_{gs}(t)+\rho\Delta\tau_{gs}(t) \tag{2}$$

where $\Delta\tau_{gs}(t)$ is the increment for the pheromone of edge (g,s) at the t-th iteration, and $\rho=0.1$ is the volatility rate of the pheromone trail [2-3].

In order to update the pheromone according to the quality of solutions found by ants, an adaptive rule for volatility of the pheromone trail is designed as follows:

$$\rho_m=L_m^{-1}/(L_m^{-1}+L_P^{-1}) \tag{3}$$

where L_m is the length of the solution S_m found by the m-th ant, and L_P is the length of the solution built based on the pheromone matrix (Equation 4).

$$s=\arg\max_{u\in J_m(r)}\{[\tau(r,u)\} \tag{4}$$

where s is the city selected as the next one to city r for any $(r,s)\in S_P$.

The motivation of the proposed rule is: better solutions should contribute more pheromone, and the worse ones contribute less. We use this rule to design a new ACO algorithm, which is similar to ant colony system (ACS) [2] except for the updating rule [2] (Equation 5). For $i=1,...,k+1$ (k ants and the best-so-far ant),

$$\tau_{gs}(t+1)=(1-\rho_i)\tau_{gs}(t)+\rho_i L_i^{-1} \tag{5}$$

where $\forall(g,s)\in S_i$ and $\rho_i=L_i^{-1}/(L_i^{-1}+L_P^{-1})$ for the t-th iteration.

3 Numerical Results

This section indicates the numerical results in the experiment that the proposed ACO algorithm is used to solve TSP problems [9]. Other ant approaches ACS [2] and the ACO [8] are also tested in the same machines as the comparison with the proposed ACO. Several TSP instances are computed by the three algorithms on a PC with an Intel Pentium 550MBHz Processor and 256MB SDR Memory, and the results are shown in Table 1. It should be noted that every instance is computed 20 times. The algorithms would not stop until a better solution could be found in 500 iterations.

Table 1 shows that the proposed ACO algorithm (PACO) performs better than ACS [2] and the ACO [8]. The shortest lengths and the average lengths obtained by PACO are shorter than those found by ACS and the ACO in all of the TSP instances. Furthermore, it can be concluded that the standard deviations of the tour lengths obtained by PACO are smaller than those of another algorithms. Therefore, we can conclude that PACO is proved to be more effective and steady than ACS [2] and the ACO [8]. Computation time cost of PACO is not less than ACS and ACO in all of the instances because it needs to compute the value of volatility rate $k+1$ times per iteration. Although all optimal tours of TSP problems cannot be found by the tested algorithms, all of the errors for PACO are much less than that for another two ACO approaches. The algorithms may make improvement in solving TSP when reinforcing heuristic strategies like ACS-3opt [2] are used.

Table 1. Comparison of the results in TSP instances

Problem	Algorithm	best	Ave	time(s)	Standard deviation
kroA100	ACS	21958	22088.8	65	1142.77
	ACO	21863	22082.5	94.6	1265.30
	PACO	21682	22076.2	117.2	549.85
ts225	ACS	130577	133195	430.6	7038.30
	ACO	130568	132984	439.3	7652.80
	PACO	130507	131560	419.4	1434.98
pr226	ACS	84534	86913.8	378.4	4065.25
	ACO	83659	87215.6	523.8	5206.70
	PACO	81967	83462.2	762.2	3103.41
lin105	ACS	14883	15125.4	88.8	475.37
	ACO	14795	15038.4	106.6	526.43
	PACO	14736	14888	112.2	211.34
kroB100	ACS	23014	23353.8	56.2	685.79
	ACO	22691	23468.1	102.9	702.46
	PACO	22289	22728	169.6	668.26
kroC100	ACS	21594	21942.6	54.8	509.77
	ACO	21236	21909.8	78.1	814.53
	PACO	20775	21598.4	114.8	414.62
lin318	ACS	48554	49224.4	849.2	1785.21
	ACO	48282	49196.7	902.7	2459.16
	PACO	47885	49172.8	866.8	1108.34

4 Discussions and Conclusions

This paper proposed an adaptive rule for volatility rate of pheromone trail, attempting to adjust the pheromone based on the solutions obtained by artificial ants. Thus, a new ACO algorithm is designed with this tuning rule. There is a special pheromone updating rule in the proposed algorithm whose framework is similar to Ant Colony System. There are some experimental comparisons among the proposed ACO approach and other methods [2, 8] in solving TSP problems. The results show the effectiveness of the proposed algorithm.

Further study is suggested to explore the better management for the optimal setting of the parameters of ACO algorithms, which will be very helpful in the application.

Acknowledgements. This work has been supported by the National Natural Science Foundation of China (60433020, 10471045), Program for New Century Excellent Talents in University (NCET-05-0734), Natural Science Foundation of Guangdong Province (031360, 04020079), Excellent Young Teachers Program of Ministry of Education of China, Fok Ying Tong Education Foundation (91005), Key Technology Research and Development Program of Guangdong Province (2005B10101010), State Key Lab. for Novel Software Technology, Nanjing University (200603), open research fund of National Mobile Communications Research Laboratory, Southeast University (A200605), Nature Science Foundation of Guangdong (05011896), and Nature Science Foundation of Education Department of Guangdong Province (Z03080).

References

1. Dorigo, M., Caro, G.D., Gambardella, L.M.: Ant algorithms for Discrete Optimization, Massachusetts Institute of Technology, Artificial Life 5: 137-172, 1999
2. Dorigo, M., Gambardella, L.M.: Ant Colony System: A Cooperative Learning Approach to the Travelling Salesman Problem, IEEE Transactions on Evolutionary Computation, 1 (1), 53-66, 1997
3. Dorigo, M., Stützle, T.: Ant Colony Optimization. MIT Press, Cambridge, MA (2004)
4. Watanabe, I., Matsui, S.L.: Improving the Performance of ACO Algorithms by Adaptive Control of Candidate Set, Evolutionary Computation, 2003. CEC '03. The 2003 Congress on Volume 2, 8-12 Dec. 2003 Page(s):1355 - 1362 Vol.2
5. Pilat, M.L., White, T.: Using Genetic Algorithms to Optimize ACS-TSP, M. Dorigo et al. (Eds.):ANTS 2002, LNCS 2463, 282-287, 2002
6. Gambardella, L.M., Dorigo, M.: Ant-Q: A Reinforcement Learning approach to the travelling salesman problem, Appeared in: Proceedings of ML-95, Twelfth Intern. Conf. On Machine Learning, Morgan Kaufmann, 252-260, 1995
7. Huang H., Yang X.W., Hao Z.F., Cai R.C.: A novel ACO algorithm with adaptive parameter, D.S. Huang et al. (Eds.): ICIC 2006, LNBI 4115, 12-21, 2006
8. Sun, J., Xiong, S.W., Guo, F.M.: A new pheromone updating strategy in ant colony optimization, Proceedings of 2004 International Conference on Machine Learning and Cybernetics, 1, 620-625, 2004
9. Reinelt, G.: TSPLIB. A traveling salesman problem library. ORSA Journal on Computing, 3 (4), 376-384, 1991

A Distributed Coordination Framework for Adaptive Sensor Uncertainty Handling

Zhifeng Dai[1], Yuanxiang Li[2], Bojin Zheng[3], and Xianjun Shen[4]

[1] School of Computer, Wuhan University
[2] State Key Lab of Software Engineering, Wuhan University
[3] College of Computer Science, South-Central University for Nationalities
[4] Department of Computer Science, Huazhong Normal University
430072 Wuhan, China
dwjse@163.com, yxli@whu.edu.cn, zhengbojin@gmail.com

Abstract. The evaluation and management of sensor uncertainty is particularly necessary in a noisy multi-sensor context. In this paper, focusing on the potential of distributed coordination among sensor nodes based on the built-in association between wireless sensor networks and multi-agent systems, meanwhile in a rough set technique sense of uncertainty, we show how an adaptive distributed coordination framework for a hierarchy of sensor data uncertainty performs local data fusion to increase the certainty of real-time sensor readings coherently, makes global rational decisions under imprecision and partial truth, and reconciles the conflicts somewhat, thus evolving adaptive and robust sensor uncertainty handling systems. Implementation results for an example sensor field demonstrate the application of our proposed approach.

Keywords: adaptive distributed coordination strategy, rough set theory, multi-agent systems, modeling sensor uncertainty, wireless sensor networks.

1 Introduction

Generally, an explicit definition of uncertainty may refer to the idea of an information gap between what we do know and what we need to know [1]. However, as a suitable mathematical formalization of uncertainty, rough set theory can quantify both what is known and what is unknown. Moreover, from rough set perspective, further research is needed to view the ways in which uncertainties are hierarchical and interconnected. Especially, so far there exists little work on incorporating coordination mechanism into adaptive sensor uncertainty handling, for which the issue of a distributed coordination framework is actually an important and novel research direction.

2 Background Concepts and Motivation

Rough set theory approaches the problem of inexact concepts in the form of a decision system (U, $A \cup \{d\}$). Here U is a non-empty finite set of objects, A is a set of condition attributes, and d is decision element. For any subset R of A, AS = (U, R),

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 1171–1174, 2007.
© Springer-Verlag Berlin Heidelberg 2007

an ordered pair, is the approximation space [2]; $[x]_R$ denotes the equivalence class of R containing x, for $x \in U$. The lower approximation of X is the set $\underline{RX} = \{x | x \in U, [x]_R \subseteq X\}$ and the upper approximation of X is the set $\overline{RX} = \{x | x \in U, [x]_R \cap X \neq \emptyset\}$. The results in $\underline{RX}$ are certain. The boundary region $\overline{RX}$ - $\underline{RX}$ contains those uncertainty results that are possible, but not certain.

Multi-agent systems are fundamentally designed for dealing with and offering architecture for complex distributed applications that require collaborative problem solving, and wireless sensor networks consist of a collection of light-weight sensors connected via wireless links to each other or to a more powerful gateway node [3].

While principles taken from the rough set theory are suitably employed for the mathematical modeling and further handling of sensor uncertainty through the use of its approximation space, the need for cooperative processing of spatially and temporally dense collected data because of limited resources, makes multi-agent system architecture an ideal candidate for analyzing and modeling distributed sensor data, and these ideas are the motivation behind this approach.

3 Modeling of Sensor Uncertainty

Herein, we identify uncertainty as a measure of the level of confidence of a sensed value, or of validity of a conclusion at a higher level. Due to hardware limitations and signal processing inaccuracies, there are always some forms of uncertainty associated with sensor measurement and the processed information. First, some uncertainties can only be inferred as the absence or incompleteness of certain observational data item values because of inaccessibility. And second, uncertainty can be seen as a scaling of the noise level of inconsistent data sets, also in the form of the binary data inconsistency failures during the actual sensor processing. Additionally, a sensor node usually faces uncertainties owing to its limited or outdated partial information about the states of complex problem solving in other nodes. Furthermore, there exists the uncertainty which arises from the issue of granularity effect. All these form a hierarchy of sensor data uncertainty.

4 Adaptive Distributed Sensor Uncertainty Coordination Frame

In order to effectively handle sensor uncertainty, as shown in Figure 1, we may use the term “agent” to mean any sensor node, and there are three kinds of hierarchically dispersed agents embedded on sensors, i.e. sensor agents, cluster agents, and the manager agent, which form a wireless sensor network into a distributed multi-agent sensor system. Furthermore, as coordination is the process of effectively managing interdependencies between activities distributed across agents [4], we propose a hierarchical distributed coordination framework for sensor uncertainty handling, which bridges the potential of adaptive coordination to reduce sensor uncertainty with that of intelligent decision in the form of rule with an appropriate level of certainty.

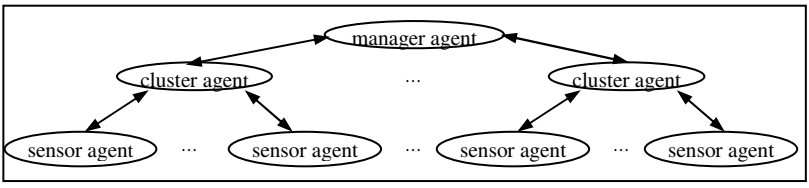


Fig. 1. Distributed multi-agent wireless sensor networks

Firstly, on the layer of local data fusion, the principle issue is to manage and minimize the measured sensor agent uncertainties through suitable fusion algorithms. On the one hand, redundant data fusion can increase the accuracy and certainty of sensed data in the case of sensor failure. On the other hand, complementary data fusion allows features to be sensed coherently through fusing the different types of data from more than one source sensors to synthesize meaningful information.

Secondly, on the layer of attribute reduction and decision rule, the major topics consist in the fact that one must not only be aware of sensor uncertainty, but also discover hidden concise relationships that agree with uncertain sensor situations. At the cluster agents, by employing the notion of reduction in a rough set sense, which is the minimal set of condition attributes that eliminates pairwise dependencies and makes the same decisions as the whole set, those redundant condition attributes along with their associated uncertainties may be removed, then the reduction of the number of attributes and further the discovery of deterministic decision rules are achieved, thus providing an appropriate model for the distributed multi-agent sensor system.

Thirdly, on the layer of global conflict resolution at the manager agent, problem of conflict resolution is of most importance due to exclusive interpretations of the same thing or conflicting sets of rules from different source sensors. In rough set approach, inconsistency is represented by the concepts of lower and upper approximation, and confliction relation is expressed by the notion of boundary region, which may be of minimal level of uncertainty by forcing a finer granulation of the partitioning. Moreover, through the use of explicit rule priorities, we can reconcile the consistency degrees about the inferences under uncertainty as far as possible.

5 Case Study: Example Implementation Results

In this section we illustrate the above presented ideas by a distributed multi-sensor paradigm for real-time forest fire detection. Herein, the sensor data can be represented as a decision table, which is constitutive of condition attributes *{temperature, relative humidity, rainfall, smoke, wind speed}* and decision attribute *{fire level}*.

Here we discuss some primary kinds of hierarchical sensor uncertainty at the cluster agents and the manager agent. A schedule of these uncertainties and the corresponding example implementation results are listed in Table 1 and Table 2, respectively. While raw uncertainty sensor data from distributed sensor agents and cluster agents have been converted into the minimal efficient sets of certain condition attribute data and deterministic decision rules at the manager agent, these demonstrate the efficiency of the adaptive distributed sensor uncertainty coordinate framework.

