ELECTRONIC

MATERIALS

SCIENCE

Metal
Gate Dielectric

n+   n+

CB

VB

Eugene A. Irene

ELECTRONIC
MATERIALS SCIENCE

ELECTRONIC MATERIALS SCIENCE

Eugene A. Irene

University of North Carolina
Chapel Hill, North Carolina

CONTENTS

PREFACE

Starting in the 1960s the field of materials science has undergone significant changes, from a field derived largely from well-established disciplines of metallurgy and ceramics to a field that includes microelectronics, polymers, biomaterials, and nanotechnology. The stringent materials requirements, such as extreme purity, perfect crystallinity and defect-free materials for the microelectronics revolution in the 1960s, were the prime movers. Major developments in other technologically significant fields, such as polymers, optics, high-strength materials that can withstand hostile environments for space and atmospheric flight, prosthetics and dental materials, and superconductivity, have along with microelectronics changed materials science from a primarily metallurgical field to a broad discipline that includes ever-growing numbers of classes of materials and subdisciplines. This book is a textbook that ambitiously endeavors to present the fundamentals of the modern broad field of materials science, electronics materials science, and to do so as a first course in materials science aimed at graduate students who have not had a previous introductory course in materials science. The book's contents derive from course notes that I have used in teaching this first course for more than 20 years at UNC.

The initial challenge in teaching a one semester first course in this broad discipline of electronics materials science is the selection of topics that provide sufficient fundamentals to facilitate further advanced study, either formally with advanced courses or via self study during the course of performing advanced degree research. It is the main intent of this book to provide fundamental intellectual "tools" for electronic materials science that can be developed through further study and research. The book is specifically directed to materials scientists who will focus on electronics and optical materials science, although with an emphasis on fundamentals, the material selected has benefited polymer and biomaterials scientists as well, enabling a wide variety of materials science, chemistry, and physics students to pursue diverse fields and qualify for a variety of advanced courses. With such a broad intent virtually all of materials science would be relevant, since modern electronics materials include many diverse materials, morphologies, and structures. However, there was a self-limiting mechanism, namely it all had to fit into one semester. Consequently fundamentals are stressed and descriptive material is limited.

The next challenge for the instructor is to consider the level of students. In materials science curricula typically found in engineering schools, a first course in materials science is usually required before the end of the second undergraduate year, so as to provide the basis for more specialized and advanced junior and senior level undergraduate courses in the various areas of materials science. Thus most introductory (first course) materials science texts are written for first or second year engineering students, and therefore assume meager mathematical experience, and only elementary chemistry and physics. In

these texts principles are often introduced using formulas that are not derived, followed by descriptive material and examples to reinforce the ideas and provide practice with problem solving. There are numerous high-quality texts available at this level. Over the years I have used a number of them either as primary texts and/or as reference materials for the materials science courses that I teach at UNC. However, the level of the available introductory texts is too low for a first course in materials science offered to graduate students and to chemistry and physics undergraduates in their senior year. For the undergraduates at UNC where there is no materials science department, the first materials science course was part of an Applied Sciences Curriculum with Materials Science (electronic materials and polymers) as a track. For the chemistry and graduate students who will do graduate level research in materials science, there are only few advanced materials courses available at UNC. Thus the first materials science course offered to these students must not only be at a higher level, it must also more completely equip the students for advanced courses and independent study in their respective research interests. This text has been written from the notes that I have generated over the years of teaching this higher level, but introductory materials science course at UNC. The notes were used to supplement and raise the level of the available introductory texts.

Chapters 1 through 11 are covered in their entirety in a single semester course at UNC. The result is a fast paced course with a dearth of descriptive material. In this course I assume that the students have had at least two semesters of calculus, general chemistry, elementary but calculus-based physics, and the equivalence of two semesters of physical chemistry, which includes thermodynamics and quantum mechanics. Most of the students taking the course have had significantly more preparation than assumed. With these assumptions I am able to move more quickly through the material. Also there is not the usual initial treatment of chemical bonding, since it is assumed that students have already had at least two chemistry courses that cover atomic and molecular structure and chemical bonding and chemical reactions. Derivations of important formulas usually omitted in a first materials course are included where it is felt that the derivation is instructive, and not simply a mathematical exercise. Nonetheless, this author believes that it is necessary to have the student reach a comfort level with some more physical and mathematical areas so that they can read original papers without trepidation. The early introduction of reciprocal space is considered essential to understand diffraction as a structural tool, and also electron band theory (as **k** space) and much of solid state physics. Reciprocal space is the natural coordinate space. The mathematical nature of diffusion is introduced to present the "flavor" of the field. Electron energy bands are treated from the Kronig-Penney model, and not simply assumed to exist from semantic arguments, as is done for typical second-year texts. The area of defects, phase equilibria, and mechanical properties are treated similarly to introductory materials science texts with the addition of some important derivations so that a students can glean an appreciation of the origin of the formulas as well as the methodology used in various fields of materials science.

I am grateful to all my students, past and present, for all their help with this textbook. It was their questions and enduring curiosity that have often driven me to seek better, clearer explanations. Over the years my graduate students have made perceptive (and usually tactful) comments about my course pointing out both strong and weak areas. During the writing and editing of this book my Ph.D. graduate students (N. Suvorova, C. Lopez, R. Shrestha, and D. Yang) and post doctoral (Dr. Le Yan) have read and commented on the many drafts. I have tried to make the changes and corrections that they suggested, but I assume responsibility for the remaining unclear discussions and errors.

1

INTRODUCTION TO ELECTRONIC MATERIALS SCIENCE

1.1 INTRODUCTION

Materials science can be thought of as a combination of the sciences of chemistry and physics within a backdrop of engineering. Chemistry helps to define the synthetic pathways, and provides the chemical makeup of a material, as well as its molecular structure. Physics provides an understanding of the ordering (or lack thereof) of atoms and/ molecules and electronic structure, and physics also provides the basic principles that enable a description of materials properties. The combined information provided by physics and chemistry about a material leads to the determination and correlation of materials properties with the process used to prepare the material, and with the materials structure and morphology. The properties once determined and understood are exploited through judicious engineering. In a sense engineering brings focus to the properties that materials possess, and to the material itself if suitable applications are found. Evidence for the leadership of engineering is witnessed by the many national goals that pervade the national research funding agencies such as nanotechnology, biotechnology, and microelectronics. In each of these fields the advantages of certain materials properties are extolled. The goals in every case include the preparation of new materials with enhanced properties for particular engineering objectives.

Materials science as we know it today finds its origins in traditional metallurgy and metallurgical engineering departments. Consequently many university materials science curricula and textbooks in use in these curricula are heavily weighted toward traditional topics related to metallurgy. More modern areas are relegated toward special topics courses and textbooks covering selected areas. This text is aimed toward electronic materials science where the engineering objective is better materials for microelectronics and photonics.

While there has been growing interest and understanding in electronic materials for centuries, there was a major revolution in electronics that began in the late 1940s with the invention of the transistor by Bardeen, Brattain, and Shockley. This invention irreversibly changed the entire electronics arena. Essentially before this time all active electronic circuits components were made of closely spaced similar metal elements (electron-emitting filaments, grids, electrodes) contained within a glass vacuum envelope, so-called vacuum tubes. These devices could switch currents, provide amplification and rectification, and along with passive components enable the construction of radios, televisions, and even analog and digital computers. About the early electronic devices based on vacuum tubes, it is amusing to recall that these early electronic marvels were all larger than today's versions. None were larger than the early (1960s) analog and digital computers that used vacuum tubes, and that filled large rooms and even entire buildings, but had less computing power that the laptop with which this text is written. Then, after the invention of the transistor, is was more than 10 years before the ideas about the solid state devices could be truly felt with the implementation of reliable discrete transistors replacing vacuum tubes on the electronics market, and in all kinds of consumer devices. During this period of incubation from invention to widespread applications, there were somewhat dormant areas of science and engineering that became very active and made major advances that were spurred on by the potential markets for the new solid state devices. First it was realized that single crystals of semiconductor electronic materials had to be made in large quantities rather than in laboratory sizes and with crystalline perfection and chemical purity never before imagined in manufacturing. Then the notion of electronic band structure that derived from the earliest days of quantum mechanics had to be modernized and understood for the new solid state electronic materials. From the new results of electronic energy band structure, doping could be understood, and the role of crystallographic defects became central to electronics materials. Lattice diffusion of dopants into crystals developed greatly in this era. It was also realized that the new class of electronic devices would require the joining of different solid state materials such as metals with semiconductors with insulators in every permutation. Thus there was renewed interest in phase equilibria, not only to understand the important metallurigical transformations that govern steel and other alloys but, with emphasis on alloys between electronically dissimilar materials and with homogeneity ranges, so as to understand atomic vacancies and correlate crystal lattice vacancies with resulting electronic properties. Along with all these advances in understanding and practice of the solid state since the invention of the transistor, another invention came to the fore that also revolutionized the way we live. This invention is the integrated circuit (IC). The integrated circuit enables the configuring of solid state electronic materials in order to fabricate devices such as transistors and rectifiers on the surface of semiconductors, and to link them all together to make a complete electronic system or subsystem to be further linked. The IC has paved the way for all the modern electronic devices especially the digital devices that perform logic and memory. In addition to enabling the efficient manufacture of multiple solid state devices, the IC paved the way for another major revolution, namely nanotechnology or nanoscience. The very heart of the IC, as it is implemented with planar technology, enables the downward size scaling to present device dimensions in the nanoscale range. The areas of electronic materials science and microelectronics are clearly the forerunners of nanotechnology, and many of the techniques developed for ICs are fully integrated into modern nanotechnology. Thus the areas of electronics materials/microelectronics and nanotechnology are intimately related in that it is clear that microelectronics is the predecessor of nanotechnology, and that advances in nanotech-

nology will undoubtedly impact microelectronics. As microelectronics took hold of all the devices we use, the area of optical devices or photonics also developed using the solid state ideas about materials as well as the ability to integrate optical and electronic devices on a chip.

The study of electronic materials science must then include the factors that enable a material to be prepared and understood, and its properties determined and optimized for defined applications, in particular, electronics and/or photonics applications. These typical factors selected for study comprise the names of Chapters 2 through 11: Structure, Diffraction, Defects, Phase Equilibria, Diffusion, Mechanical Properties (two chapters), Electronic Structure, Electronic Properties, and Devices. Many of these topics and chapters have the same names one finds in traditional materials science texts, and that is no accident. It is clear that a foundation in traditional materials science is implicit in electronics materials science. The difference is in emphasis, since as a practical matter one text or one course cannot do it all. In the following paragraphs the reasons are discussed why these headings are chosen for a study of electronics materials science, and the emphasis is explained.

1.2   STRUCTURE AND DIFFRACTION

Materials science is often described as being comprised of structure-property relationships. In this context structure refers not only to the arrangement of the basic building blocks, or long-range ordering but also to the chemical structure or short-range ordering. This more complete notion of ordering is discussed early in Chapter 2 of this text with the appropriate nomenclature, and this theme is revisited many times throughout the book. Different structures can represent both different chemical bonding and different arrangements of atoms and/or molecules, and possibly even different states of aggregation (roughness, large grained, etc.). All these structural aspects can lead to different properties, including electronic and optical properties. It is important to use a consistent nomenclature to identify the unique structural features so that materials scientists communicate in a standard language. These topics are discussed in Chapter 2 on the structure of solids.

In Chapter 3 on diffraction we study the determination of crystal structure. The basic idea that underlies this important family of techniques, diffraction techniques, is the principle of superposition. It will be seen in the text that much of the fundamentals of materials science can be understood by referring to a few the basic tenets of chemistry and physics. Among the tenets that are continually revisited is the superposition principle that is used for diffraction, mechanical properties, and electronic structure (with the first review of this tenet in Chapter 3 and again more thoroughly in Chapter 9). For example, the nature of a wave function that is used to describe an electron can be understood by considering the wave function to be made up of many waves in a complex blend, namely the notion of modulation.

Later in Chapter 3 the concept of reciprocal space is introduced. The idea follows from the notion that it is important in science to operate in the coordinate space most appropriate to the system. It is found that for crystal structure obtained by diffraction, reciprocal distances correlate the structure with diffraction experiments.

From a study of structure and diffraction one may glean the erroneous idea that only, or at least mostly, crystalline materials are important in materials science and electronic materials science. This is far from the truth, but it is a natural tendency that follows from

paying close and early attention to only perfect crystals. In fact a large fraction of useful materials in all fields are not crystalline at all (e.g., the dielectrics used in microelectronic ICs), and another large fraction is partially crystalline (alloys used for contacts in microelectronics) or at least defective in their crystalline nature. However, the nonperfectly crystalline materials are more difficult to describe universally and simply. That is to say, each material must be described using a number of structural aspects where crystallinity may be one of the important aspects. However, as is usual in science, the ideal state is the easiest to describe thoroughly, and this is the reason why virtually all studies of materials science commence with a discussion of ideal or perfect crystals.

Also electronic structure that is discussed in Chapter 9 on electronic structure is important for determining many properties particularly electrical properties. It will be seen in Chapter 9 that the structure of the material will greatly influence the electronic structure and in turn the electronic and optical properties.

1.3   DEFECTS

To dispel the misleading attention to perfect crystals, in Chapter 4 on defects in solids we look at different kinds of defects. The definitions for several of the more common material defects are discussed. It has been found over and over that simple structural defects such as substitutional and interstitial defects can alter electrical properties and mass transport via diffusion by orders of magnitude, while at the same time hardly affect the melting point or the thermal conductivity for a material. Furthermore line defects are implicated as the main factor in the plastic deformation of crystalline materials. The notion of grain boundaries as the boundaries in between single crystal grains is also implicated in the mechanical properties of materials and in electronic properties of polycrystalline semiconductors. Thus both the structure and its level of perfection provide a backdrop from which the behavior and properties of a material are understood, particularly, electronic materials.

Also in Chapter 4 another fundamental tenet of materials science is introduced and used liberally in following chapters. This tenet is the Boltzmann distribution from which both equilibrium thermodynamics and activation energies, or energy barriers, for processes can be understood. This concept is introduced by considering a simple two allowed state problem, and assessing how two energetically distinct states separated by a difference in energy, ΔE, can be populated. The result is a familiar exponential term $e^{-\Delta E/kT}$ often referred to as the Boltzmann factor. However, in the field of chemical kinetics an Arrhenius factor with the same form as the Boltzmann factor is often discussed in relation to the velocity of chemical reactions, but the Arrhenius factor is often introduced without adequate discussion about its origin, or at best as an empirical result. The importance of this idea is such that it is introduced and discussed early in the text. Furthermore the laws of thermodynamics derive from the average or statistical nature of atoms or compounds that comprise a material. This statistical notion is crucial toward the understanding the average properties of a macroscopic piece of a material that contains a large number of atoms and/or molecules. Such thermodynamics properties include the phase of the material, the vapor pressure, and decomposition temperature. On the other hand, quantum mechanics may be required to understand the properties that depend on the specific interactions of atoms and/or molecules within a material such as the absorption or emission of light and the electronic and thermal conductivity.

1.4   DIFFUSION

In virtually all solid state reactions and transformation, matter moves; that is, atoms and/or molecules are transported to and from the reaction site. Often in the solid state that motion is by a random process, and such random processes are termed diffusive processes. Early in Chapter 5 on diffusion in solids the form for a variety of diffusion equations are compared, and it is observed that seemingly unrelated phenomena are governed by equations with the same form, namely there is a flux in response to a force. That flux (with units of amount/area · time) can be matter, heat, charge, energy, and so on. Even the famous Schroedinger equation of quantum mechanics (see Chapter 9) has the form of a diffusion equation. Although only mass diffusion is covered in Chapter 5, heat transport, for example, involves the solution of similar equations.

In the field of mass diffusion many treatments deal purely with the underlying physics that enable random matter transport, while other approaches deal exclusively with the mathematics of solving the differential diffusion equations. In Chapter 5 both areas are addressed. In addition another fundamental tenet in materials science is introduced, namely the random walk problem. While applied strictly to diffusion in this chapter, the random walk problem yields insight into how random processes can yield simple understandable results precisely because of the assumed randomness of the system. This is a powerful idea that helps hone the intuition of a materials scientist who must often deal with seemingly unsolvable problems involving randomness and complexity. In the field of electronic materials diffusion plays a central role that includes the transport of dopants, other point defects (vacancies and impurities, and electronic carrier diffusion in electronic and optical devices.

1.5   PHASE EQUILIBRIA

Traditional introductory materials science texts usually cover the topic of phase equilibria adequately for understanding electronic materials. The main reason is based on the fact that most introductory materials science texts emphasize metallurgical materials, namely metals and alloys, even though these texts have often been modernized with the addition of polymers and electronic materials. Metallurgy deals extensively with mixed composition alloys such as steel. An understanding of steel and other important alloys requires a detailed knowledge of the phase diagram for the system, in order to know under what conditions to expect certain alloy phases and the composition of the phases. However, oftentimes advanced physics and chemistry courses spend little time on this topic, and while some forms of phase equilibrium are covered in undergraduate chemistry courses, solid state phase diagrams are often barely mentioned. It is clear, however, that modern trends in materials science and electronic materials science include complex materials that can have several phases and wide homogeneity (stoichiometry) ranges. Included in the kinds of electronic and photonic materials in which phase equilibria are important are modern binary semiconductors that are used extensively for both electronic and optical devices, ceramic superconductors, alloy superconductors, magnetic alloys, high dielectric constant insulators, and polymer blends.

In Chapter 6 on phase equilibria we provide simple derivations of the Gibbs phase rule and the lever rule and outlines the procedure to estimate phase diagrams from known thermodynamic data. All materials scientists deal with the formation of phases from some primal state, and hence often the initial stage of phase formation, nucleation

becomes important in determining final product morphologies. For this reason nucleation is added in the chapter. An understanding of nucleation phenomena is also important to the understanding of the processes that are used to prepare the thin films used for most modern electronic and optical devices.

1.6   MECHANICAL PROPERTIES

In the first of the two chapters on mechanical properties the emphasis is the development of the basic ideas and the resulting relationships among the elastic constants. In Chapter 7 on the elasticity property of solids, these constants are used to describe the behavior of materials that deform elastically, which means that as forces are applied, the material deforms, but the material returns to its original state as the forces are removed. Most materials exhibit this behavior when small forces are applied for short periods of time. There is more interest when larger forces are applied that leave a material permanently deformed or even causes fracture of the material, since deformation and failure relate the usefulness of a material for fabricating products such as cars, bridges, and homes. However, as was the case for structure, first the simpler ideal case of elasticity is considered and then consideration is given to a more complicated behavior called plasticity. In Chapter 8 on the plasticity property of solids the underlying ideas are presented for permanent deformation or plasticity. The implication of dislocations for the plastic deformation of crystalline materials is discussed and creep is briefly discussed. In this chapter the deformation of noncrystalline materials such as polymers is discussed, and several models that are used to interpret the mechanical response of these kinds of materials are developed.

In microelectronics and photonics many of the devices are constructed by layering films of dissimilar materials. Therefore differences in thermal expansion as well as chemical incompatibilities at the interfaces can lead to performance and reliability issues for the devices. Furthermore many of the extreme structural features and extremely small sizes of features of the modern devices can exacerbate the mechanical issues that may exist for planar and larger devices. In addition the applications of forces on a crystal lattice can alter the atomic spacing and therefore affect the electronic nature, meaning the electronic energy band structure, of a material. A full analysis of these complicated structural and electronic issues is beyond the scope of this text, but a first-order treatment of the important relationships properties is essential so that advanced study and appreciation of the implications of mechanical properties can be accomplished.

Many modern microelectronics products such as computer chips are fabricated from thin films of dissimilar materials. Also, once the layered structures are formed, the products go through various temperature cycles as part of the further processing. These structures are prone to the development of stresses that can lead to device failure and to shorter useful lifetimes. Consequently the mechanical issues of thermal expansion, stresses, and defect formation that are crucial to further study of electronic material reliability are covered in these two chapters.

1.7   ELECTRONIC STRUCTURE

In Chapter 9 on electronic structure we consider another aspect of the structure of materials, namely the electronic structure. The basic ideas relating to electronic structure

include a consideration of the arrangement of atoms and molecules as was introduced in Chapters 2 and 3 plus the addition of a consideration of the interactions of the atoms or molecules in their various structural motifs. The interactions among atoms and molecules is handled using quantum mechanics. Quantum mechanics enables chemists to estimate, if not calculate, the structure of many important molecules using the Schrödinger equation. Similarly quantum mechanics enables the calculation of the allowed and disallowed energies for the electrons in an array of atoms or molecules in condensed phases, such as liquids or solids. The allowed energies are called energy bands, and the disallowed energies are called the forbidden energy gaps (FEG) or simply band gaps. An old (1931) but useful model for the calculation of electronic energy band structure for solids is presented, the Kronig-Penney (KP) model. Despite its simplicity the KP model contains many of the important physical ideas that are used in more modern models, but without difficult mathematics. Consequently the KP model is useful as a vehicle to understand the origin of allowed electronic energy bands and gaps, but the KP model does not enable quantitative estimations of energy bands. Nonetheless, many important conclusions can be made regarding the electronic structure of materials using the KP model. Associated with the energy band structure is an extensive nomenclature and representation language, and this language is introduced to describe electron energy band structure. In this chapter there is heavy reliance on the structural ideas and reciprocal space that were introduced in Chapters 2 and 3.

It is clear that fundamental to understanding electronic and optical properties of solids and the devices is the electronic energy band structure; thus Chapters 10 and 11 make heavy use of the ideas developed in this chapter. Furthermore modern ideas about nanotechnology that include quantum well structures, quantum dots, and other small intricate structures are understood in terms of the energy band structure and the comparisons that are made to larger devices.

1.8 ELECTRONIC PROPERTIES AND DEVICES

In Chapter 10 on electronic properties we make heavy use of the results from Chapter 9, in particular, the electronic energy band structure, and adds to this development the use of the statistics for electrons, namely Fermi statistics. An estimate is made about the number of electronic states for materials, the so-called density of states (DOS) is calculated. From the energy band structure, the density of states (DOS), and the probability for occupancy, the Fermi-Dirac distribution function, the electronic arrangement for solids is deduced. From this arrangement the electronic nature of the materials is revealed, and resulting properties are understood. The different kinds of electronic materials are also discussed: conductors, semiconductors, superconductors, and non-conductors. Electronic conduction is treated both classically and in terms of quantum mechanical ideas. For superconduction the popular BCS theory is introduced. Lastly in Chapter 10 the electronic nature of organic materials is introduced, and since many of the organic materials in use are amorphous, the electronic nature of amorphous materials is discussed. In the final chapter, Chapter 11 on junctions, devices, and the nanoscale, we reach a point where we can distill the ideas developed in Chapters 9 and 10 that are fundamental to designing and understanding electronic and optical devices. Virtually all modern electronic and optical devices use the junctions of materials. Thus in Chapter 11 we commence with junctions and the electronics implications of joining dissimilar materials. From junctions, passive devices that do not change flowing currents or applied

potentials can be constructed such as thermocouples and solid state refrigerators. Then, using various junctions, this chapter introduces electronic devices that are important in today's microelectronic technology such as diodes, solar cells, transistors, and the devices that comprise computer chips. The basis ideas about optical devices are introduced with examples. The last section deals with nanotechnology and the kinds of devices that will emerge from ongoing research in fabricating nanoscale structures from materials.

1.9 ELECTRONIC MATERIALS SCIENCE

Modern science and technology requires highly trained materials scientists who can function in diverse areas such as metallurgy, biology, ceramics, electronics, and optics, to name several fields. It is clear that there are many commonalities in the fields. For example, for all solid state materials, structure with all its implications is important. For biology, molecular structure is more important than is electronic energy band structure at this juncture in development. That is not to say that with the development of biomaterials and nanotechnology the future will bring bio-inspired electronic and optical devices. For many fields structural defects are important as are mechanical properties. For the fields of electronics and optics, electronic structure and properties are fundamental to understand the resulting devices. However, defects and mechanical interactions are also crucial. Thus topics in this text were chosen more as a matter of practicality, in that to adequately cover all areas of importance to electronic materials would result in an impractically large text. Careful choices had to be made in selecting the most germane material for electronic materials science.

2

STRUCTURE OF SOLIDS

2.1 INTRODUCTION

As the study of materials progresses in successive chapters, the importance of structure in dictating many of the materials properties will become clearer. Knowledge of structure along with chemical composition comprises the most fundamental properties known about materials, and both kinds of properties are required to complete the characterization of a material. A chemist as a molecular scientist typically focuses attention on the atomic composition and molecular structure of the chemical or molecule under study. Molecular structure refers to the arrangement of the atoms in a particular molecule. In addition to composition, a materials scientist must not only know structure at the molecular level but also at higher levels such as the arrangement of molecules, namely whether the molecules are ordered (or not) on scales larger than the molecular size. This is so because a given material with a specified composition can, and often does, exhibit widely differing properties that are related to the structure. A simple example of this is water, H_2O, as shown in Figure 2.1. Water, in the solid, liquid, and gaseous states possesses different structures, widely different properties, but the same chemical composition. Figure 2.1*a* displays the structure of a molecule of water while Figure 2.1*b* displays what state (solid, liquid, vapor) and structure of water exist at various pressures and temperatures. It is possible to have both a variation of the molecular structure (the relationship of the H's and O's) and a variation in the arrangement of the water molecules (the relationship of the H_2O units). Figure 2.1*b* illustrates several solid phases of water (X1, 1C, and 1H) that exist at high pressures and low temperatures. While this example seems simple enough, it is not. In this example the differences between different states of matter were compared, thereby exaggerating the structural dissimilarities. However, we could have chosen to discuss only solid H_2O, and its different structures as can be
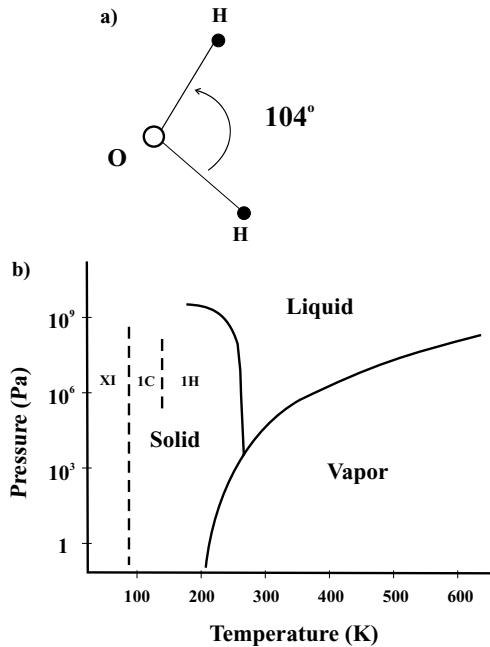
Figure 2.1 (*a*) A water molecule; (*b*) phase diagram of water showing three solid phases.

obtained by preparing ice under different conditions. One finds many properties that differ with structure, but there are also some properties that do not depend strongly on structure. An important objective of materials science is to understand structure-property relationships, namely why such correlations may or may not obtain. In this way a materials engineer can rationalize materials properties and design materials with optimum properties for a specific application.