Table 1. A schedule of hierarchical sensor uncertainties

Sensor uncertainty class	Site	Handling method
failure uncertainty	cluster agents	redundant data fusion
incompleteness uncertainty	cluster agents	complementary data fusion
inconsistency data item uncertainty	cluster agents	prior data fusion
condition attribute uncertainty	cluster agents, manager agent	the notion of reduction
hidden relationship uncertainty	cluster agents, manager agent	the notion of reduction
conflicting rule uncertainty	manager agent	conflict resolution

Table 2. Summary of kinds of uncertainty handling in distributed multi-agent sensor system

Agent class	Node num	Failure uncertainty		Incompleteness uncertainty		Inconsistency data item uncertainty		Condition attribute uncertainty		Hidden relationship uncertainty	Conflicting rule uncertainty	
		total	disposal	total	disposal	total	disposal	total	reduction		total	disposal
cluster1	40	4	4	12	12	4	4	5	2	3	0	0
cluster2	50	6	6	20	20	8	6	5	2	3	0	0
cluster3	36	4	4	10	10	2	2	5	2	3	0	0
cluster4	52	8	8	24	24	6	4	5	2	3	0	0
manager	178	0	0	0	0	0	0	5	2	3	2	2

6 Conclusions

Considering there is some room for a mixture of approaches for the sensor uncertainty problem based on rough set theory and multi-agent system, we have proposed and demonstrated an adaptive distributed coordination frame reflecting the handling of multiple levels of sensor uncertainty, along with emphasis on how to make a complex multi-agent sensor system self-repairing and behave deterministically to some extent.

Acknowledgments. The authors would like to thank the National Science Foundation of P.R.C for financial support under Grant No. 60473014 for this research.

References

1. Hatfield A. J., Hipel K W.: Understanding and Managing Uncertainty and Information. IEEE International Conference on Systems, Man, and Cybernetics, Vol. 5. IEEE Computer Society (1999) 1007-1012
2. Theresa Beaubouef, Frederick Petry: Vague Regions and Spatial Relationships: A Rough Set Approach. Proceedings of the Fourth International Conference on Computational Intelligence and Multimedia Applications. IEEE Computer Society (2001) 313-317
3. Chris Giannella, Ruchita Bhargava, Hillol Kargupta: Multi-agent Systems and Distributed Data Mining. In: Klusch M. et al. (eds.): CIA 2004, LNAI 3191. Springer- Verlag (2004) 1-15
4. Nagendra Prasad M. V., Lesser Victor R.: Learning Situation-Specific Coordination in Cooperative Multi-agent Systems. Autonomous Agents and Multi-agent Systems, 2. Kluwer Academic (1999) 173-207

A Heuristic Particle Swarm Optimization for Cutting Stock Problem Based on Cutting Pattern

Xianjun Shen[1,2], Yuanxiang Li[2], Jincai Yang[1], and Li Yu[2]

[1] Department of Computer Science
Central China Normal University, Wuhan 430079, China
[2] State Key Lab of Software Engineering
Wuhan University, Wuhan 430072, China
xjshen@mail.ccnu.edu.cn

Abstract. A heuristic particle swarm optimization (HPSO) is proposed as a solution to one-dimensional cutting stock problem (1D-CSP), which incorporate genetic operators into particle swarm optimization (PSO). In this paper, a heuristic strategy that is based on the results of analysis of the optimal cutting pattern of particles with successful search processes is described, which process a global optimization problem of the cutting-stock as a sequential optimization problem by multiple stages. During every sequential stage, the best cutting pattern for the current situation is researched and processed. This strategy is repeated until all the required stocks have been generated. The simulation results prove the effectiveness of the proposed methodology.

Keywords: Heuristic particle swarm optimization, One-dimensional cutting stock problem, Genetic operators.

1 Introduction

The one-dimensional cutting stock problem (1D-CSP) is one of the representative combinatorial optimization problems, which arises in many industries [1]. This paper proposes a heuristic particle swarm optimization (HPSO) incorporating genetic algorithm (GA). The main idea of HPSO is to process a global optimization problem of the cutting-stock as a sequential optimization problem by multiple stages. During every sequential stage, the best cutting pattern for the current situation is researched and processed. This stage processing is repeated until all the required stocks have been generated. The experiment result shows that the HPSO can obtain satisfying effect.

2 One-Dimensional Cutting Stock Problem

In the pattern-oriented approach, at first, order lengths are combined into cutting patterns, however, it is impractical to consider all feasible cutting patterns, and

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 1175–1178, 2007.
© Springer-Verlag Berlin Heidelberg 2007

we would try to find a set of n better cutting patterns yielding small deviations from the order demands. In a succeeding step, the frequencies are determined that are necessary to satisfy the order demands.

One of the most important costs for 1D-CSP is the amount of residual pieces of processed stock, called trim loss, which are usually treated as waste product. Hence, the problem of minimizing the total trim loss (or the number of processed stock) is considered as the most important factor.

To define 1D-CSP, it is given a sufficient number of stocks, which have the same length L, and m types of products with given lengths $(l_1, l_2, ..., l_m)$ and their demands $(d_1, d_2, \ldots, d_m)$. A cutting pattern is a combination of products cut from a stock. A cutting pattern is described as $p_j = (a_{1j}, a_{2j}, ..., a_{mj})$, where a_{ij} is the number of product i cut from one stock roll. It is suppose that have found n type optimal cutting patterns, the first cutting pattern is cut the quantity of the m-th type product is a_{m1}, and its the trim loss is b_1. The second cutting pattern is cut the quantity of the m-th type product is a_{m2}, and its the trim loss of the second is b_2. Thus, the cutting pattern n is cut the quantity of the m-th type product is a_{mn}, and its the trim loss is b_n [2]. The mathematics model of one-dimensional cutting stock problem is formulated as follows:

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ a_{m1} & a_{m2} & \ldots & a_{mn} \end{bmatrix} \tag{1}$$

The number of stock that use n type different cutting pattern is as follow:

$$X = (x_1, x_2, \ldots, x_n)^T \tag{2}$$

The trim loss of cutting pattern p_j is as follows.

$$b_j = L - (a_{1j}l_1 + a_{2j}l_2 + ... + a_{mj}l_m) \quad j = 1, 2, ..., n \tag{3}$$

Where a_{mj} is the number of m-th type product that use cutting pattern p_j. Now the object function of the pattern-oriented approach is formulated as follows.

$$f(x) = \min(\sum_{j=1}^{n} b_j x_j) \tag{4}$$

The last stock has the maximum remainder and can in general be used further, thus it is not considered as waste. The summation of trim loss in one-dimensional cutting stock problem should be subtracted the remainder of the last stock.

$$f(x) = \min(\sum_{j=1}^{n} b_j x_j - L')(1 \leq i \leq k) \tag{5}$$

Where L' denotes indicate the longest remainder of a cutting plan which cumulate consecutive residual lengths in one stock which could be used later.

3 Heuristic Particle Swarm Optimization

PSO is a swarm intelligence optimization algorithm. The information exchange takes place only among the particle's own experience and the experience of the best particle in the swarm. HPSO model is still based on PSO mechanism, but use genetic operator as the updating operator. The best position of each particle would update by crossover operator and mutation operator.

In heuristic particle swarm optimization, each potential optimal cutting plan, called a particle, the current position of the particle is denoted x_i^t. The particles have memory and each particle keep track of its previous best position (denoted $xbest_i^t$) and its corresponding fitness. The particle with greatest fitness is called the global best and its position is called the global best position ($gbest$). Each particle towards it's $xbest_i^t$ and the $gbest$ position. The current number of iteration generation is represented by the symbol t.

The main steps in HPSO Process are described as follows:

Step 1. Initialize the swarm randomly.

Step 2. For each particle of swarm, Choose x_i^t and $xbest_i^t$, and generate $x_i^{'t}$ by crossover operator.

Step 3. For each particle of swarm, Choose $x_i^{'t}$ and x_{gbest}, and generate $x_i^{''t}$ by crossover operator.

Step 4. For each particle of swarm, Choose $x_i^{''t}$, and generate x_i^{t+1} by mutation operator.

Step 5. Calculate the fitness each particle.

Step 6. Update the best position of each particle so far.

Step 7. Update the global best position of the whole swarm so far.

Step 8. Repeat Step 2 to Step 6, Choose the best cutting pattern for the current situation.

Step 9. Process cutting stocks by the best cutting pattern, subtract the quantities of the every type pieces that have been cut, and then update each particle of swarm.

Step 10. If all the required stocks have been not generated, go to step 2.

Step 11. Output the final results of HPSO.

4 Experiments Results

One of the most important costs for 1D-CSP is the amount of residual pieces of processed stock, called trim loss. The purpose of HPSO is that search a cutting plan that has the least trim loss The benchmark instance was given in [3]. The optimal cutting plan was given in [3] by a hybrid genetic algorithm which the longest remainder of stock is 2746mm. The total trim loss of other stocks is 2411mm.

According to the results in Table.1, HPSO has attained a satisfying optimal cutting plan. The optimal cutting plan is given in table 1 which the longest remainder of stock is 133mm. The total trim loss of other stocks is 163mm. The average availability ratio of other stocks is 99.86%. The result is satisfactory that overall stocks needed by all cuts are minimized.

Table 1. Solution of instance with HPSO

No.	Stock length	Pieces length (Pieces amount)	Trim loss	Availability ratio	Pattern number
1	6000	1694(2) 1687 925	0	100	2
2		885(3) 855(2) 828 807	0	100	2
3		1464 978 925 889(2) 855	0	100	2
4		1416 1394 1389 984 817	0	100	1
5		984 925 828 817(3) 811	1	99.99	1
6		2137 1494 1387 978	4	99.93	1
7		1422 1392 1343 978 861	4	99.93	1
8		1094(3) 978 925 811	4	99.93	1
9		1494 1422 1419 828(2)	9	99.85	1
10		2144 1446 1416 861	133	97.78	1
11	8000	2144 1296 1167 889 885 811 808	0	100	2
12		1541 1464 1167(2) 925(2) 811	0	100	1
13		1389 1167 1094 984 889 861 808(2)	0	100	1
14		2137 1167(2) 1107 808 807(2)	0	100	1
15		1422 1392 984 861(3) 811 808	0	100	1
16		1494(2) 1416 1392(2) 811	1	99.99	1
17		2137 1426 1416 1400 811 808	2	99.97	1
18		2137 1676 1464 1419 1296	8	99.90	1
19		2144 1676 1422 1389 1337	32	99.60	1
20		1389 1167 1107 861(3) 828(2)	98	98.78	1
21	9000	1081(4) 1034(2) 984 817 807	0	100	4
22		978(3) 925 906 889 855(2) 828 808	0	100	1

5 Conclusion

This paper presents a heuristic particle swarm optimization algorithm to solve a one-dimensional cutting stock problem, which merges crossover operator, mutation operator, and heuristic strategy based on best cutting pattern to explore the best cutting plan. The simulation result shows the proposed heuristic algorithm is applied successfully to solve 1D-CSP.

References

1. Gradišar, M., Trkman, P.: A combined approach to the solution to the general one-dimensional cutting stock problem. Computers and Operations Research **32**(7) (2005) 1793–1807
2. Jia, Z., Yin, G., Hu, X., Shu, B.: Optimization for one-dimensional cutting stock problem based on genetic. Journal of Xian'an Jiaotong University **36** (2002) 967–970
3. Peiyong, L.: Optimization for variable inventory of one-dimensional cutting stock. Mechanical Science and Technology **22(supplement)** (2002) 80–86

Theory of Evolutionary Algorithm: A View from Thermodynamics

Yuanxiang Li[1], Weiwu Wang[1], Xianjun Shen[2], Weiqin Ying[1], and Bojin Zheng[3]

[1] State Key Lab of Software Engineering
Wuhan University, Wuhan, China, 430072
{yxli@whu.edu.cn,wangww@mail.whut.edu.cn,weiqinying@yahoo.com.cn}
[2] Department of Computer Science
Central China Normal University, Wuhan, China, 430079
xjshen@mail.ccnu.edu.cn
[3] College of Computer Science
South-Central University For Nationalities, Wuhan, China, 430074
zhengbojin@gmail.com

Abstract. It is recognized that evolutionary algorithms are inspired from evolutionary biology. In this paper, we set up a thermodynamic model of evolutionary algorithm. This model is intuitive and has a solid foundation in thermodynamics. It is our first step towards a unified theory of evolutionary algorithms.

Keywords: Evolutionary Algorithm, Thermodynamics, Thermodynamic Model of Evolutionary Algorithm.

1 Introduction

The term Evolutionary Algorithm (EA) stands for a family of stochastic problem solvers based on principles that can be found in biological evolution. Within this paradigm, achieving a solution to a given problem is seen as a survival task: possible solutions compete with the others for survival, and this competition is the driving force behind the progress that supposedly leads to an optimal solution [1].

There are numerous successful applications of EA in business and industry, but such successes are not fully understood. It has been proposed a lot of theories such as schema theory [2], Markov chains theory [3], dimensional analysis [4], order statistics [5], quantitative genetics [6], quadratic dynamical systems [7] and statistical physics [8]. But the problem of these theories is either that they are too complicated to understand or they do not apply to algorithms that are widely used. Is there an intuitive and unified theory of EA? In this paper, we present our preliminary result of this research. It is the thermodynamic model of EA.

2 Basis of Thermodynamics

Thermodynamics is a branch of physics which deals with the energy and work of a system. Thermodynamics deals only with the large scale response of the

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 1179–1182, 2007.
© Springer-Verlag Berlin Heidelberg 2007

system. Small scale interactions are described by kinetic theory. These both theories compliment the other; some principles are more easily understood in terms of thermodynamics and some other principles are more easily explained by kinetic theory. The association between thermodynamics and kinetic theory can be derived by statistical mechanics.

2.1 Thermodynamic System

A thermodynamic system is a quantity of matter of fixed identity, around which we can draw a boundary. The boundaries may be fixed or moveable. Work or heat can be transferred across the system boundary. Everything outside the boundary is called the surroundings.

2.2 The Zeroth Law

The zeroth law of thermodynamics begins with a simple definition of thermodynamic equilibrium. Thermodynamic equilibrium can be observed as follow: if two of these objects are brought into physical contact there is initially a change in the property of both objects, but, eventually, the change in property stops. The objects are said to be in thermodynamic equilibrium.

2.3 The First Law

The first law of thermodynamics is the application of the conservation of energy principle to heat and thermodynamic processes. Any thermodynamic system in an equilibrium state possesses a state variable called the internal energy. Between any two equilibrium states, the change in internal energy is equal to the difference of heat transfer into the system and work done by the system.

There are three ways that heat may be transferred between substances at different temperatures: conduction, convection, and radiation. The flow of heat by conduction occurs via collisions between molecules in the substance and the subsequent transfer of translational kinetic energy. Convection is the flow of heat through a bulk, macroscopic movement of matter from a hot region to a cool region, as opposed to the microscopic transfer of heat between molecules involved with conduction. The third and last form of heat transfer is radiation, which in this context means light. For example, from the sun to the earth through mostly empty space - such a transfer cannot occur via convection nor conduction, which require the movement of material from one place to another or the collisions of molecules within the material. From above, the change of internal energy is caused by the changes of micro states of molecules.

2.4 The Second Law

The first law of thermodynamics allows for many possible states of a system to exist. But, experience indicates that only certain states occur. This leads to the second law of thermodynamics and the definition of another state variable called

entropy. Entropy can be defined as a measure of the multiplicity of a system. For a system of a large number of molecules, like a mole of molecules, the most probable state will be found in the state of highest multiplicity. Accordingly, the distribution at the time is the maximum entropy distribution.