The H_2O example indicates the fact that more than one level of ordering is important. On the atomic scale the chemical bonding between atoms is the same, or nearly the same, for many structurally different forms of H_2O. This chemical bonding level of structure is termed short-range order, or local structure, as opposed to the long-range ordering of the H_2O molecules in ice crystals. Short-range order is intimately related to chemical bonding and hence dictates stoichiometry. Long-range order refers to the arrangement of the chemical building blocks that may be molecular or atomic. This chapter endeavors to first describe order, then structure and the nomenclature used to indicate the kind of structure for solids. Many important materials do not possess order; hence we must also consider the kinds of disordered materials. The implications of structure are included in all the remaining chapters of this text mostly explicitly, but also implicitly.

2.2 ORDER

There are several, sometimes confusing terms related to order that require immediate attention. In the discussion above the notions of long- and short-range order were intro-
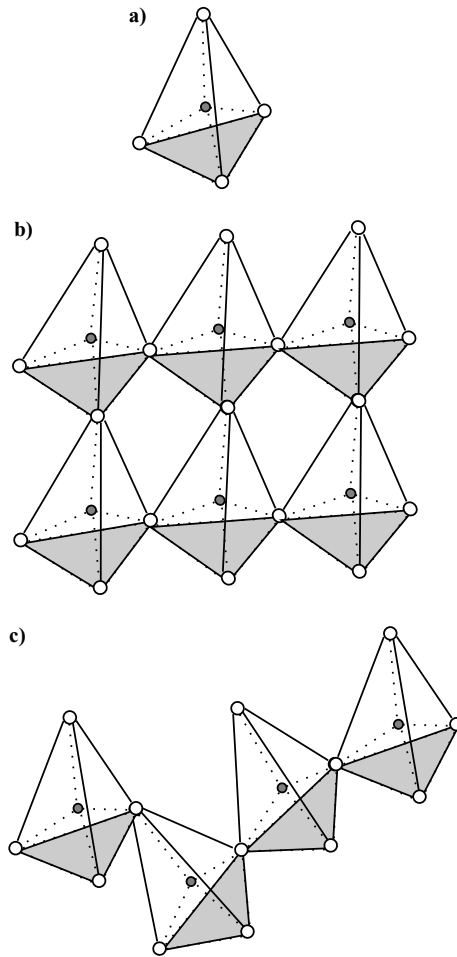
Figure 2.2 (*a*) An SiO₄ tetrahedron; (*b*) an ordered array SiO₄ tetrahedra; (*c*) a disordered array of SiO₄ tetrahedra.

duced. These concepts as well as a few other related concepts are further illustrated using Figure 2.2 for solid silicon dioxide, SiO_2. First we see that in Figure 2.2*a*, which represents a building block tetrahedron for SiO_2, the ratio of silicon atoms (shaded circles) to oxygen atoms (open circles) in the three-dimensional (3-D) representation is $\frac{1}{4}$ for a single isolated tetrahedral structural unit. The 3-D bonding in SiO_2 is tetrahedral, which means that surrounding each Si there are four O's located at the apices of a tetrahedron and with the tetrahedral O–Si–O angle of 109°54′. These SiO₄ building blocks are then assembled to create the 3-D solid SiO_2 material. This assembly is seen in Figures 2.2*b* and *c* where the O's at the apices of the tetrahedrons bridge to adjacent Si's yielding an overall stoichiometry of Si/2O's or SiO_2. The individual SiO₄ tetrahedra each composed of Si atoms tetrahedrally surrounded by O's have considerable local or short-range order, and they comprise the basic building blocks of SiO_2. However, the tetrahedra that are joined

through the bridging Os at the apices of the tetrahedra can exist in a range of angles; that is to say, there can be a wide distribution of Si–O–Si angles. If the distribution is very narrow, then the tetrahedra are all arranged in an orderly fashion and the material has long-range as well as short-range order as shown in Figure 2.2*b*. With the addition of long-range order the material is called "crystalline," and the possible crystal structures will be discussed later. If the distribution of Si–O–Si angles is wide, then the tetrahedra are arranged haphazardly although each tetrahedron is the same as all others. This material with short-range but not long-range order is called noncrystalline, or glassy. This is seen in Figure 2.2*c* where no apparent repeat is seen in the frame shown. Another possibility is that there are regions of crystalline order, but each region is unaligned with an adjacent region that is also crystalline. This kind of material is called polycrystalline or a polycrystalline aggregate. Each grain of the polycrystalline aggregate is itself a single crystal. Last, the material may have neither short- nor long-range order, and this material is called amorphous. Combinations of these types are also possible in that a material maybe part crystalline and part amorphous. While the distinctions made with these definitions for crystalline, noncrystalline, glassy, and amorphous are consistent, it is often found that the terms noncrystalline and amorphous refer to materials with no long-range order and the terms are used interchangeably, and glassy is used to describe amorphous or noncrystalline oxide glasses. In this text we will use the term amorphous to describe materials without long-range order.

The unifying theme for crystalline materials is the long-range ordering that can be thought of as a uniform translation of a basic building block. In this way one imagines that an entire macroscopic piece of a material is built simply by discrete translations of a basic building unit through three dimensions. We return to this point below.

2.3 THE LATTICE

The language used for describing crystal structures helps one understand the differences among the variety of possible structures. This language commences with the mathematical notion of a point lattice. Figure 2.3*a* shows a lattice to be an array of points in space so arranged that each point has identical surroundings. The smallest unit, or unit cell, can be obtained by constructing planes through points, and the lines resulting from the intersection of the planes at lattice points define the unit cell. Figure 2.3*a* shows a unit cell in darker outline and defined by the cell parameters **a**, **b**, **c** and angles (not shown) α, β, γ called lattice parameters. The angles are defined using Figure 2.3*b* where α is the angle between vectors **a** and **b**, β the angle between **a** and **c**, and γ is the angle between **b** and **c**. It should be noticed that the unit cell so defined embodies the symmetry of the entire lattice. The entire lattice can be generated by simply translating the unit cell by |**a**| in the **a** direction, by |**b**| in the **b** direction, and by |**c**| in the **c** direction. Thus translation becomes an important operation in understanding the long-range ordering represented by the lattice.

The question as to how many different kinds of unit cells are necessary to fill all space by translation and how to accomplish this for all possible symmetries is a solved mathematical question for which we herein accept the solution without proof. The lattices that accomplish this task are called Bravais lattices, and there are 14 such Bravais lattices, as shown in Figure 2.4.

These 14 Bravais lattices are organized into 7 crystal systems according to the basic symmetry that the lattice possesses: cubic, tetrahedral, hexagonal (or trigonal),

a)

b)
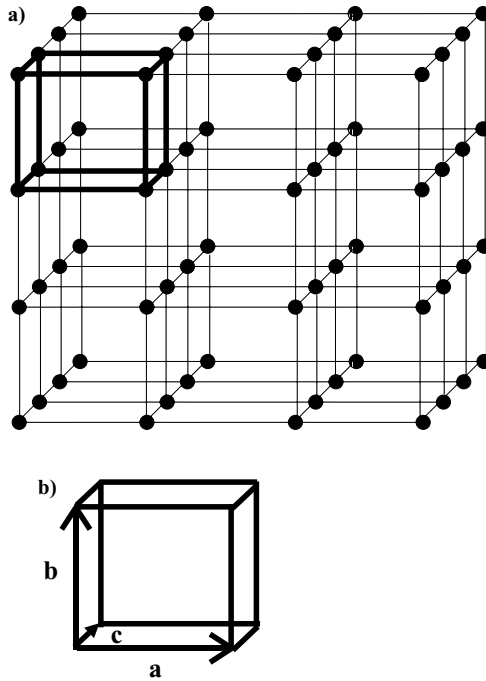
Figure 2.3 (*a*) Mathematical point lattice with a unit cell outlined; (*b*) unit cell with lattice vectors indicated.

Cubic

Tetrahdral

Orthorhombic

Hexagonal

Rhombohedral

Monoclinic

Triclinic
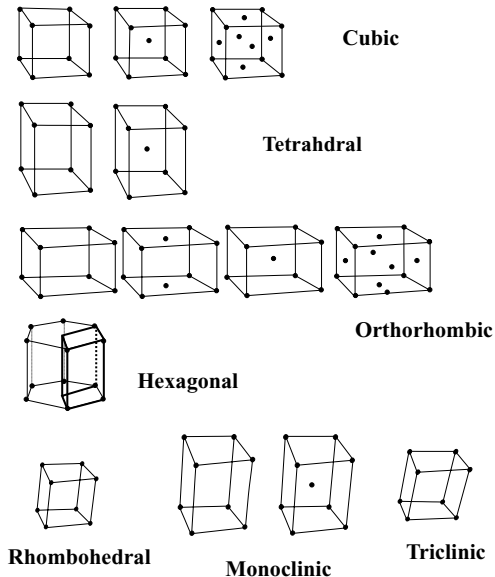
Figure 2.4 Seven crystal systems and 14 Bravais lattices.

Table 2.1   Seven crystal systems in terms of lattice parameters

Crystal System	Unit Cell Vectors	Unit Cell Angles
Cubic	$\mathbf{a} = \mathbf{b} = \mathbf{c}$	$\alpha = \beta = \gamma = 90°$
Tetragonal	$\mathbf{a} = \mathbf{b} \neq \mathbf{c}$	$\alpha = \beta = \gamma = 90°$
Orthorhombic	$\mathbf{a} \neq \mathbf{b} \neq \mathbf{c}$	$\alpha = \beta = \gamma = 90°$
Hexagonal (trigonal)	$\mathbf{a} = \mathbf{b} \neq \mathbf{c}$	$\alpha = \beta = 90°,\ \gamma = 120°$
Monoclinic	$\mathbf{a} \neq \mathbf{b} \neq \mathbf{c}$	$\alpha = \beta = 90° \neq \gamma$
Triclinic	$\mathbf{a} \neq \mathbf{b} \neq \mathbf{c}$	$\alpha \neq \beta \neq \gamma \neq 90°$
Rhombohedral	$\mathbf{a} = \mathbf{b} = \mathbf{c}$	$\alpha = \beta = \gamma \neq 90°$

Table 2.2   Seven crystal systems in terms of minimum distinguishing symmetry elements

Crystal System	Minimum Symmetry Elements
Cubic	4 Threefold rotation axes
Tetragonal	1 Fourfold rotation axis
Orthorhombic	3 Twofold rotation axes
Hexagonal (trigonal)	1 Sixfold rotation axis
Monoclinic	1 Twofold rotation axis
Triclinic	None
Rhombohedral	1 Threefold rotation axis

orthorhombic, rhombohedral, monoclinic, and triclinic. Some of these systems can have different lattices: simple or primitive (P), body centered (BC), and face centered (FC). Later we will discuss face-centered cubic structures (abbreviated FCC), and body-centered cubic structures (BCC), among others. Table 2.1 summarizes the description of the unit cells for the seven crystal systems.

In addition there are other basic symmetry operations that distinguish each of the seven crystal systems. Symmetry operations bring a lattice point into coincidence. Table 2.2 shows distinguishing or minimum symmetry elements that define and distinguish each of the seven crystal systems. Of course, the more highly symmetrical systems contain the symmetry elements of the lower symmetry ones.

Different types of lattices that make up the 14 Bravais lattices can be obtained by several fundamental translations of a primitive lattice position by the unit cell parameter(s) or fractions thereof. If we start on one corner of a primitive unit cell and assign this position the coordinates 0, 0, 0, then other major translations are

$$\text{Body centered: } 0,0,0 \rightarrow \frac{1}{2},\frac{1}{2},\frac{1}{2}$$

$$\text{Face centered: } 0,0,0 \rightarrow 0,\frac{1}{2},\frac{1}{2} \quad \frac{1}{2},0,\frac{1}{2} \quad \frac{1}{2},\frac{1}{2},0$$

$$\text{Base centered: } 0,0,0 \rightarrow \frac{1}{2},\frac{1}{2},0$$

The fractions correspond to fractions of the $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$ lattice parameters for the specific crystal system. Figure 2.5 shows the major translations for a cubic unit cell. Figure 2.5*a*
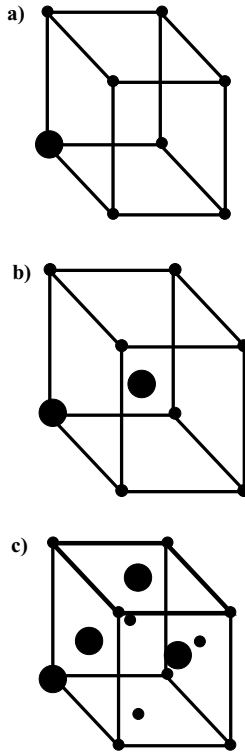
Figure 2.5 (*a*) Primitive, (*b*) body-centered, and (*c*) face-centered cubic unit cells showing unique lattice positions (large filled circles).

for a primitive cell shows that any apex is indistinguishable from the others. Thus there is only one unique position that is reproduced by a translation of **a**, and this position is denoted by the coordinates 0, 0, 0. Figure 2.5*b* for a BCC shows that there are the same corner positions summarized by 0, 0, 0, but there is also a unique center cell position labeled $\frac{1}{2}, \frac{1}{2}, \frac{1}{2}$ that cannot be generated starting from 0, 0, 0 and a translation of **a**. For the FCC Figure 2.5*c* shows that there are the corner positions summarized by 0, 0, 0 and there are six face positions. However, only three of the face positions need to be specified, since the others are obtained from a translation by the lattice parameter **a**. These major translations are useful for generating the unique positions in a lattice structure. The major translations will be used in the following chapter when we consider the scattering of radiation from the unique lattice positions of crystals and the different phases produced therefrom (i.e., diffraction). It should be noted that in many instances the magnitudes of lattice parameters are indicated with a zero subscript as a_0.