3 Thermodynamic Model of Evolutionary Algorithm

There are many similarities between thermodynamic system and EA population. Firstly, a thermodynamic system is made up by a number of molecules, while the EA population is also made up by a number of individuals. Secondly both the molecules and individuals work randomly.

In the thermodynamic model of EA, a population is a thermodynamic system. An individual is a molecule. The chromosome of the individual presents the micro state of the molecule. The energy of the molecule is the fitness of the individual, which is decided by the problem. The sum of all individual fitness is the fitness of the population. It is the internal energy of the system.

For a minimization problem, the task is to find the micro state of the molecule with the lowest energy, which is referred as the ground state. For a given micro state, it is easy to assign an energy, but it is hard to construct a micro state for a given energy. The way to find the ground state is using the search operator.

There are two kinds of search operators in EA, especially for GA. One is mutation. The other is crossover. In the thermodynamic model of EA, mutation is similar to radiation and crossover is similar to conduction. Both mutation and crossover are means of heat transferring.

The survival strategy and update strategy in the thermodynamic model of EA are the applications of the conservation of energy principle and the conservation of mass principle. Before the application of search operator, the parents, who are selected by selection operator according to the distribution of individuals, possess some amount of energy. After the application of search operator, the children possess other amount of energy. The survivals are the whole of parent or the whole of children who possess less energy. The winers take place of the losers. The replacement leads to a new distribution of individuals and a new state of the system. The change of system state leads to the change of system energy. The change of system energy is the work done by the algorithm.

If the given time of running is long enough, the population and algorithm get into thermodynamic equilibrium. We can analyze the property, performance and efficiency of the algorithm by the termination state of the population.

The summary of thermodynamic model of evolutionary algorithm is show in Tab.1.

4 Future Work

In this paper, we set up a thermodynamic model of EA. It has a solid foundation in thermodynamics. It can be taken as a reference model of EA to compare the

Table 1. Summary of thermodynamic model of evolutionary algorithm

EA	Thermodynamics
population	system
individual	molecule
chromosome	micro state
fitness function	energy function
individual fitness	molecular energy
population fitness	internal energy
selection	distribution
crossover	conduction
mutation	radiation
optimal solution	ground state

difference between EAs. In the future, we hope we can propose a unified theory of EA based on the thermodynamic model of EA.

Acknowledgment

This paper was supported by the National Natural Science Foundation of China under Grant No.60473014, and the Specialized Research Fund for the Doctoral Program of Higher Education of China under Grant No.20030486049.

References

1. Eiben, A.E., Rudolph, G.: Theory of evolutionary algorithms: a bird's eye view. Theoretical Computer Science **229**(1–2) (1999) 3–9
2. Holland, J.: Adaptation in Natural and Artificial Systems. University of Michigan Press, Ann Arbor, USA (1975)
3. Nix, A.E., Vose, M.D.: Modeling genetic algorithms with markov chains. Annals of Mathematics and Artificial Intelligence **5**(1) (1992) 79–88
4. Goldberg, D., Deb, K., Thierens, D.: Toward a better understanding of mixing in genetic algorithms. Journal of the Society for Instrumentation and Control Engeering **32**(1) (1993) 10–16
5. Beyer, H.G.: Toward a theory of evolution strategies: the (μ,λ)-theory. Evolutionary Computation **2**(4) (1994) 381–407
6. Mühlenbein, H., Schlierkamp-Voosen, D.: Predictive models for the breeder genetic algorithm I: Continuous parameter optimization. Evolutionary Computation **1**(1) (1993) 25–49
7. Vose, M.D., Liepins, G.E.: Punctuated equilibria in genetic search. Complex Systems **5**(1) (1991) 31–44
8. Prügel-Bennett, A., Shapiro, J.L.: An analysis of genetic algorithms using statistical mechanics. Physical Review Letters **72** (1994) 1305–1309

Simulated Annealing Parallel Genetic Algorithm Based on Building Blocks Migration

Zhiyong Li and Xilu Zhu

School of computer and communication, Hunan University, China, 410082

Abstract. Through analyzing schema theorem and building blocks theory, propose parallel genetic algorithms based on building blocks migration. Relying on convergence condition, receive building blocks from other populations. Using simulated annealing method prevent the density of good schema to increase greatly which will result in premature convergence. Theory analysis and experimental results show that the method not only reduce ineffective migration and decrease communication costs, but also lower the possibility of occurring premature and assure the capability of global convergence.

Keywords: parallel genetic algorithm, schema theorem, building blocks theory, simulated annealing, migration strategy.

1 Introduction

In recent years, many specialists have been researching and made significant contributions in PGA[1]. However, How to improve algorithm performance[2] and the extensibility of PGA are still barriers blocking PAG advancing[3].

In this paper, we introduce a new migration strategy which base on building blocks theory and integrate Boltzmann survivor mechanism to reduce communication cost and enhance coarse-grained PGA searching efficiency.

2 Building Block Operator

To guarantee to find global solution, the population should contain at least one same-defining-position-non-overlapping set, which can completely cover the whole solution area.

Definition 1. If population exists a building blocks[4] set $\{H_1, H_2 \cdots H_n\}$, and satisfy $H_a \cap H_b = \varnothing$, where $\forall a, b \in \{1,2 \cdots k\}$. The intersection is null if and only if no identical value in the same locus. Therefore the set mentioned above is non-overlapping BBs set.

Set threshold θ_0, $S(1) = \sum_{j=1}^{n} x_{ij}$ where x_{ij}=1. $P(1) = S(1)/n$ .If $P(1) > \theta_0$, the value 1 is accepted to make up BBs, otherwise1-P(1)> θ_0, the value 0 is accepted to make up BBs, else accept *.

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 1183–1185, 2007.
© Springer-Verlag Berlin Heidelberg 2007

3 An Outline of Simulated Annealing Parallel Genetic Algorithm Based on Building Blocks Migration

STEP 1. Initialize population $S^L = \{0,1\}^L$ randomly, and calculate population fitness.
STEP 2. If existing migrations do follows; else return the global solution.
STEP 3. If i_{loop}, when children fitness lower than parents, we consider premature convergence in GA process, and do STEP 4.else continue GA process.
STEP 4. When GA proceeds slowly, it means that subpopulation converge to a local optimum, we extract building block H_l ,and select BBs which nearly or exactly match the principle of non-overlapping.
STEP 5. Use Metropolis strategy[5] to accept building blocks which migrate from other population. The equation is described below by:

$$p_{accept} = \begin{cases} \min(1, \exp(\frac{f_{migrator} - \overline{f}}{T})) > random[0,1) & f_{migrator} \le \overline{f} \\ 1 & f_{migrator} > \overline{f} \end{cases} \quad (1)$$

As the temperature of annealing fall, the probability of accepting bad solution diminishes. The Metropolis avoids GA to trap in local optimum.

4 Communication Costs

In coarse grained PGA, overall cost V[6] estimated casually is described below as:

$$V = [\frac{g}{MI} - 2] \times \sum_{j=1}^{m} r_i s_i b \quad (2)$$

However the overall cost of BBsAPGA described below by equation:

$$V' = \left[\frac{g}{C}\right] \times mb \quad (3)$$

C denote premature interval. Obviously, $C = nMI$.

$$\Delta V = V - V' = [\frac{g}{MI} - 2] \times \sum_{i=1}^{g} r_i s_i b - [\frac{g}{C}] \times mb \quad (4)$$

Therefore, BBsAPGA effectively lower communication cost, decrease ineffective number of migration.

5 Experimental Results

To investigate the parallel efficiency of PGA and genetic operators, we conduct experiments on a cluster. All computations are performed on HP P4(3G),512 memory.

$$f = 100 - [(4 - 2.1x^2 + \frac{x^4}{3})x^2 + xy + (-4 + 4y^2)y^2] \quad x, y \in [-2.048, 2.048] \tag{5}$$

Where the function has 6 local optimums, two of them are global optimum

$$f = \cos(5x)\cos(5y)e^{-0.001(x^2+y^2)} \quad x, y \in [-2.048, 2.048] \tag{6}$$

Where the function has many local optimum, f(0,0)=1 is the global optimum

Table 1. Experimental results

experiment	PGA migration	BBsAPGA migration	PGA convergence rate	BBsAPGA convergence rate
1	670	31	55	33
2	1170	33	95	73

The result of experiment manifest BBsAPGA effectively lower communication cost, decrease ineffective number of migration. Meanwhile this algorithm overcome premature convergence phenomena, accelerate evolution, enhance global searching.

6 Conclusions

Analyze schema theorem and building blocks theory deduced by schema theorem, introduce building block operator. Integrating Boltzmann survivor mechanism propose a simulated annealing parallel genetic algorithm based on building blocks. The BBsAPGA reduce ineffective migration, decrease cost of communication. Evidently it lower premature probability and ensure GA global convergence.

References

1. Muhlenbein, H.,Schomisch, M. and Born,J.: The parallel genetic algorithm as function optimizer. Proceedings of the Fourth International Conference on Genetic Algorithms (1991) 619–632.
2. Nowostawski M., Poli R.: Parallel genetic algorithm taxonomy. Proceedings of the Third International conference on knowledge-based intelligent information engineering systems (KES'99) (1999) 88–92.
3. Cantu-Paz E.: A survey of parallel genetic algorithms. Calculateurs Paralleles, Vol. 10. Hermes Paris (1998) 141–171.
4. Lee A.: The schema theorem and Price's theorem. Foundations of Genetic Algorithms . Morgan Kaufmann San Francisco (1995) 23–49
5. Emile Aarts, Jan Korst.: Simulated annealing and Boltzmann machines : a stochastic approach to combinatorial optimization and neural computing. John Wiley and Sons New York (1989)
6. Baowen Xu, Yu Guan, Zhenqiang Chen.: Parallel Genetic Algorithms with Schema Migration. Proceedings of 26th International Computer Software and Applications Conference (COMPSAC 2002) (2002) 879–884

Comparison of Different Integral Performance Criteria for Optimal Hydro Generator Governor Tuning with a Particle Swarm Optimization Algorithm

Hongqing Fang[1], Long Chen[2], and Zuyi Shen[2]

[1] College of Electrical Engineering, Hohai University, Nanjing, 210098 Jiangsu, P.R. China
fanghongqing@sohu.com
[2] College of Water Conservancy & Hydropower Engineering, Hohai University, Nanjing, 210098 Jiangsu, P.R. China
{cdalong,richardshen}@hhu.edu.cn

Abstract. In this paper, the particle swarm optimization (PSO) algorithm with constriction factor approach (CFA) is proposed to optimal hydro generators governor Proportional-Integral-Derivative (PID) gains for small hydraulic transients. And four different integral performance criteria of turbine speed deviation such as integrated absolute error (IAE), integral of time weighted absolute value of error (ITAE), integral of squared error (ISE) and integral of time weighted squared error (ITSE) have been taken as fitness function respectively, and the differences have been investigated and compared. A step speed disturbance test on no-load operation has been performed. The digital simulation results show that the proposed PSO method has stable convergence characteristic and good computational ability, and it can effectively optimal hydro generators governor PID gains. And the dynamic performance of hydro generators governor system for small hydraulic transients is much better if ITAE criterion is applied as fitness function.

Keywords: Particle swarm optimization; Constriction factor approach; Integral performance criteria; Hydro generators governor; PID tuning.

1 Introduction

In modern hydroelectric power plants, conventional PID controller is widely applied in hydro generators governor systems [1]. There are any methods have been reported for improving the setting performance of hydro generators governor system parameters, such as simplex method, orthogonal test method, genetic algorithm (GA) and so on [2]. However, these methods have different disadvantages. Particle swarm optimization (PSO) is characterized as a simple concept, easy to implement, and computationally efficient [3]. In fact, PSO has obtained more and more attentions in the researching areas of electric power systems [4]-[5]. In this paper, the PSO algorithm with constriction factor approach (CFA-PSO) [6] is proposed to optimal hydro generators governor PID gains for small hydraulic transients. And four different integral performance criteria of turbine speed deviation such as IAE, ITAE, ISE and ITSE have been taken as fitness function respectively, and the differences have been investigated and compared.

Y. Shi et al. (Eds.): ICCS 2007, Part IV, LNCS 4490, pp. 1186–1189, 2007.
© Springer-Verlag Berlin Heidelberg 2007

2 Hydro Generators Governor System

The typical hydro generators governor system is shown in Fig. 1, which consists of PID controller, electro-hydraulic servo system, hydro turbine system, generator and load. The definitions of the parameters, in Fig. 1, could be found in [1].

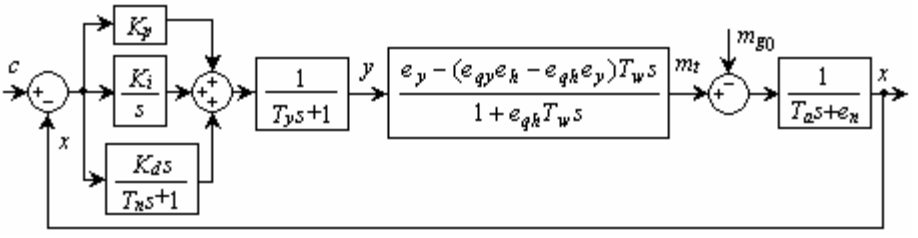


Fig. 1. Typical hydro generators governor system

3 Implementation of a PSO-PID Controller for Hydro Generators Governor System

The velocity of the *k*th dimension of the *i*th particle is updated in current iteration cycle ($t+1$) in CFA-PSO can be expressed as:

$$V_{i,k}(t+1)=\chi(V_{i,k}(t)+c_1 r_1(P_{i,k}(t)-X_{i,k}(t))+c_2 r_2(P_{g,k}(t)-X_{i,k}(t))) \quad (1)$$

And the position of the *k*th dimension of the *i*th particle is updated as:

$$X_{i,k}(t+1)=X_{i,k}(t)+V_{i,k}(t+1) \quad (2)$$

The definitions of the parameters of CFA-PSO could be found in [6]. Usually, c_1 and c_2 are set as $c_1=c_2=2.05$, thus, $\chi=0.729$.

Table 1. Parameters for hydro generators governor system

T_y	T_a	T_w	e_y	e_{qy}	e_h	e_{qh}	e_n
0.1	6.0	1.5	1.0	1.0	1.5	0.5	1.5

A PID controller optimized with the CFA-PSO algorithm was developed, which was called PSO-PID controller. The integral performance criteria (IAE, or ITAE, or ISE, or ITSE criteria) of turbine speed deviation x must be minimized by CFA-PSO algorithm. The CFA-PSO algorithm for the PID gains optimizing was implemented using Microsoft Visual Basic 6.0 programming language on a PC with Intel Pentium III 1GHz processor. The population size of CFA-PSO algorithm is 15 and the maximum allowable number of iterations is 30. The data of hydro generators governor system are given in Table 1, the time step of the simulation for small hydraulic transients is 0.02 sec, the

simulation is done for 30 sec, and the lower and upper bounds of PID gains are all set as [0, 10]. The optimizing processes have been done off-line.

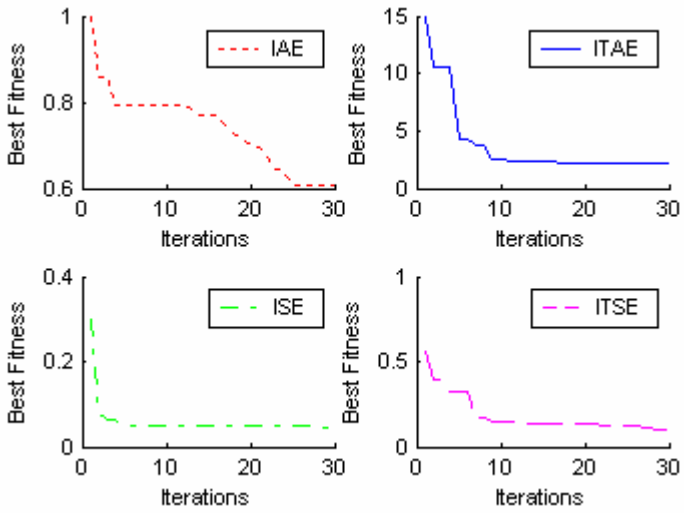


Fig. 2. Best fitness convergence

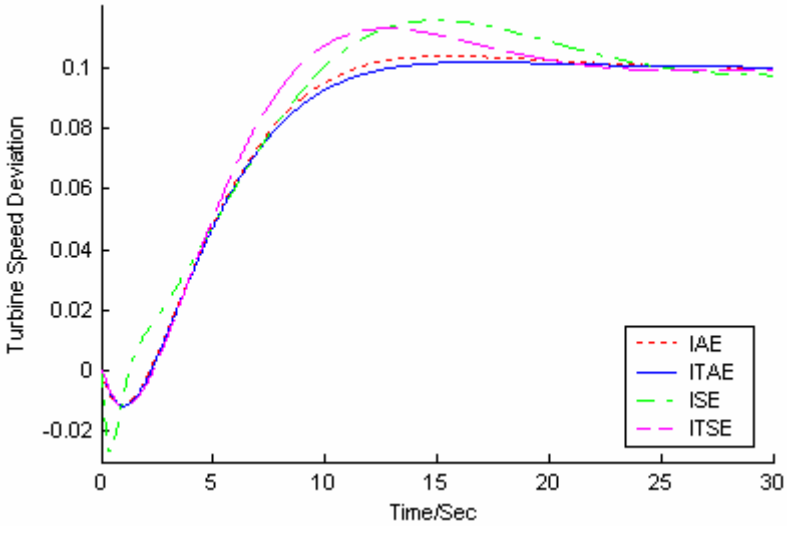


Fig. 3. Speed response to a 10% step increasing in turbine speed set point

A 10% step turbine speed deviation increasing was applied to the turbine speed reference, i.e., $c = 0.1$. Since all the four integral performance criteria of turbine speed deviation x have been taken as fitness function respectively, thus, there have four different optimal PID gains sets, shown in Table 2. From the simulations results, it

could be found that if ITAE criterion was taken as fitness function, hydro generators governor system will get the smallest overshoot, i.e., $M_p = 0.102$, and the shortest settling time, i.e., $t_s = 11.717\,\text{sec}$ .The traces of the best fitness convergence of the four integral performance criteria are shown in Fig. 2. It could be found that all of the optimizing processes have downward trends and are convergence in not more than 30 iterations. The fastest one is ITAE criterion was taken as fitness function, which is not more than 16 iterations, and the slowest one is ITSE criteria was taken as fitness function, which is not more than 29 iterations. And Fig. 3 shows the turbine speed deviation response traces using the four different optimal PID gains sets. It is clear that the hydro generators governor system has the best dynamic performance if ITAE criterion was applied as fitness function than that of IAE, ISE and ITSE criterion.

Table 2. Results for a 10% step turbine speed increasing

Criteria	K_p	K_i	K_d	T_n	M_p	t_s	J
J_{IAE}	0.921	0.279	0.066	0.007	0.104	20.735	0.606
J_{ITAE}	0.933	0.263	0.0	0.0	0.102	11.717	2.228
J_{ISE}	0.741	0.356	0.946	0.095	0.117	33.110	0.049
J_{ITSE}	0.895	0.340	0.0	0.0	0.113	20.455	0.109

4 Conclusion

This paper, has shown the effectively design method for determining the PID gains for hydro generators governor system applying CFA-PSO algorithm and it is clear that the ITAE criterion is most suitable for the application as fitness function.

References

1. IEEE Committee Report: Hydraulic turbine and turbine control models for system dynamic studies. IEEE Trans. Power Systems. 1(1992) 167–178
2. Li Z., and Malik O. P.: An orthogonal test approach based control parameter optimization and its application to a hydro-turbine governor. IEEE Trans. Energy Conversion, 4(1997) 388–393
3. Kennedy J., and Eberhart R.: Particle swarm optimization. IEEE Int. Conf. Neural Networks. Purth, Australia, (1995) 1942–1948
4. Abido M. A.: Optimal design of power-system stabilizers using particle swarm optimization. IEEE Trans. Energy Conversion. 3(2002) 406–413
5. Gaing Z. L.: Particle swarm optimization to solving the economic dispatch considering the generator constraints. IEEE Trans. Power Systems. 3(2003) 1187–1195
6. Eberhart R. C. and Shi Y.: Comparing inertia weights and constriction factors in particle swarm optimization. 2000 Congress on Evolutionary Computation. San Diego, (2000) 84–88

Author Index

Abánades, Miguel A. II-227
Abbate, Giannandrea I-842
Abdullaev, Sarvar R. IV-729
Abdullah, M. I-446
Acevedo, Liesner I-152
Adam, J.A. I-70
Adriaans, Pieter III-191, III-216
Adrianto, Indra I-1130
Agarwal, Pankaj K. I-988
Ahn, Chan-Min II-515
Ahn, Jung-Ho IV-546
Ahn, Sangho IV-360
Ahn, Sukyoung I-660
Ahn, Woo Hyun IV-941
Ahn, Young-Min II-1202, II-1222
Ai, Hongqi II-327
Al-Sammane, Ghiath II-263
Alexandrov, Vassil I-747, II-744, II-768, II-792
Alfonsi, Giancarlo I-9
Alidaee, Bahram IV-194
Aliprantis, D. I-1074
Allen, Gabrielle I-1034
Alper, Pinar II-712
Altintas, Ilkay III-182
Álvarez, Eduardo J. II-138
An, Dongun III-18
An, Sunshin IV-869
Anthes, Christoph II-752, II-776
Araz, Ozlem Uzun IV-973
Archip, Neculai I-980
Arifin, B. II-335
Aristov, V.V. I-850
Arslanbekov, Robert I-850, I-858
Arteconi, Leonardo I-358
Aslan, Burak Galip III-607
Assous, Franck IV-235
Atanassov, E. I-739
Avolio, Maria Vittoria I-866
Awan, Asad I-1205

Babik, Marian III-265
Babuška, I. I-972
Bação, Fernando II-542
Bae, Guntae IV-417
Bae, Ihn-Han IV-558
Baek, Myung-Sun IV-562
Baek, Nakhoon II-122
Bai, Yin III-1008
Bai, Zhaojun I-521
Baik, Doo-Kwon II-720
Bajaj, C. I-972
Balas, Lale I-1, I-38
Baliś, Bartosz I-390
Balogh, Zoltan III-265
Baloian, Nelson II-799
Bang, Young-Cheol III-432
Bao, Yejing III-933
Barabási, Albert-László I-1090
Barabasz, B. I-342
Barrientos, Ricardo I-229
Barros, Ricardo III-253
Baruah, Pallav K. I-603
Bashir, Omar I-1010
Bass, J. I-972
Bastiaans, R.J.M. I-947
Baumgardner, John II-386
Baytelman, Felipe II-799
Bayyana, Narasimha R. I-334
Bechhofer, Sean II-712
Beezley, Jonathan D. I-1042
Bei, Yijun I-261
Bell, M. I-1074
Belloum, Adam III-191
Bemben, Adam I-390
Benhai, Yu III-953
Benkert, Katharina I-144
Bennethum, Lynn S. I-1042
Benoit, Anne I-366, I-591
Bervoets, F. II-415
Bhatt, Tejas I-1106
Bi, Jun IV-801
Bi, Yingzhou IV-1061
Bidaut, L. I-972
Bielecka, Marzena II-970
Bielecki, Andrzej II-558
Black, Peter M. I-980
Bo, Hu IV-522

Bo, Shukui III-898
Bo, Wen III-917
Bochicchio, Ivana II-990, II-997
Bosse, Tibor II-888
Botana, Francisco II-227
Brendel, Ronny II-839
Bressler, Helmut II-752
Brewer, Wes II-386
Brill, Downey I-1058
Brooks, Christopher III-182
Browne, J.C. I-972
Bu, Jiajun I-168, I-684
Bubak, Marian I-390
Bungartz, Hans-Joachim I-708
Burguillo-Rial, Juan C. IV-466
Burrage, Kevin I-778
Burrage, Pamela I-778
Bushehrian, Omid I-599
Byon, Eunshin I-1197
Byrski, Aleksander II-928
Byun, Hyeran IV-417, IV-546
Byun, Siwoo IV-889

Cai, Guoyin II-569
Cai, Jiansheng III-313
Cai, Keke I-684
Cai, Ming II-896, III-1048, IV-725, IV-969
Cai, Ruichu IV-1167
Cai, Shaobin III-50, III-157
Cai, Wentong I-398
Cai, Yuanqiang III-1188
Caiming, Zhang II-130
Campos, Celso II-138
Cao, Kajia III-844
Cao, Rongzeng III-1032, IV-129
Cao, Suosheng II-1067
Cao, Z.W. II-363
Carmichael, Gregory R. I-1018
Caron, David I-995
Catalyurek, Umit I-1213
Cattani, Carlo II-982, II-990, II-1004
Cecchini, Arnaldo I-567
Čepulkauskas, Algimantas II-259
Cetnarowicz, Krzysztof II-920
Cha, Jeong-won IV-721
Cha, JeongHee II-1
Cha, Seung-Jun II-562
Chai, Lei IV-98
Chai, Tianfeng I-1018
Chai, Yaohui II-409
Chai, Zhenhua I-802
Chai, Zhilei I-294
Chakraborty, Soham I-1042
Chandler, Seth J. II-170
Chandola, Varun I-1222
Chang, Ok-Bae II-1139
Chang, Jae-Woo III-621
Chang, Moon Seok IV-542
Chang, Sekchin IV-636
Chang, Yoon-Seop II-562
Chaoguang, Men III-166
Chatelain, Philippe III-1122
Chaturvedi, Alok I-1106
Chawla, Nitesh V. I-1090
Che, HaoYang III-293
Chen, Bin III-653
Chen, Bing III-338
Chen, Changbo II-268
Chen, Chun I-168, I-684
Chen, Gang I-253, I-261, III-1188
Chen, Guangjuan III-984
Chen, Guoliang I-700
Chen, Jianjun I-318
Chen, Jianzhong I-17
Chen, Jiawei IV-59, IV-98
Chen, Jin I-30
Chen, Jing III-669
Chen, Juan IV-921
Chen, Ken III-555
Chen, Lei IV-1124
Chen, Ligang I-318
Chen, Liujun IV-59
Chen, Long IV-1186
Chen, Qingshan II-482
Chen, Tzu-Yi I-302
Chen, Weijun I-192
Chen, Wei Qing II-736
Chen, Xiao IV-644
Chen, Xinmeng I-418
Chen, Ying I-575
Chen, Yun-ping III-1012
Chen, Yuquan II-1186, II-1214
Chen, Zejun III-113
Chen, Zhengxin III-852, III-874
Chen, Zhenyu II-431
Cheng, Frank II-17
Cheng, Guang IV-857
Cheng, Jingde I-406, III-890
Cheng, T.C. Edwin III-338

Cheng, Xiaobei III-90
Chi, Hongmei I-723
Cho, Eunseon IV-713
Cho, Haengrae IV-753
Cho, Hsung-Jung IV-275
Cho, Jin-Woong IV-482
Cho, Ki Hyung III-813
Cho, Sang-Young IV-949
Cho, Yongyun III-236
Cho, Yookun IV-905
Choe, Yoonsik IV-668
Choi, Bum-Gon IV-554
Choi, Byung-Uk IV-737
Choi, Han-Lim I-1138
Choi, Hyoung-Kee IV-360
Choi, HyungIl II-1
Choi, Jaeyoung III-236
Choi, Jongsun III-236
Choi, Kee-Hyun II-952
Choi, Myounghoi III-508
Choo, Hyunseung I-668, II-1226, III-432, III-465, IV-303, IV-336, IV-530, IV-534, IV-538, IV-550
Chopard, Bastien I-922
Chou, Chung-I IV-1163
Choudhary, Alok III-734
Chourasia, Amit I-46
Chrisochoides, Nikos I-980
Christiand II-760
Chtepen, Maria I-454
Chu, Chao-Hsien III-762
Chu, You-ling IV-1163
Chu, Yuan-Sun II-673
Chuan, Zheng Bo II-25
Chung, Hee-Joon II-347
Chung, Hyunsook II-696
Chung, Min Young IV-303, IV-534, IV-550, IV-554
Chung, Seungjong III-18
Chung, Tai-Myoung III-1024
Chung, Yoojin IV-949
Cianni, Nathalia M. III-253
Ciarlet Jr., Patrick IV-235
Cisternino, Antonio II-585
Claeys, Filip H.A. I-454
Clark, James S. I-988
Clatz, Olivier I-980
Clegg, June IV-18
Clercx, H.J.H. I-898
Cline, Alan II-1123
Coen, Janice L. I-1042
Cofiño, A.S. III-82
Cole, Martin J. I-1002
Cong, Guodong III-960
Constantinescu, Emil M. I-1018
Corcho, Oscar II-712
Cornish, Annita IV-18
Cortial, J. I-1171
Costa-Montenegro, Enrique IV-466
Costanti, Marco II-617
Cox, Simon J. III-273
Coyle, E. I-1074
Cuadrado-Gallego, J. II-1162
Cui, Gang IV-1021
Cui, Ruihai II-331
Cui, Yifeng I-46
Cui, Yong IV-817
Curcin, Vasa III-204
Cycon, Hans L. IV-761

D'Ambrosio, Donato I-866
Dăescu, Dacian I-1018
Dai, Dao-Qing I-102
Dai, Kui IV-251
Dai, Tran Thanh IV-590
Dai, Zhifeng IV-1171
Danek, Tomasz II-558
Danelutto, Marco II-585
Dang, Sheng IV-121
Dapeng, Tan IV-957
Darema, Frederica I-955
Darmanjian, Shalom I-964
Das, Abhimanyu I-995
Day, Steven I-46
Decyk, Viktor K. I-583
Degond, Pierre I-939
Delu, Zeng IV-283
Demeester, Piet I-454
Demertzi, Melina I-1230
Demkowicz, L. I-972
Deng, An III-1172
Deng, Nai-Yang III-669, III-882
Deng, Xin Guo II-736
Dhariwal, Amit I-995
Dhoedt, Bart I-454
Di, Zengru IV-98
Díaz-Zuccarini, V. I-794
DiGiovanna, Jack I-964
Diller, K.R. I-972
Dimov, Ivan I-731, I-739, I-747