The number of lattice points for a unit cell, N, is calculated by counting the points that bound and are interior to the cell and then considering the sharing of points by adjacent cells. For example, the eight lattice points at the cell corners in the unit cell in Figure 2.3*a* are each shared by eight adjacent cells (N_c), the points on the face of a cell by two cells (N_f), and of course, the interior points (N_i) belong solely to the cell in question. Hence the following relationship summarizes this:

$$N = N_{\mathrm{i}} + \frac{N_{\mathrm{f}}}{2} + \frac{N_{\mathrm{c}}}{8} \qquad (2.1)$$

A primitive unit cell is defined as a cell that contains one lattice point. The density of lattice points may be calculated by considering the number of lattice points for the cell divided by the unit cell volume. If, as we show later, atoms or molecules are associated with lattice points, then the theoretical density of a material can be obtained. For example, for a cubic unit cell of dimension a_0, the volume of the cell is a_0^3. If the cell contains N atoms of the type with a molecular weight of M (g/mole), then the density is given as

$$\rho = \frac{N(\text{number}) \cdot M(\text{g/mole})}{N_0(\text{number per mole or Avogadro's number}) \cdot a_0^3} \qquad (2.2)$$

This theoretical density is sometimes called the X-ray density, and it can be compared with measured density. The difference is a measure of the perfection of a material. As will be covered in Chapter 3, when one uses X-ray diffraction to measure atomic positions, an average position or structure is measured. Local vacant positions and sparse impurities are ignored. Thus the X-ray density is based on the overall structure without imperfections.

2.4   CRYSTAL STRUCTURE

The mathematical lattices displayed in Figure 2.4 serve as the starting point for understanding crystal structures. They provide the smallest number of allowed symmetries in terms of easily imagined unit cells that are necessary to fill and thus define all space. But the information from these mathematical lattices is insufficient to describe real crystal structures. What is lacking is called the "basis." The basis is the atoms, or molecules, that comprise the real material and that are in some fixed relation to the lattice points. For example, the simplest case is the monatomic elemental solid where one atom resides exactly at each lattice point of one of the Bravais lattices. The Bravais lattice then becomes the crystal structure. The crystal structures for the elements are of this type. However, more complicated materials such as the SiO_2, as was discussed above, have the basic building blocks such as SiO_4 tetrahedra associated with the lattice points; more complex materials such as proteins and DNA have more elaborate building blocks associated with (not necessarily at) the lattice points of the Bravais lattices, thereby yielding a crystal structure. Thus the defining relationship for crystal structure is

$$\text{Crystal structure} = \text{Point lattice} + \text{Basis} \qquad (2.3)$$

Figure 2.6 illustrates the formula above. The structure shown at the right is made up of an array of the building blocks. The building blocks are the triangles of three open circles (e.g., to model a triangular molecule) and the array at the left is a Bravais lattice (primitive cubic). There is the same relationship between each building block and the array or lattice. For most elements the basis is unity, which means there is literally an atom at the lattice points.

If a structure is considered where the basis is known to be atoms or molecules, then the ideal density can be calculated. Depending on the lattice type, the number of atoms
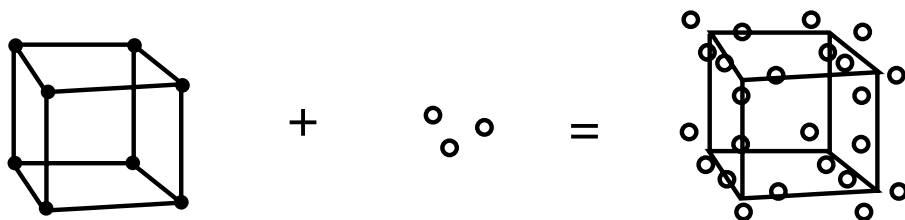
Figure 2.6   Lattice plus basis yields a crystal structure.

or molecules, N, in the structure can be calculated from formula (2.1) for N. With knowledge of the identity of the basis elements or molecules, the atomic or molecular weights, MW, are known. From N and MW for each species present in the unit cell, the mass of atoms in the cell is calculated. Now, if the unit cell parameters are known, then the volume of the cell is also calculated. Thus the mass divided by the volume for the unit cell yields the density for the structure. As was discussed above, the density calculated in this way is considered to be ideal, since it assumes that all the unit cells are as perfect as the one used for the calculation. Later we consider that defects can occur and alter the ideality. For example, suppose that one in one hundred lattice sites are vacant. This will alter the actual density by 1%. Similarly the presence of impurities, either as substitutes for atoms or in addition to, will alter the ideal density. Therefore the differences between ideal and real densities can signal and quantify the presence of some sort of lattice imperfection.

2.5   NOTATION

As one's understanding of structure deepens, it becomes increasingly important to be able to discuss specific planes and directions in the various crystalline materials. This importance derives from the fact that the chemical bonding that, being directional, is often different in different directions and on different planes in a crystal structure. Thus it is not surprising that many material properties are different in different bonding directions. Some of these properties, along with crystallographic differences, will be discussed in later chapters. In order to deal with directional differences, a methodology to name directions and planes in crystalline materials is in common use.

2.5.1   Naming Planes

The accepted system for naming planes is the Miller index notation. Naming planes is linked with finding the intersections of the planes with the basic lattice vectors that define the fundamental Bravais lattice for a structure. However, simply using intersections is sometimes cumbersome because interesting planes are often parallel to one or more of the unit cell lattice vectors. In this case the intersection is at infinity, and either the word or infinity symbol ∞ needs to be carried along in the nomenclature. In order to obviate this situation, the reciprocals of the intercepts are taken so that $1/\infty$ becomes 0. Fractions obtained after taking the reciprocal of the intercepts are cleared, and the resulting sets of usually three whole numbers (an exception to three is covered below) are placed
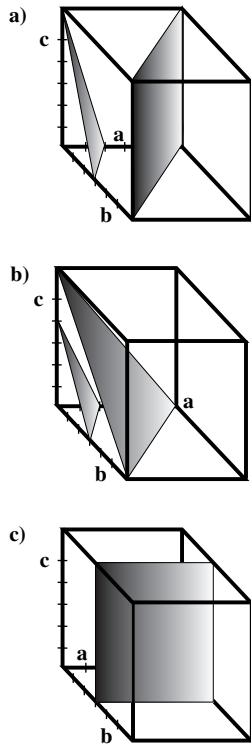
Figure 2.7 Examples of planes showing intercepts on lattice vectors.

in between rounded brackets ( ) indicative of specific planes. Figure 2.7 illustrates the Miller index system.

Figure 2.7a shows the **a**, **b**, **c** axes with the larger diagonal plane intercepts of **a** = 1, **b** = 1, and **c** = ∞ (i.e., the plane is parallel to the **c** axis). The reciprocals are 1/1, 1/1, and 1/∞, which yield the (110) plane. The small plane has intercepts of **a** = 1/3, **b** = 1/2, and **c** = 1. The corresponding reciprocals are 3, 2, 1, so the plane is the (321) plane. In Figure 2.7b the larger plane has intercepts of 1, 1, 1, so the plane is (111). The smaller plane has intercepts of 1/3, 1/2, 2/3. The reciprocals are 3, 2, 3/2 and, upon clearing fractions, becomes the (643) plane. Figure 2.7c shows the shadowed plane with intercepts of ∞, 1/2, ∞, which yields the (020) plane. We can imagine the planes perpendicular to and bisecting the shaded (020) plane. These planes would be the either the (200) plane or the (002) plane. These three planes comprise a family of planes denoted by {200}. Similarly in Figure 2.7c the planes that bound the figure are {100}, namely the family of (100) planes.

The hexagonal system often uses an additional index, meaning four indexes rather than three, as (h, k, i, l). The new index i is symmetrically related to the first two as $-i = h + k$. Because this fourth index is not unique, it is sometimes omitted or replaced by a period as ($h \, k \, . \, l$) to indicate hexagonal symmetry.

Because the Miller indexes are obtained from the reciprocals of the intercepts, the planes with the smallest intercepts (relative to a lattice parameter) have the largest Miller
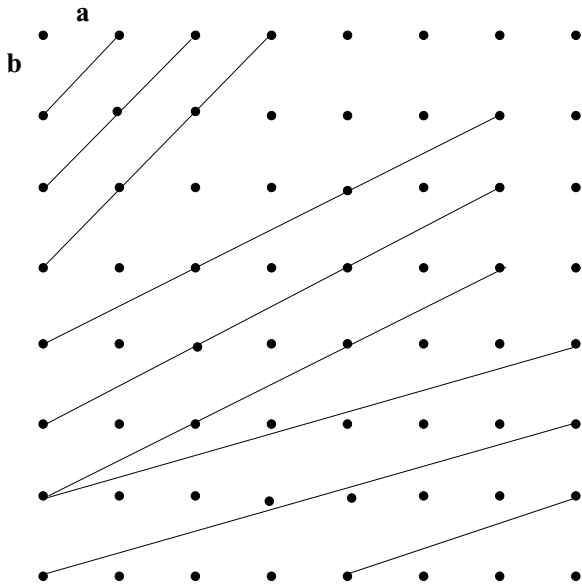
Figure 2.8 Two-dimensional cubic lattice showing different low and high Miller index planes.

indexes. Low index planes are the most common ones found in nature, and hence the intercepts being fractional intercepts correspond with the lattice parameters. Figure 2.8 shows a 2-D projection of the low and high index planes. There are three sets of planes shown: (11), (12), and (17). Notice that the low index planes also contain the greatest number of lattice points per unit length in 2-D (area in 3-D). These planes with the highest atom/molecule concentration also possess the highest bond density and thus electron density. Therefore all those properties that correlate with atom, bond, and/or electron density are determined by the low index planes of the material. It is easy to see why when describing the properties of a crystal it is important to also specify the direction in which the property was measured and the appropriate plane involved.

2.5.2 Lattice Directions

In order to name a direction, one must first construct a line parallel to the direction to be named, but that intersects the origin of the lattice vectors. Then at any point on the constructed line a perpendicular is dropped to each lattice vector. The intercepts to the lattice vectors cleared of fractions are the direction indexes. An example is shown in Figure 2.9a. The line l is drawn from the origin parallel to the line whose direction is to be determined (?). The intercepts on the **a**, **b**, **c** axes are noted to be **a** = 1, **b** = 1, **c** = 1/2. The intercepts are cleared of fractions yielding the direction [221]. Any intercepts consistent with being parallel to the direction in question will work. Figure 2.9b shows a cubic unit cell with low index planes and directions. Notice that the directions in square brackets are perpendicular to the planes in rounded brackets with the same indices.
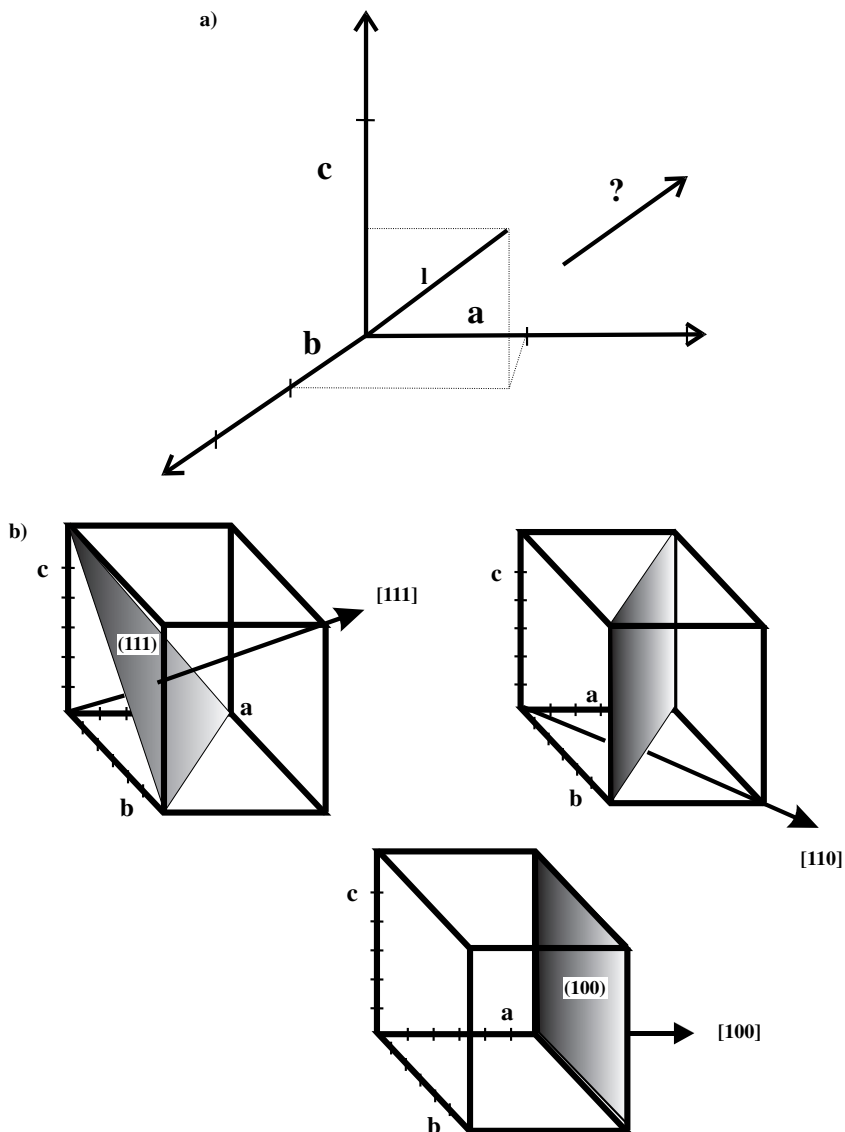
Figure 2.9 Procedure to index a direction: (*a*) The direction desired to be indexed (?) with a line (1) drawn parallel to the desired line and with intercepts; (*b*) several directions with associated planes.