Ding, Dawei III-347
Ding, Lixin IV-1061
Ding, Maoliang III-906
Ding, Wantao III-145
Ding, Wei III-1032, IV-129, IV-857
Ding, Yanrui I-294
Ding, Yong IV-1116
Ding, Yongsheng III-74
Ding, Yu I-1197
Diniz, Pedro I-1230
Dittamo, Cristian II-585
Doboga, Flavia II-1060
Dobrowolski, Grzegorz II-944
Dong, Jinxiang I-253, I-261, II-896, II-1115, III-1048, IV-725, IV-969
Dong, Yong IV-921
Dongarra, Jack II-815
Dongxin, Lu III-129
Dostert, Paul I-1002
Douglas, Craig C. I-1002, I-1042
Downar, T. I-1074
Dreżewski, Rafał II-904, II-920
Dressler, Thomas II-831
Du, Xu IV-873
Du, Ye III-141
Duan, Gaoyan IV-1091
Duan, Jianyong II-1186
Dunn, Adam G. I-762
Dupeyrat, Gerard. IV-506

Efendiev, Yalchin I-1002
Egorova, Olga II-65
Eilertson, Eric I-1222
Elliott, A. I-972
Ellis, Carla I-988
Emoto, Kento II-601
Engelmann, Christian II-784
Eom, Jung-Ho III-1024
Eom, Young Ik IV-542, IV-977
Ertoz, Levent I-1222
Escribano, Jesús II-227
Espy, Kimberly Andrew III-859
Ewing, Richard E. I-1002

Fabozzi, Frank J. III-937
Fairman, Matthew J. III-273
Falcone, Jean-Luc I-922
Fan, Hongli III-563
Fan, Ying IV-98
Fan, Yongkai III-579
Fang, F. II-415
Fang, Fukang IV-59
Fang, Hongqing IV-1186
Fang, Hua III-859
Fang, Li Na II-736
Fang, Lide II-1067
Fang, Liu III-1048
Fang, Yu III-653
Fang, Zhijun II-1037
Fang-an, Deng III-453
Farhat, C. I-1171
Farias, Antonio II-799
Fathy, M. IV-606
Fedorov, Andriy I-980
Fei, Xubo III-244
Fei, Yu IV-741
Feixas, Miquel II-105
Feng, Huamin I-374, II-1012, III-1, III-493
Feng, Lihua III-1056
Feng, Y. I-972
Feng, Yuhong I-398
Ferrari, Edward I-1098
Fidanova, Stefka IV-1084
Field, Tony I-111
Figueiredo, Renato I-964
Fischer, Rudolf I-144
Fleissner, Sebastian I-213
Flikkema, Paul G. I-988
Flórez, Jorge II-166
Fortes, Jos A.B. I-964
Frausto-Solís, Juan II-370, IV-981
Freire, Ricardo Oliveira II-312
Frigerio, Francesco II-272
Frolova, A.A. I-850
Fu, Chong I-575
Fu, Hao III-1048
Fu, Qian I-160
Fu, Shujun I-490
Fu, Tingting IV-969
Fu, Xiaolong III-579
Fu, Yingfang IV-409
Fu, Zetian III-547
Fuentes, D. I-972
Fujimoto, R.M. I-1050
Fukushima, Masao III-937
Fürlinger, Karl II-815
Furukawa, Tomonari I-1180
Fyta, Maria I-786

Gallego, Samy I-939
Gálvez, Akemi II-211
Gang, Fang Xin II-25
Gang, Yung-Jin IV-721
Gao, Fang IV-1021
Gao, Liang III-212
Gao, Lijun II-478
Gao, Rong I-1083
Gao, Yajie III-547
Garcia, Victor M. I-152
Gardner, Henry J. I-583
Garre, M. II-1162
Garšva, Gintautas II-439
Gautier, Thierry II-593
Gava, Frédéric I-611
Gawroński, P. IV-43
Geiser, Jürgen I-890
Gelfand, Alan I-988
Georgieva, Rayna I-731
Gerndt, Michael II-815, II-847
Gerritsen, Charlotte II-888
Ghanem, Moustafa III-204
Ghattas, Omar I-1010
Gi, YongJae II-114
Gibson, Paul II-386
Gilbert, Anna C. I-1230
Goble, Carole II-712, III-182
Goda, Shinichi IV-142
Goderis, Antoon III-182
Goey, L.P.H. de I-947
Golby, Alexandra I-980
Goldberg-Zimring, Daniel I-980
Golubchik, Leana I-995
Gombos, Daniel I-1138
Gómez-Tato, A. III-637
Gong, Jian IV-809
Gong, Jianhua III-516, III-563
González-Castaño, Francisco J. III-637, IV-466
Gonzalez, Marta I-1090
Gore, Ross I-1238
Goto, Yuichi I-406
Gould, Michael II-138
Govindan, Ramesh I-995
Grama, Ananth I-1205
Gregor, Douglas I-620
Gregorio, Salvatore Di I-866
Gu, Guochang III-50, III-90, III-137, III-157, III-178
Gu, Hua-Mao III-591
Gu, Jifa IV-9
Gu, Jinguang II-728
Gu, Yanying IV-312
Guan, Ying I-270
Guang, Li III-166
Guang-xue, Yue IV-741
Guensler, R. I-1050
Guermazi, Radhouane III-773
Guibas, L.J. I-1171
Guo, Bo IV-202
Guo, Jiangyan III-370
Guo, Jianping II-538, II-569
Guo, Song III-137
Guo, Yan III-1004
Guo, Yike III-204
Guo, Zaiyi I-119
Guo, Zhaoli I-802, I-810
Gurov, T. I-739
Gutiérrez, J.M. III-82
Gyeong, Gyehyeon IV-977

Ha, Jong-Sung II-154
Ha, Pan-Bong IV-721
Haase, Gundolf I-1002
Habala, Ondrej III-265
Hachen, David I-1090
Haffegee, Adrian II-744, II-768
Hagiwara, Ichiro II-65
Hall, Mary W. I-1230
Hamadou, Abdelmajid Ben III-773
Hammami, Mohamed III-773
Hammond, Kevin II-617
Han, Houde IV-267
Han, Hyuck II-577, III-26, IV-705
Han, Jianjun I-426, IV-965
Han, Jinshu II-1091
Han, Ki-Joon I-692, II-511
Han, Ki-Jun IV-574
Han, Kyungsook I-78, I-94, II-339
Han, Lu IV-598
Han, Mi-Ryung II-347
Han, Qi-ye III-1012
Han, SeungJo III-829, IV-717
Han, Shoupeng I-1246
Han, Yehong III-444
Han, Youn-Hee IV-441
Han, Young-Ju III-1024
Han, Yuzhen III-911
Han, Zhangang IV-98
Hansen, James I-1138

Hao, Cheng IV-1005
Hao, Zhenchun IV-841
Hao, Zhifeng IV-1167
Hasan, M.K. I-326
Hasegawa, Hiroki I-914
Hatcher, Jay I-1002, I-1042
Hazle, J. I-972
He, Gaiyun II-1075
He, Jing II-401, II-409
He, Jingsha IV-409
He, Kaijian I-554, III-925
He, Tingting III-587
He, Wei III-1101
He, X.P. II-1083
He, Yulan II-378
He, Zhihong III-347
Heijst, G.J.F. van I-898
Hermer-Vazquez, Linda I-964
Hertzberger, Bob III-191
Hieu, Cao Trong IV-474
Hill, Chris I-1155, I-1163
Hill, Judith I-1010
Hinsley, Wes I-111
Hiroaki, Deguchi II-243
Hirose, Shigenobu I-914
Hluchy, Ladislav III-265
Hobbs, Bruce I-62
Hoekstra, Alfons G. I-922
Hoffmann, C. I-1074
Holloway, America I-302
Honavar, Vasant I-1066
Hong, Choong Seon IV-474, IV-590
Hong, Dong-Suk II-511
Hong, Helen II-9
Hong, Jiman IV-905, IV-925, IV-933
Hong, Soon Hyuk III-523, IV-425
Hong, Weihu III-1056
Hong, Yili I-1066
Hongjun, Yao III-611
Hongmei, Liu I-648
Horák, Bohumil II-936
Hose, D.R. I-794
Hou, Jianfeng III-313, III-320, III-448
Hou, Wenbang III-485
Hou, Y.M. III-1164
How, Jonathan I-1138
Hsieh, Chih-Hui I-1106
Hu, Bai-Jiong II-1012
Hu, Jingsong I-497
Hu, Qunfang III-1180
Hu, Ting IV-1029
Hu, Xiangpei IV-218
Hu, Xiaodong III-305
Hu, Yanmei I-17
Hu, Yi II-1186, II-1214
Hu, Yincui II-569
Hu, Yuanfang I-46
Hua, Chen Qi II-25
Hua, Kun III-867
Huajian, Zhang III-166
Huan, Zhengliang II-1029
Huang, Chongfu III-1016, III-1069
Huang, Dashan III-937
Huang, Fang II-523
Huang, Han IV-1167
Huang, Hong-Wei III-1114, III-1180
Huang, Houkuan III-645
Huang, Jing III-353
Huang, Kedi I-1246
Huang, LaiLei IV-90
Huang, Lican III-228
Huang, Linpeng II-1107
Huang, Maosong III-1105
Huang, Minfang IV-218
Huang, Mingxiang III-516
Huang, Peijie I-430
Huang, Wei II-455, II-486
Huang, Yan-Chu IV-291
Huang, Yong-Ping III-125
Huang, Yu III-257
Huang, Yue IV-1139
Huang, Z.H. II-1083
Huang, Zhou III-653
Huashan, Guo III-611
Huerta, Joaquin II-138
Huh, Eui Nam IV-498, IV-582
Huh, Moonhaeng IV-889
Hui, Liu II-130
Hunter, M. I-1050
Hur, Gi-Taek II-150
Hwang, Chih-Hong IV-227
Hwang, Hoyoung IV-889, IV-897
Hwang, Jun IV-586
Hwang, Yuan-Chu IV-433
Hwang, Yun-Young II-562

Ibrahim, H. I-446
Iglesias, Andres II-89, II-194, II-235
Inceoglu, Mustafa Murat III-607
Ipanaqué, R. II-194

Iskandarani, Mohamed I-1002
İşler, Veysi II-49
İnan, Asu I-1, I-38
Ito, Kiichi IV-74

Jackson, Peter III-746
Jacob, Robert L. I-931
Jagannathan, Suresh I-1205
Jagodziński, Janusz II-558
Jaluria, Y. I-1189
Jamieson, Ronan II-744
Jang, Hyun-Su IV-542
Jang, Sung Ho II-966
Jayam, Naresh I-603
Jeon, Jae Wook III-523, IV-425
Jeon, Keunhwan III-508
Jeon, Taehyun IV-733
Jeong, Chang Won III-170
Jeong, Dongwon II-720, III-508, IV-441
Jeong, Seung-Moon II-150
Jeong, Taikyeong T. IV-586
Jeun, In-Kyung II-665
Jho, Gunu I-668
Ji, Hyungsuk II-1222, II-1226
Ji, Jianyue III-945
Ji, Youngmin IV-869
Jia, Peifa II-956
Jia, Yan III-717, III-742
Jian, Kuodi II-855
Jian-fu, Shao III-1130
Jiang, Changjun III-220
Jiang, Dazhi IV-1131
Jiang, Hai I-286
Jiang, He III-293, III-661
Jiang, Jianguo IV-1139
Jiang, Jie III-595
Jiang, Keyuan II-393
Jiang, Liangkui IV-186
Jiang, Ming-hui IV-158
Jiang, Ping III-212
Jiang, Shun IV-129
Jiang, Xinlei III-66
Jiang, Yan III-42
Jiang, Yi I-770, I-826
Jianjun, Guo III-611
Jianping, Li III-992
Jiao, Chun-mao III-1197
Jiao, Licheng IV-1053
Jiao, Xiangmin I-334
Jiao, Yue IV-134
Jin, Hai I-434
Jin, Ju-liang III-980, III-1004
Jin, Kyo-Hong IV-721
Jin, Li II-808
Jin, Shunfu IV-210, IV-352
Jing, Lin-yan III-1004
Jing, Yixin II-720
Jing-jing, Tian III-453
Jinlong, Zhang III-953
Jo, Geun-Sik II-704
Jo, Insoon II-577
Johnson, Chris R. I-1002
Jolesz, Ferenc I-980
Jones, Brittany I-237
Joo, Su Chong III-170
Jordan, Thomas I-46
Jou, Yow-Jen IV-291
Jung, Hyungsoo III-26, IV-705
Jung, Jason J. II-704
Jung, Kwang-Ryul IV-745
Jung, Kyunghoon IV-570
Jung, Soon-heung IV-621
Jung, Ssang-Bong IV-457
Jung, Woo Jin IV-550
Jung, Youngha IV-668
Jurenz, Matthias II-839

Kabadshow, Ivo I-716
Kacher, Dan I-980
Kakehi, Kazuhiko II-601
Kalaycı, Tahir Emre II-158
Kambadur, Prabhanjan I-620
Kanaujia, Atul I-1114
Kaneko, Masataka II-178
Kang, Dazhou I-196
Kang, Hong-Koo I-692, II-511
Kang, Hyungmo IV-514
Kang, Lishan IV-1116, IV-1131
Kang, Mikyung IV-401
Kang, Min-Soo IV-449
Kang, Minseok III-432
Kang, Sanggil III-836
Kang, Seong-Goo IV-977
Kang, Seung-Seok IV-295
Kapcak, Sinan II-235
Kapoor, Shakti I-603
Karakaya, Ziya II-186
Karl, Wolfgang II-831
Kasprzak, Andrzej I-442
Kawano, Akio I-914

Kaxiras, Efthimios I-786
Ke, Lixia III-911
Keetels, G.H. I-898
Kempe, David I-995
Kennedy, Catriona I-1098
Kereku, Edmond II-847
Khan, Faraz Idris IV-498, IV-582
Khazanchi, Deepak III-806, III-852
Khonsari, A. IV-606
Ki, Hyung Joo IV-554
Kikinis, Ron I-980
Kil, Min Wook IV-614
Kim, Deok-Hwan I-204
Kim, Beob Kyun III-894
Kim, Byounghoon IV-570
Kim, Byung-Ryong IV-849
Kim, ByungChul IV-368
Kim, ChangKug IV-328
Kim, Changsoo IV-570
Kim, Cheol Min III-559
Kim, Chul-Seung IV-542
Kim, Deok-Hwan I-204, II-515, III-902
Kim, Do-Hyeon IV-449
Kim, Dong-Oh I-692, II-511
Kim, Dong-Uk II-952
Kim, Dong-Won IV-676
Kim, Eung-Kon IV-717
Kim, Gu Su IV-542
Kim, GyeYoung II-1
Kim, H.-K. I-1050
Kim, Hanil IV-660
Kim, Hojin IV-865
Kim, Hyogon IV-709
Kim, Hyun-Ki IV-457, IV-1076
Kim, Jae-gon IV-621
Kim, Jae-Kyung III-477
Kim, Jee-Hoon IV-344, IV-562
Kim, Ji-Hong IV-721
Kim, Jihun II-347
Kim, Jinhwan IV-925
Kim, Jinoh I-1222
Kim, Jong-Bok II-1194
Kim, Jong Nam III-10, III-149
Kim, Jong Tae IV-578
Kim, Joongheon IV-385
Kim, Joung-Joon I-692
Kim, Ju Han II-347
Kim, Jungmin II-696
Kim, Junsik IV-713
Kim, Kanghee IV-897
Kim, Ki-Chang IV-849
Kim, Ki-Il IV-745
Kim, Kilcheon IV-417
Kim, Kwan-Woong IV-328
Kim, Kyung-Ok II-562
Kim, LaeYoung IV-865
Kim, Minjeong I-1042
Kim, Moonseong I-668, III-432, III-465
Kim, Myungho I-382
Kim, Nam IV-713
Kim, Pankoo III-829, IV-660, IV-925
Kim, Sang-Chul IV-320
Kim, Sang-Sik IV-745
Kim, Sang-Wook IV-660
Kim, Sanghun IV-360
Kim, Sangtae I-963
Kim, Seong Baeg III-559
Kim, Seonho I-1222
Kim, Shingyu III-26, IV-705
Kim, Sung Jin III-798
Kim, Sungjun IV-869
Kim, Sung Kwon IV-693
Kim, Sun Yong IV-360
Kim, Tae-Soon III-902
Kim, Taekon IV-482
Kim, Tai-Hoon IV-693
Kim, Ung Mo III-709
Kim, Won III-465
Kim, Yong-Kab IV-328
Kim, Yongseok IV-933
Kim, Young-Gab III-1040
Kim, Young-Hee IV-721
Kisiel-Dorohinicki, Marek II-928
Kitowski, Jacek I-414
Kleijn, Chris R. I-842
Klie, Hector I-1213
Kluge, Michael II-823
Knight, D. I-1189
Knüpfer, Andreas II-839
Ko, Il Seok IV-614, IV-729
Ko, Jin Hwan I-521
Ko, Kwangsun IV-977
Koda, Masato II-447
Koh, Kern IV-913
Kolobov, Vladimir I-850, I-858
Kondo, Djimedo III-1130
Kong, Chunum IV-303
Kong, Xiangjie II-1067
Kong, Xiaohong I-278
Kong, Yinghui II-978

Kong, Youngil IV-685
Koo, Bon-Wook IV-562
Koo, Jahwan IV-538
Korkhov, Vladimir III-191
Kot, Andriy I-980
Kotulski, Leszek II-880
Kou, Gang III-852, III-874
Koumoutsakos, Petros III-1122
Koźlak, Jarosław II-872, II-944
Krile, Srecko I-628
Krishna, Murali I-603
Krömer, Pavel II-936
Kryza, Bartosz I-414
Krzhizhanovskaya, Valeria V. I-755
Kuang, Minyi IV-82
Kuijk, H.A.J.A. van I-947
Kułakowski, K. IV-43
Kulikov, Gennady Yu. I-136
Kulvietienė, Regina II-259
Kulvietis, Genadijus II-259
Kumar, Arun I-603
Kumar, Vipin I-1222
Kurc, Tahsin I-1213
Kusano, Kanya I-914
Küster, Uwe I-128
Kuszmaul, Bradley C. I-1163
Kuzumilovic, Djuro I-628
Kwak, Ho Young IV-449
Kwak, Sooyeong IV-417
Kwoh, Chee Keong II-378
Kwon, B. I-972
Kwon, Hyuk-Chul II-1170, II-1218
Kwon, Key Ho III-523, IV-425
Kwon, Ohhoon IV-913
Kwon, Ohkyoung II-577
Kyriakopoulos, Fragiskos II-625

Laat, Cees de III-191
Laclavik, Michal III-265
Laganà, Antonio I-358
Lai, C.-H. I-294
Lai, Hong-Jian III-377
Lai, K.K. III-917
Lai, Kin Keung I-554, II-423, II-455, II-486, II-494, III-925, IV-106
Landertshamer, Roland II-752, II-776
Lang, Bruno I-716
Lantz, Brett I-1090
Larson, J. Walter I-931
Laserra, Ettore II-997
Laszewski, Gregor von I-1058
Lawford, P.V. I-794
Le, Jiajin III-629
Lee, Bong Gyou IV-685
Lee, Byong-Gul II-1123
Lee, Chang-Mog II-1139
Lee, Changjin IV-685
Lee, Chung Sub III-170
Lee, Donghwan IV-385
Lee, Edward A. III-182
Lee, Eun-Pyo II-1123
Lee, Eung Ju IV-566
Lee, Eunryoung II-1170
Lee, Eunseok IV-594
Lee, Haeyoung II-73
Lee, Heejo IV-709
Lee, HoChang II-162
Lee, Hyun-Jo III-621
Lee, Hyungkeun IV-482
Lee, In-Tae IV-1076
Lee, Jae-Hyung IV-721
Lee, Jaeho III-477
Lee, Jaewoo IV-913
Lee, JaeYong IV-368
Lee, Jang-Yeon IV-482
Lee, Jin-won IV-621
Lee, Jong Sik II-966
Lee, Joonhyoung IV-668
Lee, Ju-Hong II-515, III-902
Lee, Jung-Bae IV-949
Lee, Jung-Seok IV-574
Lee, Junghoon IV-401, IV-449, I-586, IV-660, IV-925
Lee, Jungwoo IV-629
Lee, K.J. III-701
Lee, Kye-Young IV-652
Lee, Kyu-Chul II-562
Lee, Kyu Min II-952
Lee, Kyu Seol IV-566
Lee, Mike Myung-Ok IV-328
Lee, Namkyung II-122
Lee, Peter I-1098
Lee, Samuel Sangkon II-1139, III-18
Lee, Sang-Yun IV-737
Lee, SangDuck IV-717
Lee, Sang Ho III-798
Lee, Sang Joon IV-449
Lee, Seok-Lae II-665
Lee, Seok-Lyong I-204
Lee, SeungCheol III-709

Lee, Seungwoo IV-905
Lee, Seung Wook IV-578
Lee, Soojung I-676
Lee, SuKyoung IV-865
Lee, Sungyeol II-73
Lee, Tae-Jin IV-336, IV-457, IV-550, IV-554
Lee, Wan Yeon IV-709
Lee, Wonhee III-18
Lee, Wonjun IV-385
Lee, Young-Ho IV-897
Lee, Younghee IV-629
Lei, Lan III-381, III-384
Lei, Tinan III-575
Lei, Y.-X. IV-777
Leier, André I-778
Leiserson, Charles E. I-1163
Lemaire, François II-268
Lenton, Timothy M. III-273
Leung, Kwong-Sak IV-1099
Levnajić, Zoran II-633
Li, Ai-Ping III-121
Li, Aihua II-401, II-409
Li, Changyou III-137
Li, Dan IV-817, IV-841
Li, Deng-Xin III-377
Li, Deyi II-657
Li, Fei IV-785
Li, Gen I-474
Li, Guojun III-347
Li, Guorui IV-409
Li, Haiyan IV-961
Li, Hecheng IV-1159
Li, Jianping II-431, II-478, III-972
Li, Jinhai II-1067
Li, Jun III-906
Li, Li III-984
Li, Ling II-736
Li, Ming I-374, II-1012, III-1, III-493
Li, MingChu III-293, III-329
Li, Ping III-440
Li-ping, Chen IV-741
Li, Qinghua I-426, IV-965
Li, Renfa III-571
Li, Rui IV-961
Li, Ruixin III-133
Li, Runwu II-1037
Li, Sai-Ping IV-1163
Li, Shanping IV-376
Li, Shengjia III-299
Li, Shucai III-145
Li, Tao IV-166
Li, Weimin III-629
Li, Wenhang III-516
Li, X.-M. II-397
Li, Xiao-Min III-381, III-384
Li, Xikui III-1210
Li, Xin II-251, III-587
Li, Xing IV-701, IV-853
Li, Xingsen III-781, III-906
Li, Xinhui IV-174
Li, Xinmiao IV-174
Li, Xinye III-531
Li, Xiong IV-121
Li, Xiuzhen II-1021
Li, Xue-Yao III-125, III-174
Li, Xuening II-1186
Li, Xueyu II-978
Li, Xuezhen I-430
Li, Yan III-603
Li, Yanhui I-196
Li, Yi III-485
Li, Yih-Lang IV-259
Li, Yiming IV-227, IV-259
Li, Ying II-1115
Li, Yixue II-363
Li, Yiyuan II-1115
Li, Yong III-50, III-157, IV-251
Li, Yuan I-1066
Li, Yuanxiang IV-997, IV-1037, IV-1124,IV-1171, IV-1175, IV-1179
Li, Yueping III-401
Li, Yun II-327
Li, Yunwei IV-598
Li, Zhiguo I-1114
Li, Zhiyong IV-1183
Li, Zi-mao IV-1045, IV-1147
Liang, L. III-988
Liang, Liang IV-202
Liang, Xiaodong III-334
Liang, Yi I-318
Liang, Yong IV-1099
Lim, Eun-Cheon III-821
Lim, Gyu-Ho IV-721
Lim, Jongin III-1040
Lim, Kyung-Sup II-1194
Lim, S.C. I-374, II-1012, III-1
Lim, Sukhyun I-505
Lim, Sung-Soo IV-889, IV-897
Lin, Jun III-579

Lin, Yachen II-470
Lin, Zefu III-945
Lin, Zhun II-1178
Lin-lin, Ci III-539
Liñán-García, Ernesto II-370
Ling, Yun III-591
Linton, Steve II-617
Liu, Caiming II-355, IV-166
Liu, Dayou I-160
Liu, Dingsheng II-523
Liu, Dong IV-961
Liu, Dongtao IV-701
Liu, E.L. III-1151
Liu, Fang IV-1053
Liu, Feiyu III-1188
Liu, Fengli III-762
Liu, Fengshan II-33
Liu, Gan I-426, IV-965
Liu, Guizhen III-313, III-320, III-362, III-440, III-457
Liu, Guobao II-97
Liu, Guoqiang III-1062
Liu, Haibo III-90, III-178
Liu, Hailing III-1205
Liu, Han-long III-1172
Liu, Hong III-329
Liu, Hong-Cheu I-270
Liu, Hongwei IV-1021
Liu, Jia I-168, III-677
Liu, Jiangguo (James) I-882
Liu, Jiaoyao IV-877
Liu, Jin-lan III-1008
Liu, Li II-17
Liu, Liangxu III-629
Liu, Lin III-980
Liu, Lingxia III-133
Liu, Ming III-1105
Liu, Peng II-523, II-896, IV-969
Liu, Qingtang III-587
Liu, Qizhen III-575
Liu, Quanhui III-347
Liu, Sheng IV-1068
Liu, Tianzhen III-162
Liu, Weijiang IV-793
Liu, Xiaojie II-355, IV-166
Liu, Xiaoqun IV-841
Liu, Xin II-657
Liu, Xinyu I-1238
Liu, Xinyue III-661
Liu, Xiuping II-33
Liu, Yan IV-59
Liu, Yijun IV-9
Liu, Ying III-685, III-781, IV-18
Liu, Yingchun III-1205
Liu, Yuhua III-153
Liu, Yunling IV-162
Liu, Zejia III-1210
Liu, Zhen III-543
Liu, Zhi III-595
Liu, Zongtian II-689
Lobo, Victor II-542
Lodder, Robert A. I-1002
Loidl, Hans-Wolfgang II-617
Loop, B. I-1074
Lord, R. II-415
Lorenz, Eric I-922
Lou, Dingjun III-401, III-410
Loureiro, Miguel II-542
Lu, Feng III-587
Lu, J.F. III-1228
Lu, Jianjiang I-196
Lu, Jianjun IV-162
Lu, Ruzhan II-1186, II-1214
Lu, Shiyong III-244
Lü, Shunying I-632
Lu, Weidong IV-312
Lu, Yi I-1197
Lu, Yunting III-410
Lu, Zhengding II-808
Lu, Zhengtian II-355
Lu, Zhongyu III-754
Luengo, F. II-89
Łukasik, Szymon III-726
Lumsdaine, Andrew I-620
Luo, Qi III-531, III-583
Luo, Xiaonan III-485
Luo, Ying II-538, II-569
Lv, Tianyang II-97

Ma, Q. I-1189
Ma, Tieju IV-1
Ma, Xiaosong I-1058
Ma, Yinghong III-444
Ma, Yongqiang III-898
Ma, Zhiqiang III-133
Macêdo, Autran III-281
Madey, Gregory R. I-1090
Maechling, Philip I-46
Maeno, Yoshiharu IV-74
Mahinthakumar, Kumar I-1058

Mahmoudi, Babak I-964
Majer, Jonathan D. I-762
Majewska, Marta I-414
Majumdar, Amitava I-46
Malekesmaeili, Mani IV-490
Malony, Allen I-86
Mandel, Jan I-1042
Mao, Cunli IV-598
Marchal, Loris I-964
Marin, Mauricio I-229
Markowski, Marcin I-442
Marques, Vinícius III-253
Marsh, Robert III-273
Marshall, John I-1155, I-1163
Martínez-Álvarez, R.P. III-637
Martino, Rafael N. De III-253
Martinovič, Jan II-936
Mascagni, Michael I-723
Matsuzaki, Kiminori II-601, II-609
Mattos, Amanda S. de III-253
Matza, Jeff III-852
Maza, Marc Moreno II-251, II-268
McCalley, James I-1066
McGregor, Robert I-906
McMullan, Paul I-538
Mechitov, Alexander II-462
Meeker, William I-1066
Méhats, Florian I-939
Mei, Hailiang III-424
Meire, Silvana G. II-138
Melchionna, Simone I-786
Meliopoulos, S. I-1074
Melnik, Roderick V.N. I-834
Memarsadeghi, Nargess II-503
Memik, Gokhan III-734
Meng, Fanjun II-478
Meng, Huimin III-66
Meng, Jixiang III-334
Meng, Wei III-299
Meng, Xiangyan IV-598
Merkevičius, Egidijus II-439
Metaxas, Dimitris I-1114
Miao, Jia-Jia III-121
Miao, Qiankun I-700
Michopoulos, John G. I-1180
Mikucioniene, Jurate II-259
Min, Jun-Ki I-245
Min, Sung-Gi IV-441
Ming, Ai IV-522
Minster, Bernard I-46
Mirabedini, Seyed Javad II-960
Missier, Paolo II-712
Mok, Tony Shu Kam IV-1099
Molinari, Marc III-273
Montanari, Luciano II-272
Monteiro Jr., Pedro C.L. III-253
Moon, Jongbae I-382
Moon, Kwang-Seok III-10
Moore, Reagan I-46
Mora, P. III-1156
Morimoto, Shoichi II-1099, III-890
Morozov, I. III-199
Morra, Gabriele III-1122
Moshkovich, Helen II-462
Mount, David M. II-503
Mu, Chengpo I-490
Mu, Weisong III-547
Mulder, Wico III-216
Mun, Sung-Gon IV-538
Mun, Youngsong I-660, IV-514
Müller, Matthias II-839
Munagala, Kamesh I-988
Muntean, Ioan Lucian I-708
Murayama, Yuji II-550