Direction indexes are enclosed in square brackets [ ]. A family of directions, such as [111], [11$\bar{1}$], [1$\bar{1}$1], and [$\bar{1}$11], where $\bar{1}$ indicates a negative value for the index, can be indicated using angular brackets as <111>. It is both interesting and useful to realize that for orthogonal Bravais lattices ($\alpha = \beta = \gamma = 90°$) the [100], [110], and [111] are perpendicular to (100), (110), and (111), respectively. Table 2.3 summarizes the kinds of brackets that are conventionally used to indicate planes, directions and families of each.

Table 2.3 Nomenclature for planes and directions

Plane	( )
Family of planes	{ }
Direction	[ ]
Family of directions	⟨ ⟩
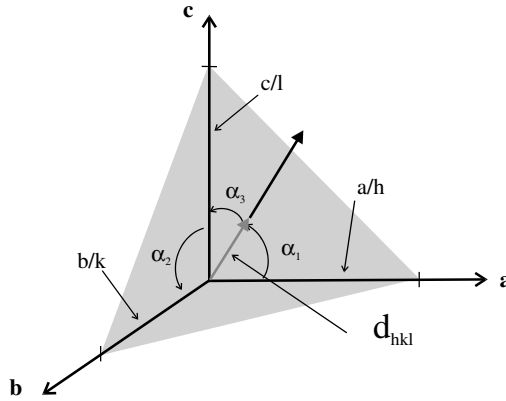
Figure 2.10 Plane (*shaded*) with angles and fractional intercepts.

2.6 LATTICE GEOMETRY

2.6.1 Planar Spacing Formulas

From Figure 2.10 it is seen that the perpendicular distance from the origin, (000) of the coordinate system to the plane shown, is labeled $\mathbf{d}_{hkl}$. $\mathbf{d}$ is the perpendicular to the plane. With the planar Miller indexes of (*hkl*), the fractional intercepts that (*hkl*) makes with the coordinate system are $\mathbf{a}/h$, $\mathbf{b}/k$, $\mathbf{c}/l$ for the axes $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$. Remember that $\mathbf{a}$ is the full lattice vector length. Assume unit length $\mathbf{a} = 1$, then $1/h$ is the fractional intercept. From this figure we can define the following angles α:

$$\alpha_1 \text{ between } \mathbf{d} \text{ and } \mathbf{a}, \quad \alpha_2 \text{ between } \mathbf{d} \text{ and } \mathbf{b}, \quad \alpha_3 \text{ between } \mathbf{d} \text{ and } \mathbf{c}$$

The direction cosines are obtained:

$$\cos\alpha_1 = \frac{\mathbf{d}}{\mathbf{a}/h}, \quad \cos\alpha_2 = \frac{\mathbf{d}}{\mathbf{b}/k}, \quad \cos\alpha_3 = \frac{\mathbf{d}}{\mathbf{c}/l} \tag{2.4}$$

For a cubic system $|\mathbf{a}| = |\mathbf{b}| = |\mathbf{c}| = a_0$ and $d = |\mathbf{d}|$,

$$\cos^2\alpha_1 + \cos^2\alpha_2 + \cos^2\alpha_3 = 1 = \frac{d^2\left(h^2 + k^2 + l^2\right)}{a_0^2} \tag{2.5}$$

Table 2.4   Planar spacing formulas for the seven crystal systems

Crystal System	Plane Spacing Formulas
Cubic	$\dfrac{1}{d^2} = \dfrac{h^2 + k^2 + l^2}{a_0^2}$
Tetragonal	$\dfrac{1}{d^2} = \dfrac{h^2 + k^2}{a_0^2} + \dfrac{l^2}{c_0^2}$
Orthorhombic	$\dfrac{1}{d^2} = \dfrac{h^2}{a_0^2} + \dfrac{k^2}{b_0^2} + \dfrac{l^2}{c_0^2}$
Hexagonal	$\dfrac{1}{d^2} = \dfrac{4}{3}\left(\dfrac{h^2 + hk + k^2}{a_0^2}\right) + \dfrac{l^2}{c_0^2}$
Rhombohedral	$\dfrac{1}{d^2} = \dfrac{(h^2 + k^2 + l^2)\sin^2\alpha + 2(hk + kl + hl)(\cos^2\alpha - \cos\alpha)}{a_0^2(1 - 3\cos^2\alpha + 2\cos^3\alpha)}$
Monoclinic	$\dfrac{1}{d^2} = \dfrac{1}{\sin^2\alpha}\left(\dfrac{h^2}{a_0^2} + \dfrac{k^2\sin^2\alpha}{b_0^2} + \dfrac{l^2}{c_0^2} - \dfrac{2hl\cos\alpha}{a_0 c_0}\right)$
Triclinic	$\dfrac{1}{d^2} = \dfrac{1}{V^2}\left(Ah^2 + Bk^2 + Cl^2 + 2Dhk + 2Ekl + 2Fhl\right)$
	$V = a_0 b_0 c_0 (1 - \cos^2\alpha_1 - \cos^2\alpha_2 - \cos^2\alpha_3 + 2\cos\alpha_1\cos\alpha_2\cos\alpha_3)^{1/2}$
	$A = b_0^2 c_0^2 \sin^2\alpha_1; \; B = a_0^2 c_0^2 \sin^2\alpha_2; \; C = a_0^2 b_0^2 \sin^2\alpha_3$
	$D = a_0 b_0 c_0^2(\cos\alpha_1\cos\alpha_2 - \cos\alpha_3); \; E = a_0^2 b_0 c_0(\cos\alpha_2\cos\alpha_3 - \cos\alpha_1)$
	$F = a_0 b_0^2 c_0(\cos\alpha_3\cos\alpha_1 - \cos\alpha_2)$

or in the more common form:

$$\frac{1}{d^2} = \frac{h^2 + k^2 + l^2}{a_0^2} \tag{2.6}$$

From Table 2.4 we can see that the algebraic method used to calculate the interplanar spacing formulas rapidly becomes tedious for the crystal systems with lower symmetry. In Chapter 3, when we cover reciprocal space that comprises the natural coordinates for the link between structure and diffraction, we will see that this procedure is simpler in reciprocal space.

2.6.2   Close Packing

In order to obtain a first-order notion of close packing, consider the primitive, body-centered and face-centered cubic cells in Figure 2.4. At each lattice position imagine a sphere (atom or molecule as the basis) of radius R. For simplicity, consider all the spheres to be equivalent. Now imagine each of the cell dimensions shrinking uniformly (but not the radius of the spheres) until the spheres just touch. It is clear that the shortest connecting dimensions are the most important, and along these directions the spheres will touch first. For the primitive cubic structure shown in Figure 2.11a, each of the eight spheres (two shown) touch six nearest neighbors, and no sphere is untouched. A tetra-

a)

$$a_o = 2R$$

b)

$$a_o = \frac{4R}{\sqrt{3}}$$

c)

$$a_o = \frac{4R}{\sqrt{2}}$$
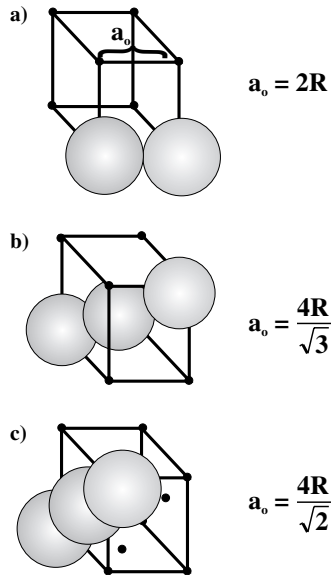
Figure 2.11   Close packing directions for (*a*) PC, (*b*) BCC, and (*c*) FCC cubic unit cells where the closely packed direction is indicated by the touching of atoms (*shaded*). The relationship between the lattice parameter a_0 and the atomic radius R is also given.

hedral shaped hole is formed at the center of this closely packed structure that has an edge length or lattice parameter of $a_0 = 2R$, where R is the radius of the spheres. For the BCC, however, the sphere at the center of the cell contacts the eight corner spheres, yielding a cell body diagonal length of $4R$ as is illustrated in Figure 2.11*b*. This closely packed structure yields a lattice parameter of $a_0 = 4R/\sqrt{3}$, and the corner spheres do not touch. At each of the six cubic faces an octahedral hole exists. The octahedron shape for the interstitial holes in the lattice is completed with the inclusion of adjacent BCC cells. For the FCC shown in Figure 2.11*c*, close packing is achieved when the face-centered spheres touch the corner spheres. Again, the corner spheres do no contact each other. The face diagonal is $4R$ and the cell dimensions are $a_0 = 4R/\sqrt{2}$.

Interestingly there are the octahedral holes at the center of each cell, and in addition at the cell edges there are tetrahedral interstitial sites formed. Tetrahedral sites have four spheres and octahedral sites have six. The existence and size of these interstices are important because transport of species can take place through the interstices and foreign species can occupy interstices. These ideas will be developed further in following chapters.

Another idea related to the concept of packing is the possibility of packing spheres in one layer upon another. As shown in Figure 2.12, the first and bottom layer of spheres (dashed) that are touching is called the A layer. The next closed packed layer is imagined to form by simply allowing the spheres for the second or B layer to fall into the troughs made by three A layer close packed spheres. Now to form the third layer, there are two possibilities. If the possibility that the third layer forms in direct correspondence to the A layer, then this third layer is also named an A layer (another A layer). The close packing of layers follows the order A B A B A. . . . This form has hexagonal symmetry and is
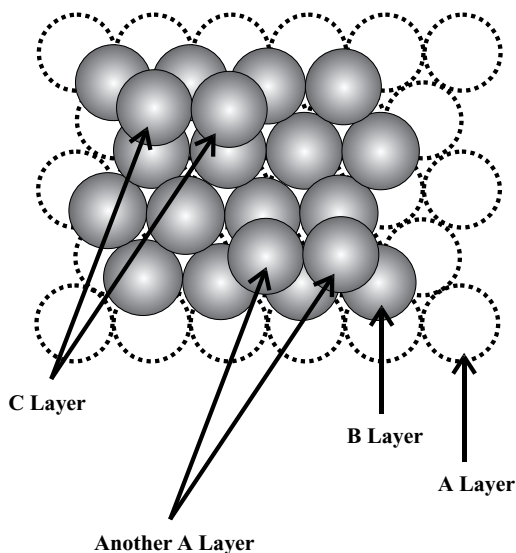
Figure 2.12   Close packing in layers where atoms are assumed to be spherical. The bottom A layer (*dashed*) is covered with B layer atoms in troughs in the A layer. The C layer is likewise formed but in two different ways.

consequently called hexagonal close packed (HCP). Alternatively, if the third layer forms in the other position, which registers neither with the A or B layers, it forms a C layer with the order A B C A B C. . . . This packing is also close packing and possesses FCC symmetry, so it is termed accordingly. Atoms of both HCP and FCC close packing are shown in Figure 2.12.