Nagel, Wolfgang E. II-823, II-839
Nah, HyunChul II-162
Nakajima, Kengo III-1085
Nakamori, Yoshiteru IV-1
Nam, Junghyun III-709
Nara, Shinsuke I-406
Narayanan, Ramanathan III-734
Narracott, A.J. I-794
Nawarecki, Edward II-944
Nedjalkov, M. I-739
Nepal, Chirag I-78, I-94
Ni, Jun III-34
Nicole, Denis A. III-273
Niemegeers, Ignas IV-312
Niennattrakul, Vit I-513
Nieto-Yáñez, Alma IV-981
Ning, Zhuo IV-809
Niu, Ben II-319
Niu, Ke III-677
Niu, Wenyuan IV-9
Noh, Dong-Young II-347
Nong, Xiao IV-393
Noorbatcha, I. II-335
Norris, Boyana I-931

Oberg, Carl I-995
Oden, J.T. I-972
Oh, Hyukjun IV-933
Oh, Jehwan IV-594
Oh, Sangchul IV-713
Oh, Sung-Kwun IV-1076, IV-1108
Ohsawa, Yukio IV-74, IV-142
Oijen, J.A. van I-947
Oladunni, Olutayo O. I-176
Oliveira, Suely I-221
Olsen, Kim I-46
Olson, David L. II-462
Ong, Everest T. I-931
Ong, Hong II-784
Oosterlee, C.W. II-415
Ord, Alison I-62
Othman, M. I-326, I-446
Ou, Zhuoling III-162
Ould-Khaoua, M. IV-606
Ouyang, Song III-289
Özışıkyılmaz, Berkin III-734

Pacifici, Leonardo I-358
Paik, Juryon III-709
Palkow, Mark IV-761
Pan, Wei II-268
Pan, Yaozhong III-1069
Pang, Jiming II-97
Pang, Yonggang III-117, III-141
Papancheva, Rumyana I-747
Parashar, Manish I-1213
Parhami, Behrooz IV-67
Park, Ae-Soon IV-745
Park, Byungkyu II-339
Park, Chiwoo I-1197
Park, Dong-Hyun IV-344
Park, Gyung-Leen IV-449, IV-586, IV-660, IV-925
Park, Hee-Geun II-1222
Park, Heum II-1218
Park, Hyungil I-382
Park, Ilkwon IV-546
Park, Jaesung IV-629
Park, Jeonghoon IV-336
Park, Ji-Hwan III-523, IV-425
Park, JongAn III-829, IV-717
Park, Keon-Jun IV-1108
Park, ManKyu IV-368
Park, Moonju IV-881
Park, Mu-Hun IV-721
Park, Namhoon IV-713
Park, Sanghun I-25
Park, Seon-Ho III-1024
Park, Seongjin II-9
Park, So-Jeong IV-449
Park, Sooho I-1138
Park, Sungjoon III-836
Park, TaeJoon IV-368
Park, Woojin IV-869
Park, Youngsup II-114
Parsa, Saeed I-599
Paszyński, M. I-342, II-912
Pathak, Jyotishman I-1066
Pawling, Alec I-1090
Pedrycz, Witold IV-1108
Pei, Bingzhen II-1214
Pei-dong, Zhu IV-393
Pein, Raoul Pascal III-754
Peiyu, Li IV-957
Peng, Dongming III-859
Peng, Hong I-430, I-497
Peng, Lingxi II-355, IV-166
Peng, Qiang II-57
Peng, Shujuan IV-997
Peng, Xia III-653
Peng, Xian II-327
Peng, Yi III-852, III-874
Peng, Yinqiao II-355
Pflüger, Dirk I-708
Pinheiro, Wallace A. III-253
Plale, Beth I-1122
Platoš, Jan II-936
Prasad, R.V. IV-312
Price, Andrew R. III-273
Primavera, Leonardo I-9
Prudhomme, S. I-972
Príncipe, José C. I-964
Pu, Liang III-867
Pusca, Stefan II-1053

Qi, Jianxun III-984
Qi, Li I-546
Qi, Meibin IV-1139
Qi, Shanxiang I-529
Qi, Yutao IV-1053
Qiao, Daji I-1066
Qiao, Jonathan I-237
Qiao, Lei III-615
Qiao, Yan-Jiang IV-138
Qin, Jun IV-1045

Qin, Ruiguo III-599
Qin, Ruxin III-669
Qin, Xiaolin II-1131
Qin, Yong IV-67, IV-1167
Qiu, Guang I-684
Qiu, Jieshan II-280
Qiu, Yanxia IV-598
Qizhi, Zhu III-1130
Queiroz, José Rildo de Oliveira II-304
Quirós, Ricardo II-138

Ra, Sang-Dong II-150
Rafirou, D. I-794
Rajashekhar, M. I-1171
Ram, Jeffrey III-244
Ramakrishnan, Lavanya I-1122
Ramalingam, M. II-288
Ramasami, K. II-288
Ramasami, Ponnadurai II-296
Ramsamy, Priscilla II-744, II-768
Ranjithan, Ranji I-1058
Ratanamahatana, Chotirat Ann I-513
Rattanatamrong, Prapaporn I-964
Ravela, Sai I-1147, I-1155
Regenauer-Lieb, Klaus I-62
Rehn, Veronika I-366
Rejas, R. II-1162
ReMine, Walter II-386
Ren, Lihong III-74
Ren, Yi I-462, I-466, II-974
Ren, Zhenhui III-599
Reynolds Jr., Paul F. I-1238
Richman, Michael B. I-1130
Rigau, Jaume II-105
Robert, Yves I-366, I-591
Roberts, Ron I-1066
Roch, Jean-Louis II-593
Rocha, Gerd Bruno II-312
Rodríguez, D. II-1162
Rodríguez-Hernández, Pedro S. III-637, IV-466
Román, E.F. II-370
Romero, David II-370
Romero, Luis F. I-54
Rong, Haina IV-243, IV-989
Rong, Lili IV-178
Rongo, Rocco I-866
Rossman, T. I-1189
Roy, Abhishek I-652
Roy, Nicholas I-1138
Ruan, Jian IV-251
Ruan, Qiuqi I-490
Ruan, Youlin I-426, IV-965
Ryan, Sarah I-1066
Ryu, Jae-hong IV-676
Ryu, Jihyun I-25
Ryu, Jung-Pil IV-574
Ryu, Kwan Woo II-122

Sabatka, Alan III-852
Safaei, F. IV-606
Sainz, Miguel A. II-166
Salman, Adnan I-86
Saltz, Joel I-1213
Sameh, Ahmed I-1205
San-Martín, D. III-82
Sanchez, Justin C. I-964
Sánchez, Ruiz Luis M. II-1004
Sandu, Adrian I-1018, I-1026
Sanford, John II-386
Santone, Adam I-1106
Sarafian, Haiduke II-203
Sarafian, Nenette II-203
Savchenko, Maria II-65
Savchenko, Vladimir II-65
Saxena, Navrati I-652
Sbert, Mateu II-105, II-166
Schaefer, R. I-342
Scheuermann, Peter III-781
Schmidt, Thomas C. IV-761
Schoenharl, Timothy I-1090
Schost, Éric II-251
Schwan, K. I-1050
Scott, Stephen L. II-784
Seinfeld, John H. I-1018
Sekiguchi, Masayoshi II-178
Senel, M. I-1074
Senthilkumar, Ganapathy I-603
Seo, Dong Min III-813
Seo, Kwang-deok IV-621
Seo, SangHyun II-114, II-162
Seo, Young-Hoon II-1202, II-1222
Seshasayee, B. I-1050
Sha, Jing III-220
Shakhov, Vladimir V. IV-530
Shan, Jiulong I-700
Shan, Liu III-953
Shan-shan, Li IV-393
Shang, Weiping III-305
Shanzhi, Chen IV-522

Shao, Feng I-253
Shao, Huagang IV-644
Shao, Xinyu III-212
Shao, Ye-Hong III-377
Shao-liang, Peng IV-393
Sharif, Hamid III-859
Sharma, Abhishek I-995
Sharma, Raghunath I-603
Shen, Huizhang IV-51
Shen, Jing III-90, III-178
Shen, Linshan III-1077
Shen, Xianjun IV-1171, IV-1175, IV-1179
Shen, Yue III-109, III-555
Shen, Zuyi IV-1186
Shi, Baochang I-802, I-810, I-818
Shi, Bing III-615
Shi, Dongcai II-1115
Shi, Haihe III-469
Shi, Huai-dong II-896
Shi, Jin-Qin III-591
Shi, Xiquan II-33
Shi, Xuanhua I-434
Shi, Yaolin III-1205
Shi, Yong II-401, II-409, II-490, II-499, III-685, III-693, III-852, III-874, III-906, III-1062
Shi, Zhongke I-17
Shi-hua, Ma I-546
Shim, Choon-Bo III-821
Shima, Shinichiro I-914
Shin, Byeong-Seok I-505
Shin, Dong-Ryeol II-952
Shin, In-Hye IV-449, IV-586, IV-925
Shin, Jae-Dong IV-693
Shin, Jitae I-652
Shin, Kwonseung IV-534
Shin, Kyoungho III-236
Shin, Seung-Eun II-1202, II-1222
Shin, Teail IV-514
Shin, Young-suk II-81
Shindin, Sergey K. I-136
Shirayama, Susumu II-649
Shiva, Mohsen IV-490
Shouyang, Wang III-917
Shuai, Dianxun IV-1068
Shuang, Kai IV-785
Shukla, Pradyumn Kumar I-310, IV-1013
Shulin, Zhang III-992
Shuping, Wang III-992
Silva, Geraldo Magela e II-304
Simas, Alfredo Mayall II-312
Simmhan, Yogesh I-1122
Simon, Gyorgy I-1222
Simutis, Rimvydas II-439
Siricharoen, Waralak V. II-1155
Sirichoke, J. I-1050
Siwik, Leszek II-904
Siy, Harvey III-790
Skelcher, Chris I-1098
Skomorowski, Marek II-970
Słota, Damian I-184
Snášel, Václav II-936
Śnieżyński, Bartłomiej II-864
Soberon, Xavier II-370
Sohn, Bong-Soo I-350
Sohn, Won-Sung III-477
Soltan, Mehdi IV-490
Song, Hanna II-114
Song, Huimin III-457
Song, Hyoung-Kyu IV-344, IV-562
Song, Jeong Young IV-614
Song, Jae-Won III-902
Song, Joo-Seok II-665
Song, Sun-Hee II-150
Song, Wang-Cheol IV-925
Song, Xinmin III-1062
Song, Zhanjie II-1029, II-1075
Sorge, Volker I-1098
Souza, Jano M. de III-253
Spataro, William I-866
Spiegel, Michael I-1238
Sreepathi, Sarat I-1058
Srinivasan, Ashok I-603
Srovnal, Vilém II-936
Stafford, R.J. I-972
Stauffer, Beth I-995
Steder, Michael I-931
Sterna, Kamil I-390
Stransky, S. I-1155
Strug, Barbara II-880
Su, Benyue II-41
Su, Fanjun IV-773
Su, Hui-Kai IV-797
Su, Hung-Chi I-286
Su, Liang III-742
Su, Sen IV-785
Su, Zhixun II-33
Subramaniam, S. I-446

Succi, Sauro I-786
Sugiyama, Toru I-914
Suh, W. I-1050
Sui, Yangyi III-579
Sukhatme, Gaurav I-995
Sulaiman, J. I-326
Sun, De'an III-1138
Sun, Feixian II-355
Sun, Guangzhong I-700
Sun, Guoqiang IV-773
Sun, Haibin II-531
Sun, Jin II-1131
Sun, Jun I-278, I-294
Sun, Lijun IV-218
Sun, Miao II-319
Sun, Ping III-220
Sun, Shaorong IV-134
Sun, Shuyu I-755, I-890
Sun, Tianze III-579
Sun, Xiaodong IV-134
Šuvakov, Milovan II-641
Swain, E. I-1074
Swaminathan, J. II-288
Szabó, Gábor I-1090
Szczepaniak, Piotr II-219
Szczerba, Dominik I-906
Székely, Gábor I-906

Tabik, Siham I-54
Tackley, Paul III-1122
Tadić, Bosiljka II-633, II-641
Tadokoro, Yuuki II-178
Tahar, Sofiène II-263
Tak, Sungwoo IV-570
Takahashi, Isao I-406
Takato, Setsuo II-178
Takeda, Kenji III-273
Tan, Guoxin III-587
Tan, Hui I-418
Tan, Jieqing II-41
Tan, Yu-An III-567
Tan, Zhongfu III-984
Tang, Fangcheng IV-170
Tang, J.M. I-874
Tang, Jiong I-1197
Tang, Liqun III-1210
Tang, Sheng Qun II-681, II-736
Tang, Xijin IV-35, IV-150
Tang, Yongning IV-857
Tao, Chen III-953
Tao, Jianhua I-168
Tao, Jie II-831
Tao, Yongcai I-434
Tao, Zhiwei II-657
Tay, Joc Cing I-119
Terpstra, Frank III-216
Teshnehlab, Mohammad II-960
Theodoropoulos, Georgios I-1098
Thijsse, Barend J. I-842
Thrall, Stacy I-237
Thurner, Stefan II-625
Tian, Chunhua III-1032, IV-129
Tian, Fengzhan III-645
Tian, Yang III-611
Tian, Ying-Jie III-669, III-693, III-882
Ting, Sun III-129
Tiyyagura, Sunil R. I-128
Tobis, Michael I-931
Tokinaga, Shozo IV-162
Toma, Ghiocel II-1045
Tong, Hengqing III-162
Tong, Qiaohui III-162
Tong, Weiqin III-42
Tong, Xiao-nian IV-1147
Top, P. I-1074
Trafalis, Theodore B. I-176, I-1130
Treur, Jan II-888
Trinder, Phil II-617
Trunfio, Giuseppe A. I-567, I-866
Tsai, Wu-Hong II-673
Tseng, Ming-Te IV-275
Tsoukalas, Lefteri H. I-1074, I-1083
Tucker, Don I-86
Turck, Filip De I-454
Turovets, Sergei I-86

Uchida, Makoto II-649
Uğur, Aybars II-158
Ülker, Erkan II-49
Unold, Olgierd II-1210
Urbina, R.T. II-194
Uribe, Roberto I-229
Urmetzer, Florian II-792

Vaidya, Binod IV-717
Valuev, I. III-199
Vanrolleghem, Peter A. I-454
Vasenkov, Alex I-858
Vasyunin, Dmitry III-191
Vehí, Josep II-166

Veloso, Renê Rodrigues III-281
Venkatasubramanian, Venkat I-963
Venuvanalingam, P. II-288
Vermolen, F.J. I-70
Vías, Jesús M. I-54
Vidal, Antonio M. I-152
Viswanathan, M. III-701
Vivacqua, Adriana S. III-253
Vodacek, Anthony I-1042
Volkert, Jens II-752, II-776
Vuik, C. I-874
Vumar, Elkin III-370

Waanders, Bart van Bloemen I-1010
Wagner, Frédéric II-593
Wählisch, Matthias IV-761
Walentyński, Ryszard II-219
Wan, Wei II-538, II-569
Wang, Aibao IV-825
Wang, Bin III-381, III-384
Wang, Chao I-192
Wang, Chuanxu IV-186
Wang, Daojun III-516
Wang, Dejun II-1107
Wang, Haibo IV-194
Wang, Hanpin III-257
Wang, Honggang III-859
Wang, Hong Moon IV-578
Wang, Huanchen IV-51
Wang, Huiqiang III-117, III-141, III-1077
Wang, J.H. III-1164, III-1228
Wang, Jiabing I-497
Wang, Jian III-1077
Wang, Jian-Ming III-1114
Wang, Jiang-qing IV-1045, IV-1147
Wang, Jianmin I-192
Wang, Jianqin II-569
Wang, Jihui III-448
Wang, Jilong IV-765
Wang, Jing III-685
Wang, Jinping I-102
Wang, Jue III-964
Wang, Jun I-462, I-466, II-974
Wang, Junlin III-1214
Wang, Liqiang III-244
Wang, Meng-dong III-1008
Wang, Naidong III-1146
Wang, Ping III-389
Wang, Pu I-1090
Wang, Qingquan IV-178
Wang, Shengqian II-1037
Wang, Shouyang II-423, II-455, II-486, III-925, III-933, III-964, IV-106
Wang, Shuliang II-657
Wang, Shuo M. III-1205
Wang, Shuping III-972
Wang, Tianyou III-34
Wang, Wei I-632
Wang, Weinong IV-644
Wang, Weiwu IV-997, IV-1179
Wang, Wenqia I-490
Wang, Wu III-174
Wang, Xianghui III-105
Wang, Xiaojie II-1178
Wang, Xiaojing II-363
Wang, Xin I-1197
Wang, Xing-wei I-575
Wang, Xiuhong III-98
Wang, Xun III-591
Wang, Ya III-153
Wang, Yi I-1230
Wang, Ying II-538, II-569
Wang, You III-1101
Wang, Youmei III-501
Wang, Yun IV-138
Wang, Yuping IV-1159
Wang, Yunfeng III-762
Wang, Zheng IV-35, IV-218
Wang, Zhengning II-57
Wang, Zhengxuan II-97
Wang, Zhiying IV-251
Wang, Zuo III-567
Wangc, Kangjian I-482
Warfield, Simon K. I-980
Wasynczuk, O. I-1074
Wei, Anne. IV-506
Wei, Guozhi. IV-506
Wei, Lijun II-482
Wei, Liu II-146
Wei, Liwei II-431
Wei, Wei II-538
Wei, Wu II-363
Wei, Yi-ming III-1004
Wei, Zhang III-611
Weihrauch, Christian I-747
Weimin, Xue III-551
Weissman, Jon B. I-1222
Wen, Shi-qing III-1172
Wendel, Patrick III-204

Wenhong, Xia III-551
Whalen, Stephen I-980
Whangbo, T.K. III-701
Wheeler, Mary F. I-1213
Wibisono, Adianto III-191
Widya, Ing III-424
Wilhelm, Alexander II-752
Willcox, Karen I-1010
Winter, Victor III-790
Wojdyła, Marek II-558
Wong, A. I-1155
Woods, John I-111
Wu, Cheng-Shong IV-797
Wu, Chuansheng IV-1116
Wu, Chunxue IV-773
Wu, Guowei III-419
Wu, Hongfa IV-114
Wu, Jiankun II-1107
Wu, Jian-Liang III-320, III-389, III-457
Wu, Jianpin IV-801
Wu, Jianping IV-817, IV-833
Wu, Kai-ya III-980
Wu, Lizeng II-978
Wu, Qiuxin III-397
Wu, Quan-Yuan I-462, I-466, II-974, III-121
Wu, Ronghui III-109, III-571
Wu, Tingzeng III-397
Wu, Xiaodan III-762
Wu, Xu-Bo III-567
Wu, Yan III-790
Wu, Zhendong IV-376
Wu, Zheng-Hong III-493
Wu, Zhijian IV-1131
Wyborn, D. III-1156

Xexéo, Geraldo III-253
Xi, Lixia IV-1091
Xia, Jingbo III-133
Xia, L. II-1083
Xia, Na IV-1139
Xia, Xuewen IV-1124
Xia, ZhengYou IV-90
Xian, Jun I-102
Xiang, Li III-1138
Xiang, Pan II-25
Xiao, Hong III-113
Xiao, Ru Liang II-681, II-736
Xiao, Wenjun IV-67
Xiao, Zhao-ran III-1214
Xiao-qun, Liu I-546
Xiaohong, Pan IV-957
Xie, Lizhong IV-801
Xie, Xuetong III-653
Xie, Yi I-640
Xie, Yuzhen II-268
Xin-sheng, Liu III-453
Xing, Hui Lin III-1093, III-1146, III-1151, III-1156, III-1205
Xing, Wei II-712
Xing, Weiyan IV-961
Xiong, Liming III-329, III-397
Xiong, Shengwu IV-1155
Xiuhua, Ji II-130
Xu, B. III-1228
Xu, Chao III-1197
Xu, Chen III-571
Xu, Cheng III-109
Xu, H.H. III-1156
Xu, Hao III-289
Xu, Hua II-956
Xu, Jingdong IV-877
Xu, Kaihua III-153
Xu, Ke IV-506
Xu, Ning IV-1155
Xu, Wei III-964
Xu, Wenbo I-278, I-294
Xu, X.P. III-988
Xu, Xiaoshuang III-575
Xu, Y. I-1074
Xu, Yang II-736
Xu, Yaquan IV-194
Xu, You Wei II-736
Xu, Zhaomin IV-725
Xu, Zhenli IV-267
Xue, Gang III-273
Xue, Jinyun III-469
Xue, Lianqing IV-841
Xue, Wei I-529
Xue, Yong II-538, II-569

Yamamoto, Haruyuki III-1146
Yamashita, Satoshi II-178
Yan, Hongbin IV-1
Yan, Jia III-121
Yan, Nian III-806
Yan, Ping I-1090
Yan, Shi IV-522
Yang, Bo III-1012
Yang, Chen III-603

Yang, Chuangxin I-497
Yang, Chunxia IV-114
Yang, Deyun II-1021, II-1029, II-1075
Yang, Fang I-221
Yang, Fangchun IV-785
Yang, Hongxiang II-1029
Yang, Jack Xiao-Dong I-834
Yang, Jianmei IV-82
Yang, Jihong Ou I-160
Yang, Jincai IV-1175
Yang, Jong S. III-432
Yang, Jun I-988, II-57
Yang, Kyoung Mi III-559
Yang, Lancang III-615
Yang, Seokyong IV-636
Yang, Shouyuan II-1037
Yang, Shuqiang III-717
Yang, Weijun III-563
Yang, Wu III-611
Yang, Xiao-Yuan III-677
Yang, Xuejun I-474, IV-921
Yang, Y.K. III-701
Yang, Young-Kyu IV-660
Yang, Zhenfeng III-212
Yang, Zhongzhen III-1000
Yang, Zong-kai III-587, IV-873
Yao, Kai III-419, III-461
Yao, Lin III-461
Yao, Nianmin III-50, III-66, III-157
Yao, Wenbin III-50, III-157
Yao, Yangping III-1146
Yazici, Ali II-186
Ye, Bin I-278
Ye, Dong III-353
Ye, Liang III-539
Ye, Mingjiang IV-833
Ye, Mujing I-1066
Yen, Jerome I-554
Yeo, Sang-Soo IV-693
Yeo, So-Young IV-344
Yeom, Heon Y. II-577, III-26, IV-705
Yi, Chang III-1069
Yi, Huizhan IV-921
Yi-jun, Chen IV-741
Yi, Sangho IV-905
Yim, Jaegeol IV-652
Yim, Soon-Bin IV-457
Yin, Jianwei II-1115
Yin, Peipei I-192
Yin, Qingbo III-10, III-149
Ying, Weiqin IV-997, IV-1061, IV-1124, IV-1179
Yongqian, Lu III-166
Yongtian, Yang III-129
Yoo, Gi-Hyoung III-894
Yoo, Jae-Soo II-154, III-813
Yoo, Kwan-Hee II-154
Yoon, Ae-sun II-1170
Yoon, Jungwon II-760
Yoon, KyungHyun II-114, II-162
Yoon, Seokho IV-360
Yoon, Seok Min IV-578
Yoon, Won Jin IV-550
Yoshida, Taketoshi IV-150
You, Jae-Hyun II-515
You, Kang Soo III-894
You, Mingyu I-168
You, Xiaoming IV-1068
You, Young-Hwan IV-344
Youn, Hee Yong IV-566
Yu, Baimin III-937
Yu, Beihai III-960
Yu, Chunhua IV-1139
Yu, Fei III-109, III-555, III-571
Yu, Jeffrey Xu I-270
Yu, Lean II-423, II-486, II-494, III-925, III-933, III-937, IV-106
Yu, Li IV-1175
Yu, Shao-Ming IV-227, IV-259
Yu, Shun-Zheng I-640
Yu, Weidong III-98
Yu, Xiaomei I-810
Yu-xing, Peng IV-393
Yu, Zhengtao IV-598
Yuan, Jinsha II-978, III-531
Yuan, Soe-Tsyr IV-433
Yuan, Xu-chuan IV-158
Yuan, Zhijian III-717
Yuanjun, He II-146
Yue, Dianmin III-762
Yue, Guangxue III-109, III-555, III-571
Yue, Wuyi IV-210, IV-352
Yue, Xin II-280
Yuen, Dave A. I-62
Yuen, David A. III-1205

Zabelok, S.A. I-850
Zain, Abdallah Al II-617
Zain, S.M. II-335
Zaki, Mohamed H. II-263

Zambreno, Joseph III-734
Zand, Mansour III-790
Zapata, Emilio L. I-54
Zechman, Emily I-1058
Zeleznikow, John I-270
Zeng, Jinquan II-355, IV-166
Zeng, Qingcheng III-1000
Zeng, Z.-M. IV-777
Zha, Hongyuan I-334
Zhan, Mingquan III-377
Zhang, Bin I-286, I-995
Zhang, CaiMing II-17
Zhang, Chong II-327
Zhang, Chunyuan IV-961
Zhang, Defu II-482
Zhang, Dong-Mei III-1114
Zhang, Fangfeng IV-59
Zhang, Gexiang IV-243, IV-989
Zhang, Guang-Zheng I-78
Zhang, Guangsheng III-220
Zhang, Guangzhao IV-825
Zhang, Guoyin III-105
Zhang, H.R. III-1223
Zhang, J. III-1093
Zhang, Jing IV-765
Zhang, Jingping II-319, II-331
Zhang, Jinlong III-960
Zhang, Juliang II-499
Zhang, Keliang II-409
Zhang, L.L. III-1164
Zhang, Li III-599
Zhang, Li-fan I-562
Zhang, Lihui III-563
Zhang, Lin I-1026
Zhang, Lingling III-906
Zhang, Lingxian III-547
Zhang, Miao IV-833
Zhang, Min-Qing III-677
Zhang, Minghua III-58
Zhang, Nan IV-35
Zhang, Nevin L. IV-26
Zhang, Peng II-499
Zhang, Pengzhu IV-174
Zhang, Qi IV-1139
Zhang, Ru-Bo III-125, III-174
Zhang, Shenggui III-338
Zhang, Shensheng III-58
Zhang, Sumei III-448
Zhang, Weifeng II-1147
Zhang, Wen III-964, IV-150
Zhang, Xi I-1213
Zhang, Xia III-362
Zhang, XianChao III-293, III-661
Zhang, Xiangfeng III-74
Zhang, Xiaoguang IV-1091
Zhang, Xiaoping III-645
Zhang, Xiaoshuan III-547
Zhang, Xuan IV-701, IV-853
Zhang, Xueqin III-615
Zhang, Xun III-933, III-964, III-1032
Zhang, Y. I-972
Zhang, Y.M. III-1223
Zhang, Yafei I-196
Zhang, Yan I-632
Zhang, Ying I-474
Zhang, Yingchao IV-114
Zhang, Yingzhou II-1147
Zhang, Zhan III-693
Zhang, Zhen-chuan I-575
Zhang, Zhiwang II-490
Zhangcan, Huang IV-1005
Zhao, Chun-feng III-1197
Zhao, Guosheng III-1077
Zhao, Hui IV-166
Zhao, Jidi IV-51
Zhao, Jie III-984
Zhao, Jijun II-280
Zhao, Jinlou III-911
Zhao, Kun III-882
Zhao, Liang III-583
Zhao, Ming I-964
Zhao, Ming-hua III-1101
Zhao, Qi IV-877
Zhao, Qian III-972
Zhao, Qiang IV-1021
Zhao, Qingguo IV-853
Zhao, Ruiming III-599
Zhao, Wen III-257
Zhao, Wentao III-42
Zhao, Xiuli III-66
Zhao, Yan II-689
Zhao, Yaolong II-550
Zhao, Yongxiang IV-1155
Zhao, Zhiming III-191, III-216
Zhao, Zun-lian III-1012
Zheng, Bojin IV-1029, IV-1037, IV-1171, IV-1179
Zheng, Di I-462, I-466, II-974
Zheng, Jiping II-1131
Zheng, Lei II-538, II-569

Zheng, Rao IV-138
Zheng, Ruijuan III-117
Zheng, SiYuan II-363
Zheng, Yao I-318, I-482
Zheng, Yujun III-469
Zhengfang, Li IV-283
Zhiheng, Zhou IV-283
Zhong, Shaobo II-569
Zhong-fu, Zhang III-453
Zhou, Bo I-196
Zhou, Deyu II-378
Zhou, Jieping III-516
Zhou, Ligang II-494
Zhou, Lin III-685
Zhou, Peiling IV-114
Zhou, Wen II-689
Zhou, Xiaojie II-33
Zhou, Xin I-826
Zhou, Zongfang III-1062
Zhu, Aiqing III-555
Zhu, Changqian II-57
Zhu, Chongguang III-898
Zhu, Egui III-575
Zhu, Jianhua II-1075
Zhu, Jiaqi III-257
Zhu, Jing I-46
Zhu, Meihong II-401
Zhu, Qiuming III-844
Zhu, Weishen III-145
Zhu, Xilu IV-1183
Zhu, Xingquan III-685, III-781
Zhu, Yan II-1067
Zhuang, Dong IV-82
Zienkiewicz, O.C. III-1105
Zong, Yu III-661
Zou, Peng III-742
Zuo, Dong-hong IV-873
Zurita, Gustavo II-799