2.7   THE WIGNER-SEITZ CELL

Up until now we have chosen the unit cell boundaries somewhat intuitively by extracting a portion of the larger lattice. It is reasonable to expect that by this method the carefully chosen pieces will reproduce the entire lattice by translation and therefore fulfill the unit cell definition. There are other methods to select the unit cell that keep the requirements the same, namely that the unit cell must contain the symmetry of the lattice and fill all space by translation. It is particularly useful in some applications to choose a cell that is primitive, a cell that contains a single lattice point. One way to do this is with a square 2-D lattice as depicted in Figure 2.13. As the figure shows, one starts at any lattice point in the 2-D array and draws lattice vectors emanating from the starting point to first nearest neighbors (solid arrows). The bisectors (dashed lines) of these vectors are constructed and extended. The area included within the bisectors forms a new unit primitive cell (shaded) and is called a Wigner-Seitz cell after the scientists who made use of this kind of cell. In 3-D the lattice vector bisectors are planes that enclose a volume surrounding the chosen initial lattice point. We will revisit this construction in Chapter 3 after reciprocal space is introduced, and we will produce a similar unit cell in reciprocal
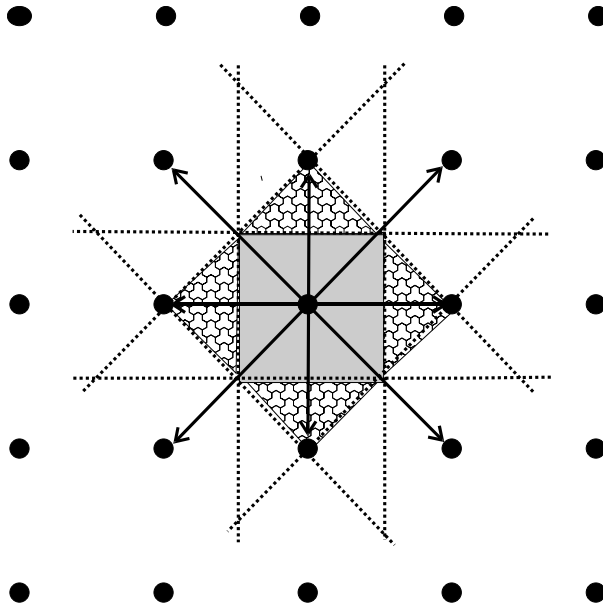
Figure 2.13   Schematic of the Wigner-Seitz cell formation. A 2-D lattice is shown with vectors drawn to nearest neighbors and next nearest neighbors. These vectors are bisected to form two (*shaded*) primitive unit cells.

space that is called a Brillouin zone. The Brillouin zone has a special significance in electron band theory, as will become apparent in Chapter 9 and subsequent chapters.

2.8   CRYSTAL STRUCTURES

2.8.1   Structures for Elements

As was mentioned above, for a crystal structure to form, both lattice and basis are required. We first consider a structure composed of atoms of the same element at all lattice sites. All the elements fit this example except for uranium, where the stable room temperature crystal structure is base-centered orthorhombic but with two atoms not at, but near, each lattice position. Many elements are cubic, either BCC or FCC. Table 2.5 gives some examples. For the most part these kinds of structures are easily visualized by simply imagining a Bravais point lattice and placing atoms at the lattice points or in some few cases near these points. In Figure 2.4 the top row gives examples of PC, BCC, and FCC metals where the basis is the shown lattice points. Figure 2.14*a* shows the diamond cubic (DC) lattice of carbon (C). This is similar in form to the FCC (open circles) but includes four extra C atoms at $\frac{1}{4}\frac{1}{4}\frac{1}{4}$, $\frac{3}{4}\frac{3}{4}\frac{3}{4}$, $\frac{3}{4}\frac{3}{4}\frac{1}{4}$, and $\frac{1}{4}\frac{1}{4}\frac{3}{4}$ that are shown as black circles for contrast. Several important semiconductors, Si and Ge, have this diamond cubic structure.

Table 2.5   Selected elements and compounds with crystal structures and lattice parameters

Substance	Structure	Lattice Parameters a, b, c, α, β, γ (nm, °)
Al	FCC	0.5311
As	Rhomb	0.4132, 57.1°
Ag	FCC	0.4086
Au	FCC	0.4079
B	Tet	0.880, 0.505
C	DC	0.3567
	Hex (graphite)	0.2461, 0.6708
Cu	FCC	0.3615
Ga	Ortho	0.4523, 0.4524, 0.7661
Ge	DC	0.5658
In	Tet	0.4598, 0.4947
Fe	α, BCC	0.2867
	β, FCC	0.3647
	γ, BCC	0.2932
Pb	FCC	0.4950
Ni	FCC	0.3524
P	Ortho	0.3314, 0.4377, 0.1048
Pt	FCC	0.3924
Si	DC	0.5431
Sn	α, DC	0.6489
	β, Tet	0.5832
CsCl	PC	0.356
GaAs	DC/zincblende	0.5653
NaCl	FCC	0.5639
SiO_2	α, Hex	0.490, 0.539

2.8.2   Structures for Compounds

Chemical compounds are characterized as being both chemically and physically distinct from their atomic constituents. This characteristic is attributed to the fact that the way chemical bonding alters the electron density distribution is dependent on both the electrons and the nuclei involved. Thus each compound is a definite and unique combination of atoms that is also unique in nature. The specific arrangement of atoms must be unique since the atomic and molecular forces are related to the particular arrangement. In fact it is probable that the crystal structures for all compounds are distinct. While this intuitive reasoning suggests considerable truth, the way the crystal structure fills space is described well within the framework already developed. When dealing with chemical compounds there are some rules to keep in mind. First, all the atoms in a given compound crystal structure must have the same symmetry. Second, when describing a crystal structure for a compound, all the repeated translations must begin and end on the same atom. While compounds present a great diversity to crystallographers, the ideas already developed plus the two rules given above greatly simplify the area. We consider a few common cases from myriad possibilities.
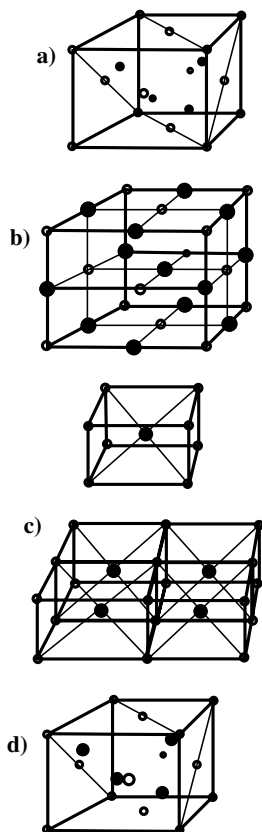
Figure 2.14 Several crystal structures: (*a*) Diamond cubic; (*b*) NaCl structure; (*c*) CsCl structure; (*d*) ZnS structure.

Inorganic solid compounds typically form simple structures. For example, solid NaCl possesses FCC structure (note that elemental Na at room temperature is BCC and Cl is a gas). What does this mean? In order to visualize this structure, it is well first to imagine the appropriate point lattice, namely the FCC Bravais lattice. Then apply the rules stated above. Figure 2.14*b* shows the NaCl structure to be two interpenetrating FCC point lattices. The basis for one is Na at the lattice points, while Cl is the basis for the other. It is sometimes difficult to determine the repeating cell because there are too few atoms in view. This can be illustrated with the CsCl structure shown in Figure 2.14*c*. In Figure 2.14*c* the top panel shows that CsCl has a BCC structure with Cl at the center, but this is not correct from both a structural and chemical point of view. Since such a repeating unit cannot faithfully reproduce the correct stoichiometry of CsCl at a ratio of 1/1, a more extended view of the CsCl geometry is shown in the lower panel. The structure here is primitive cubic with two interpenetrating primitive cubic lattices, the first with Cs as the basis and the other with Cl. Not only are the NaCl and CsCl structures interesting in themselves, but there are numerous other chemical compounds that have the same structures. KCl, CaSe, and PbTe have the NaCl structure, while CsBr and NiAl as well

as many important ordered alloys (discussed below) display the CsCl structure. In Figure 2.14*d* once again is shown a FCC structure that appears similar to the DC discussed above and shown in Figure 2.14*a* for C. However, Figure 2.14*d* gives the so-called ZnS structure in which two FCC lattices interpenetrate, one for Zn and the other for S. It should now be clear that compound structures can be difficult to visualize from a small sampling size.

Table 2.5 lists some commonly used materials in elemental and compound form with structures and lattice parameters.

2.8.3 Solid Solutions

In addition to pure elements that have a definite crystal structure and chemical compounds that are distinguished by having both a definite stoichiometry and a definite crystal structure, there are other kinds of crystalline materials for which knowledge of structure allows one to understand the materials properties and reactions. Solid solutions are prominent in this category of materials, and the solid solution is defined as a phase with variable stoichiometry (see Chapter 6 on phase equilibria). Structurally the solid solution retains the crystal structure of the host or solvent, and the other minor constituents or solutes add either substitutionally or interstitially to the host lattice. In the former case as depicted in Figure 2.15*a* some of the host species (open circles) are randomly replaced by solute species (filled circles) at the lattice sites. For an interstitial solid solution the solute species are positioned in the interstices of the host lattice rather than at the lattice sites. The arrangement of solute species is typically random on the host lattice. However, under certain conditions for substitutional solid solutions and for specific materials, the solute atoms take regular positions in the host lattice. A second lattice arises as shown in Figure 2.15*b* (the ordering of the solute-filled circles) due to this additional ordering. This situation is termed a super lattice because a new ordering is added to the already existing order, which in this case exists in the host lattice. For substitutional solutions, the host and guest atoms need to be similar in size and charge for any substantive solubility (CuNi solutions as mentioned below display 100% solubility due to the similarities of the atoms). For interstitial solutions, the guest typically needs to be sufficiently small to fit in the interstices. A notable example of an interstitial solid solution is steel, which is essentially an interstitial solution of C atoms in the interstices in the Fe lattice. Later, in Chapter 11, we will see other kinds of superlattices that are composed of repetitive thin film layers. This ordering yields unusual quantum mechanical behavior.

Solid solutions are also referred to as alloys. The stoichiometry for solid solutions or alloys can vary sometimes to all proportions. For example, Cu and Ni can mix in all proportions, but the question arises as to what structure results, Cu or Ni? In fact both Cu and Ni have the same structure, FCC. In addition both elements have nearly the same lattice constant of 0.36 nm for Cu and 0.35 nm for Ni and similar atomic radii (0.25 nm) and electronic structure. Thus intuition suggests that these elements are continuously miscible in all proportions. In the case of interstitial alloys, typically the solute is smaller than the host so that the host interstices can house solute without considerable distortion to the host lattice. One technologically significant example is steel composed of C in Fe, as mentioned above. C with a radius of about 0.15 nm somehow fits into the octahedral holes in BCC Fe (α-Fe) that has a diameter of about 0.072 nm in the Fe lattice. Due to the size of C relative to the size of the octahedral hole, the solubility of C is small, and this results in considerable local distortion of the Fe lattice. It should be
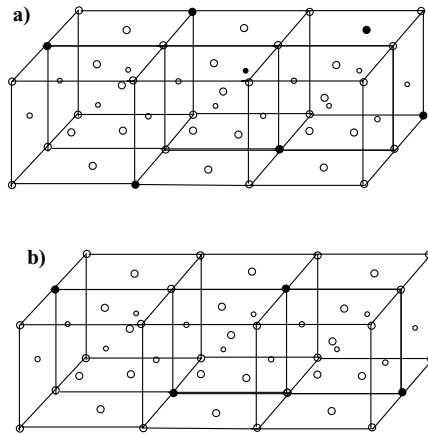
Figure 2.15 Alloy formation with guest atoms (*filled circles*) taking position on the host structure (*open circles*): (*a*) Random distribution of guest atoms; (*b*) ordered distribution forming a superlattice.

remembered that however convenient intuition here is at predicting the results, all solute atoms add energy to the system by distorting the original lattice, even where atomic sizes are close. The solubility limit in such an addition process can, in principle, be obtained by determining how much additional strain the host lattice can withstand before breaking down or rejecting further foreign species.

RELATED READING

C. R. Barrett, W. D. Nix, and A. S. Tetelman. 1973. *The Principles of Engineering Materials*. Prentice Hall, Englewood Cliffs, NJ. A readable elementary text for a first course in materials science.

E. D. Cullity. 1956. *Elements of X-ray Diffraction*. Addison Wesley, Reading, MA. All editions of this book contain voluminous structure information in readable text form, and in appendixes the X-ray diffraction techniques are discussed at length.

P. A. Thornton and V. J. Colangelo. 1985. *Fundamental of Engineering Materials*. Prentice Hall, Englewood Cliffs, NJ. A readable elementary text for a first course in materials science.

EXERCISES

1. (a) In the cubic structure calculate the angles between the following planes:
   (100) and (100) there are 2; (100) and (110) there are 2; (110) and (110) there are 3
   (111) and (100) there are 1; (111) and (110) there are 2; (111) and (111) there are 3.
   Hint: The angle between planes is the angle between plane normals.
   (b) Illustrate the angles between the (100) and (110) with sketches.

2. (a) For cubic symmetry draw the (111) bounded by [110].
   (b) Identify each of the three bounding directions.
   (c) Show mathematically that [111] is perpendicular to each direction in (b).

3. Show using sketches or math that the [111] is perpendicular to (111) in a cubic system but not otherwise.

4. Sketch and shade the following planes: (111), (110), (100), (220), (330), (113), and (115). Make a statement on the relationship between the Miller indexes and the distance from the origin for the planes sketched.

5. For FCC and BCC metals calculate the lattice parameter in terms of the atomic radii of the atoms. Sketch the structures and indicate the close packed directions.

6. For Cu with a cubic structure, $MW = 63.54\,g/mol$, $a_0 = 0.3168\,nm$ and density $= 8.92\,g/cm^3$
   (a) Determine whether Cu is PC, BCC, or FCC.
   (b) Calculate the atomic radius for Cu.

7. For BCC Fe with atomic radius is $0.1238\,nm$:
   (a) Calculate a_0. (b) Calculate the ρ for a perfect crystal of Fe. (c) Look up a value for π_{Fe} and compare with the calculated value and discuss reasons for the difference.

8. Calculate the % volume change for a substance assuming constant atomic radius:
   (a) If a material were to change from PC to BCC upon heating (b) and then from BCC to FCC if heated further.

9. Calculate the atomic area density (#/area in nm^2) on the (111), (110), and (100) planes for any FCC and BCC element that you choose.

10. Make careful sketches of an octahedral (six nearest neighbors) and a tetrahedral (four nearest neighbors) site in the FCC.

11. Show that low Miller index planes have a higher lattice point density that high index planes.

12. Determine the specific Bravais lattice for a cubic metal (atomic wt $= 192\,g/mol$, $a_0 = 0.3839\,nm$) with a density of $22.5\,g/cm^3$.

13. Discuss briefly the difference between short- and long-range order in any material you choose (SiO_2, etc.). Use a labeled sketch.

14. For a BCC with $a_0 = 0.1\,nm$, calculate the atomic radius assuming close packing.

15. Given the density, Δ, for an FCC element with a known atomic weight, calculate the Bragg angle for (111).

16. Explain (briefly) the structural difference(s) between a compound and an alloy.

17. What are the Miller indexes for the intercepts: $a = 2$, $b = 1$, and $c = 0.5$?

3

DIFFRACTION

3.1 INTRODUCTION

The study of diffraction in materials science is fundamental and important for several reasons. First and foremost is the determination of the structure of solids. The lattice type and the precise locations of atoms or molecules relative to the lattice, the basis, are accessed using diffraction techniques. A large number of studies of structure and properties have attested to a clear link between structure and properties. Thus the establishment of sturcture-property relationships has become the defining feature of the field of materials science.

Of special importance to electronic materials scientists is the study of electronic transport, which is essentially the manner in which electrons interact with the solid (i.e., the interaction of electrons with the atoms/molecules). The interaction of the electronic wave functions of the atoms/molecules in the lattice gives rise to the electronic energy band structure (the subject of Chapter 9), and thus establishes the electronic transport. The modern description of electronic band structure uses terminology derived from the structure of solids. Hence the knowledge of structure looms fundamentally important. Also related to electronic properties is the subject of defects. As will be covered in later chapters (particularly Chapter 4), defects in crystalline solids are defined relative to a perfect defect-free structure. Hence the ideal or perfect structure needs to be established, in order to then establish clear notions about defects. The X-ray diffraction techniques discussed here typically ignore defects and yield the ideal structure. There are, however, diffraction techniques that can determine crystallographic abnormalities. Furthermore defects dominate the electronic properties of semiconductors. The mechanical properties of solids depend on chemical bonding, which in crystalline solids is highly oriented. Thus the mechanical properties become first order functions of structure. Finally, surfaces in elec-

tronics devices are of tremendous significance, and the structure of surfaces, while often related to underlying bulk structure, is often different from bulk structure. Hence an understanding of the many properties that depend on structure of the surface requires the determination of surface structure. All the structural information crucial to virtually all the subfields of materials science are accessible using diffraction techniques.

Diffraction measures the effects on amplitude and phase of the interaction of monochromatic electromagnetic radiation (emr) and matter waves with a solid. Historically, as well as scientifically, the diffraction of X rays is of great significance. The wavelength of X rays ranges from fractions of a nm to a full nm, which is of the order of the size of atomic and interplanar spacings in typical crystal structures. Thus, as we will shortly see, the X-ray wavelength range is ideal for maximizing the phase changes that occur when the emr is scattered from atoms in a lattice. At this point it is useful to calculate the approximate wavelength for photons, electrons, and neutrons. For photons the formula is

$$E = h\nu = \frac{hc}{\lambda} \tag{3.1}$$

Using dimensions for λ of about 0.1 nm or 1 Å (1×10^{-10} m), in this formula, we calculate E for a photon (where $c = 1 \times 10^8$ m/s) as follows:

$$E = \frac{6.63 \times 10^{-34} \text{ J} \cdot \text{s} \cdot 3 \times 10^8 \text{ m/s}}{1 \times 10^{-10} \text{ m}} = 1.989 \times 10^{-15} \text{ J}$$

The conversion $1 \text{ eV} = 1.602 \times 10^{-19}$ J is then used to convert E to eV so that $E = 12415.7$ eV. Solving for λ in Å yields the approximate result:

$$\lambda = 12.40 \times 10^3 \text{ V}^{-1} \tag{3.2}$$

For particles with mass such as electrons where $m_e = 9.11 \times 10^{-31}$ kg, the deBroglie relationship for matter waves is used as follows:

$$\lambda = \frac{h}{p} = \frac{h}{mv}$$
$$\text{Kinetic energy} = eV = \frac{1}{2}mv^2, \quad v = \sqrt{\frac{2eV}{m}} \tag{3.3}$$

then

$$\lambda = \frac{h}{m\sqrt{\frac{2eV}{m}}} = \sqrt{\frac{150}{V}} \quad \text{for V in volts}$$

So electrons at 1×10^6 V yields a wavelength, $\lambda \approx 0.01$ Å. Likewise, for neutrons where $m_n = 1.68 \times 10^{-27}$ kg,

$$E = \frac{1}{2}mv^2, \quad \lambda = \frac{h}{\sqrt{2mE}} \tag{3.4}$$

at $E = kT$, $\lambda = 1$ Å, $T \approx 300 - 400$ K. Thus electrons in the appropriate energy range, and thermal neutrons also yield a useful range of wavelengths for diffraction.

The production of high-intensity monochromatic X rays is a well-developed technology. Suitable available wavelengths with high-wavelength precision are used to render X rays nearly ideal for diffraction from solids. However, the diffraction of other kinds of waves in addition to X rays is important, and among these are matter waves from electrons and neutrons. Particle diffraction depends on the energy derived from wavelengths of particles through the de Broglie relationship, as illustrated above. The main difference between particle and nonparticle emr diffraction is the efficiency of the diffraction, which is typically greater for particles. But other than this, an understanding of the fundamental aspects of X-ray diffraction underlies an understanding of all the other diffraction phenomena. Therefore, in the following treatment of diffraction, X-ray diffraction is emphasized as a way to discovering fundamental ideas, and later other diffraction results will be discussed in terms of the obtained results.

Before beginning the serious treatment of how the emr interaction with a lattice reveals structural information, it is useful to briefly look at diffraction results from various techniques illustrated in Figure 3.1. The array of lines, arcs, and spots on a photographic plate are typical of experimental results of diffraction from crystalline materials. Figure 3.1a shows so-called powder diffraction results taken from crystalline material ground to a fine powder before diffraction. Figure 3.1b is diffraction taken from a single crystal where the definite arrangement of spots is indicative of the crystal structure of the material. Figure 3.1c shows two electron diffraction micrographs where the left-hand image is for a polycrystalline material displaying rings while the right side shows both single crystal spots and polycrystalline rings. The raw diffraction data in this figure bespeaks of an underlying order. In this chapter the physics that gives rise to the result is explored, as this will naturally yield the ability to interpret the results in terms of underlying crystal structures. A powerful structural tool.

3.2   PHASE DIFFERENCE AND BRAGG'S LAW

Figure 3.2 shows three rays (A, B, and C) of monochromatic radiation (**E** field) traveling to the right and in phase at plane 11′ where the rays originate. Rays A and B travel the same distance and are in phase at every point equidistant from the source at plane 11′ because the rays are parallel. For example, at plane 22′ the rays A and B are in phase. Ray C, which is also traveling parallel to A and B, is in phase at 22′. However, ray C travels a greater distance to plane 33′. At plane 33′ the phase of C is different from the phases at 22′. This long-winded example is used simply to show that monochromatic waves produced in phase at the source will remain in phase for an equal distance of travel, but each ray may not remain in phase if waves are added or compared at different distances of travel. Let us assume that each ray has the same electric field (**E**) field ($\mathbf{E} = \mathbf{E}_A = \mathbf{E}_B = \mathbf{E}_C$) and consider a detector placed at 22′. The detector "sees" the resultant electric field disturbance (the superposition of the waves). Since the three waves traveled exactly the same path, they all arrive at 22′ in phase and the electric fields add to 3**E** because the path difference is 0. However, if for wave C we put its detector at 33′ where ray C is 180° out of phase with A or B, but leave the detectors for A and B at 22′, the total **E** field will be 2**E** − 1**E** = 1**E**. We will see that when emr of suitable wavelengths (i.e., for λ's of the size of the lattices) is scattered from crystal structures, the path and, consequently, phase differences will cause **E** field differences that lead to intensity
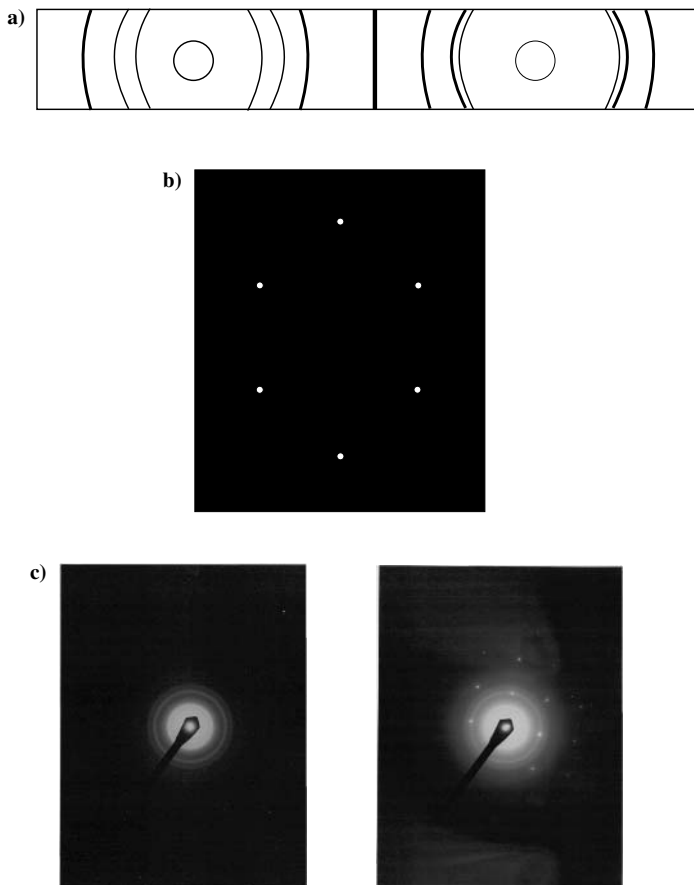
Figure 3.1 Various diffraction patterns: (*a*) Powder X-ray diffraction; (*b*) Laue back reflection X-ray diffraction; (*c*) transmission electron microscopy diffraction of polycrystalline material (*left*) and polycrystalline material on a single crystal substrate.
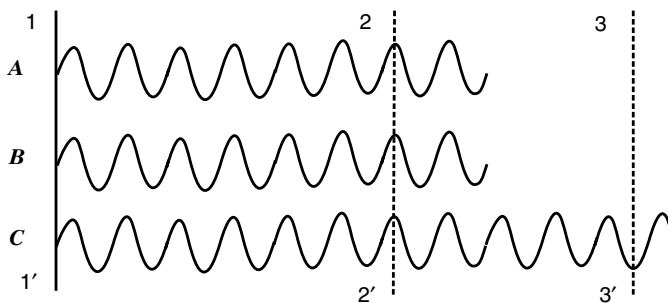
Figure 3.2 Three waves *A*, *B*, *C* in phase at the origin 11′ and 22′. Wave *C* displays different phase for different distance traveled to 33′.

differences that are actually measured. The process of interaction of emr with a crystal structure needs preliminary consideration. The electric field oscillations of a periodic X-ray emr interact with the electron clouds of the atoms in a lattice of atoms or the crystal structure. In general, a fraction of the emr field is absorbed and part is re-emitted. That part that is re-emitted unaltered with respect to the initial phase and wavelength is the only part from which diffraction will take place. This so-called unmodified emr retains phase coherency with the incident emr. The unmodified component of the scattered radiation resembles the reflection of light from a mirror, and for this reason diffraction is often spoken about as "reflection" of X rays. However, despite the common terminology that X-ray diffraction is reflection, the scattering of X rays from materials is decidedly different from light reflection. The coherently scattered component of the incident X rays is a small fraction of the total X-ray intensity, typically less than 1%, compared to true optical reflection where nearly 100% can be obtained using good mirrors. The coherently scattered emr has no phase change, whereas reflection involves a phase change. For reflection, the angle of incidence is the same as the angle of reflection, whereas for diffraction, specific angles are derived. Furthermore reflection is a surface effect whereas diffraction is a bulk effect. So, although common parlance may relate reflection and diffraction, the basic physics is decidedly different.

Figure 3.3 shows the interaction of X-ray emr with three equidistant rows (planes but seen end on as a row of atoms) of atoms (filled circles) where each plane is separated by distance x. Rays 1, 2, and 3 are incident at θ and, after scattering at atoms at O, E, and B, propagate to the detector at the far right that collects all the scattered rays. First, we consider two parallel rays 1 and 3 incident at θ (as measured from the plane surface) onto the top and bottom planes, respectively. From point O on the top plane a perpendicular is drawn to the incident beams 2 at D and 3 at A. This is similarly done to exiting rays 2' and 3' at F and C, respectively. Now the perpendicular line ODA forms an incident ray front up to which the three incident rays are in phase and have traveled the same distance from the source. Likewise, line OFC forms a front in the scattered rays, to the right of which the rays travel the same distance to the detector. In between these two fronts (ODA and OFC) the rays travel different distances that can result in phase differences.
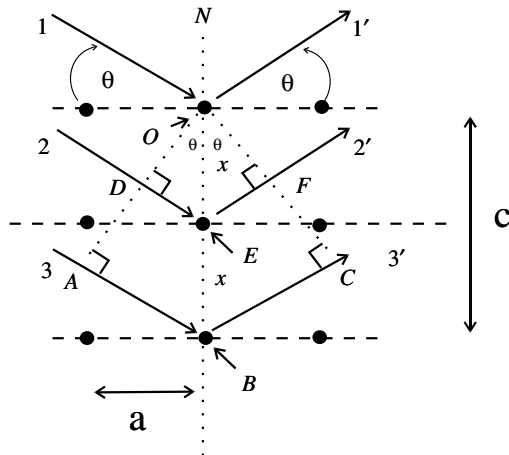

**Figure 3.3** Three waves impinging on three parallel equidistant planes.

Considering rays 1 and 3, we see that before scattering, ray 3 travels a distance AB more than ray 1, and that after scattering, ray 3 travels a distance BC more than ray 1. So the total path difference, δ, between rays 1 and 3 is $AB + BC$. δ is the path difference between incident and scattered rays. If this path difference is an integral number, n, of wavelengths, λ, or $n\lambda$, then rays 1 and 3 that become 1′ and 3′, respectively, are exactly in phase at the exit wave front plane at angle θ. Under this condition each segment AB and BC is equal to $2 \times \sin(\theta)$. This simple construction yields the usual form for Bragg's law:

$$n\lambda = 2d \sin(\theta) \tag{3.5}$$

where d is the interplanar spacing ($2x$ in the figure for rays 1 and 3) between the top and bottom planes.

Under the conditions shown in Figure 3.3, notice that for scattering from the top layer, surface scattering, at equal incident and scattered angles there is no path difference between incident and scattered waves. Consequently there is no diffraction from a surface under these conditions. However, unlike the incident radiation, it is not required to measure the scattered emr at the same angle θ. In principle, the detector can be placed to receive scattered radiation at any angle. For example, let us assume that in Figure 3.3, δ is not equal to $n\lambda$ but is $\lambda/2$. Then the intensity in the scattered wave front is identically 0, since the electric fields are 180° out of phase and add to 0. Now, if the detector is moved to different scattering angles, then different intensities are obtained due to the different paths. Maxima in intensity will appear sharply at certain angles, and these are termed Bragg angles. In actual diffraction experiments many rays are considered from many scattering centers (atoms), so the sharpness of the in-phase scattering at specific angles is great. To obtain diffraction from a surface, the emr is brought to the surface at normal incidence, as shown in Figure 3.4. In this case all the incident rays are in phase at the surface atoms. The detector scans hemispherically above the surface for backscattered radiation, and the maxima are noted. The angles at which maxima are obtained are used in the Bragg relationship to determine interplanar distances and indexes. If emr or matter waves that do not penetrate the surface atom layer are used, such as the matter waves from low-energy electrons, then the resulting diffraction is representative of the surface structure. This technique is called low-energy electron diffraction or LEED.

For typical diffraction problems it is often convenient to define the incident angle, θ, from the normal to the top plane, and then the diffraction angle, the angle between the incident and diffracted beam, is reported as 2θ, where θ is called the Bragg angle.
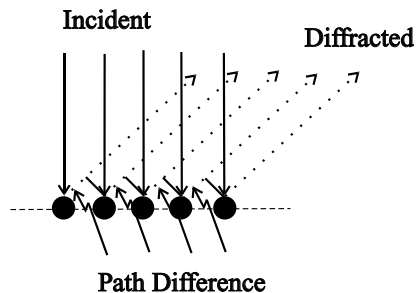
Figure 3.4   Incident waves at normal incidence to surface atom, diffracted at various angles.

Consider the Bragg equation (3.5) and a cubic lattice where from Table 2.4

$$\frac{1}{d^2} = \frac{\left(h^2 + k^2 + l^2\right)}{a_0^2} \tag{3.6}$$

Then, after squaring the Bragg equation and substituting the formula for d^2, the following formula is obtained and used to interpret measured scattering angles:

$$\sin^2(\theta) = \frac{\lambda^2\left(h^2 + k^2 + l^2\right)}{4a_0^2} \tag{3.7}$$

where 2θ is measured, and λ is known. From the measured Bragg angles the kind of unit cell (as determined by the hkl values for the planes diffracting) and size can be determined. However, in order to obtain the specific position of scattering sites (i.e., the atomic positions), more information is required. This information is contained in the intensities of the scattered radiation.

To see how atom positions affect scattered intensities, we return to Figure 3.3 and focus on the middle row of atoms that is exactly halfway between the top and bottom rows. We set the conditions of distances and angles so that the path difference for rays 1 and 3 scattered from the top and bottom planes yield a 0 path difference in the scattered waves when $ABC = n\lambda$. This situation should yield a maximum in the scattered intensity if we consider only the top and bottom planes and rays 11′ and 33′. However, we have the presence of the middle plane in Figure 3.3 to consider, and it is positioned at exactly half the distance from the top and bottom planes. The path difference of the ray scattered from the top plane to the middle plane is exactly $\frac{1}{2}$ that to the bottom plane so that from a similar triangles argument the following is obtained:

$$DE + EF = \frac{1}{2}(AB + BC) \tag{3.8}$$

Thus the coherent component of the emr scattered from the middle plane is 180° out of phase from the scattered radiation of both the top and bottom planes. With the radiation from the top and bottom planes arranged to be exactly in phase at θ, the addition of the middle plane reduces the scattered radiation. The addition of the new plane in between the top and bottom changes the originally set up maximum at θ to some lesser intensity. If this were the body-centered atom in a BCC or face-centered atoms in the FCC, one can imagine a drastic alteration for certain scattering angles relative to a simple cubic structure. This result will be made clearer when we consider how to quantify the scattered intensities and thereby determine atom positions, but it should be evident here that the specific atom positions greatly alter the scattered intensities. For this reason atom positions can be conversely determined from an analysis of the intensities. Before we continue with solving this problem, we need to take a step back and consider the different scattering intensities from an electron, from an atom, and then from an array of atoms.

3.3 THE SCATTERING PROBLEM

We first consider scattering from a gas and then from atoms on a crystal lattice. All gas particles (atoms and/or molecules) or scatterers in the gas are at random positions at any

instant, and so the phases of the scattered emr at the detector are also random. The intensity at a detector is given as

$$I = \mathbf{E}\mathbf{E}^* \tag{3.9}$$

where $\mathbf{E}$ is the electric field of the emr. For N scatterers in the gas

$$I = N\mathbf{E}^2 \tag{3.10}$$

in the case of a pure gas of one kind of atom or molecule where the scattering from each particle is the same. However, if the phases at the detector are coherent, then $\mathbf{E}$ for N scatterers is $N\mathbf{E}$ and the intensity at the detector is given as

$$I = (N\mathbf{E})^2 = N^2\mathbf{E}^2 \tag{3.11}$$

The huge difference, in coherence, of course, becomes evident when N is of order $\sim10^{20}$. But even a small amount of scattered radiation (it is an inefficient process) is made significant by coherency. This fact renders diffraction an important and measurable phenomenon.

3.3.1 Coherent Scattering from an Electron

The X-ray emr propagates from $-x$ to $+x$ and interacts with an electron at O, as shown in Figure 3.5. The scattered intensity, I, is measured at a point in the XZ plane, P, a distance OP ($OP = r$) from the electron, and at an angle 2θ determined from the direction of the incident radiation, I_0, from $-x$ toward x to the direction OP, or simply it is the angle between the transmitted and scattered radiation. β is the angle between X and P directions and is 90° in this example, since P was chosen to be in the XZ plane. The oscillating electric field of the emr in the YZ plane with random polarization induces an oscillation in the electron which in turn re-emits the emr. The coherent scattered portion of this event is governed by the Thompson equation:
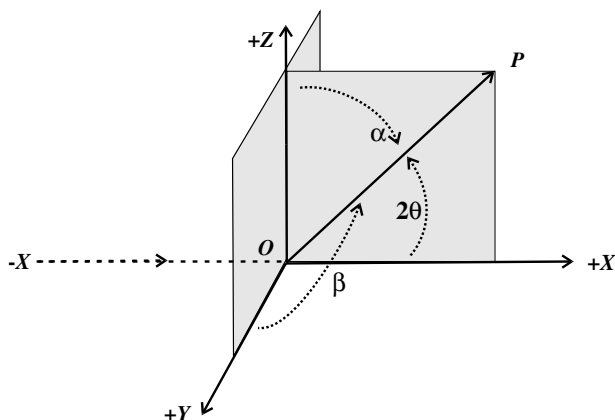
Figure 3.5 Electromagnetic radiation travelling from -X to X, stikes and electron at O. The radiation is measured at P in the XZ plane.

$$I = \frac{I_0 K \sin^2(\alpha)}{r^2} \tag{3.12}$$

where K is a constant that is proportional to $1/m^2$ where m here is the electron mass, and α is the angle between the direction of vibration of the electric field, **E**, of the incident emr, and the scattering direction, OP in Figure 3.5. With $K \approx 7.94 \times 10^{-30}\,\text{m}^2$, $I_p/I_0 \approx 7.94 \times 10^{-26}$ at 1 cm in the forward direction. According to this relationship, for emr propagating in the x direction, $I = 0$ at $\alpha = 0$ (in the Z direction). For a transverse wave, characteristic of an X ray, the **E** field direction is orthogonal to the propagation direction, which in Figure 3.5 is the YZ plane. Thus there is 0 scattered intensity in the plane normal to the propagation direction. The maximum scattered intensity, where $\sin(\alpha) = 1$ at 90° and 270°, is in the forward and backward scattering directions, $+OX$ and $-OX$.

For I_0 in the OX direction (coming from $-X$) and encountering an e^- at O, we can calculate I at P in the XZ plane. (Note that for simplicity the point P is in the XZ plane. In reality the detector can be anywhere.) Consider an unpolarized I_0 that has random **E** oscillating in the YZ plane but propagating in the $+X$ direction. This beam can be resolved into y and z components as

$$\mathbf{E}^2 = \mathbf{E}_y^2 + \mathbf{E}_z^2$$

$$\mathbf{E}_y^2 = \mathbf{E}_z^2 + \frac{1}{2}\mathbf{E}^2 \tag{3.13}$$

$$I_{oy} = I_{oz} = \frac{1}{2}I_0$$

We can write

$$I_{py} = I_{oy}\frac{K}{r^2}\sin^2\beta = I_{oy}\frac{K}{r^2}, \qquad \text{for } \beta = 90°$$

$$I_{pz} = I_{oz}\frac{K}{r^2}\sin^2\alpha, \qquad \alpha = 90° - 2\theta \tag{3.14}$$

with the identity $\sin(a - b) = \sin(a)\cos(b) - \cos(a)\sin(b)$ and with $a = 90°$ and $b = 2\theta$, $\sin^2(\alpha) = \cos^2(2\theta)$. Using the relationship above for I, we obtain

$$I_p = I_{py} + I_{pz} = I_0\frac{K}{r^2}\left[\frac{1 + \cos^2 2\theta}{2}\right] \tag{3.15}$$

This is the Thompson equation for scattering from an electron. The important message is that the intensity measured at any position is a function of θ simply due to the nature of the scattering physics of transverse electromagnetic radiation waves. Also scattering in the forward and reverse directions is strongest and weakest at 90° to the direction of propagation. As a result the intensities at specific Bragg angles must be corrected for this θ dependence before the intensity changes due to atomic position can be determined.

It is also possible to have incoherent or Compton scattering from an electron. Then some of the energy of the incident photon will convert to kinetic energy for the electron. The increase in λ for the incident photon will depend on the scattering angle:

$$\Delta\lambda = 0.0486 \sin^2 \theta \qquad (3.16)$$

With random phase, compton modified radiation is produced, so there is no diffraction. Incoherent scattering is usually a problem for loosely bound electrons (typically occurring for atoms of low atomic number, Z, e.g., C), which form a background to coherently scattered intensities.

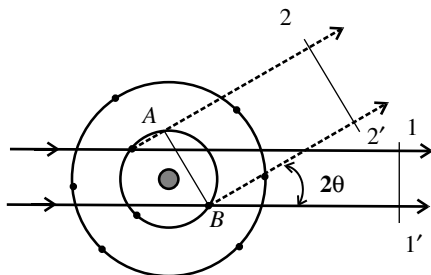### 3.3.2 Coherent Scattering from an Atom

Scattering from an atom is easily understood by considering that an atom has a number of electrons associated with it and each atom will scatter emr, as discussed above. The radiation coherently scattered from the nucleus is very small because, as mentioned above, $K \propto 1/m^2$ and is small for the relatively large mass of the nucleus compared to the electron mass. An atomic scattering factor f is defined as

$$f = \frac{I_{atom}}{I_{electron}} \qquad (3.17)$$

Equation (3.17) compares the atomic scattered intensity to the intensity scattered from a single electron. f is given by the atomic number (Z) for scattering in the direction of propagation of the emr, since Z determines the number of electrons, and in that direction no phase differences exist for the emr scattered from the electrons of an atom. In Figure 3.6 this plane is labeled 11′. However, at all other angles (e.g., plane 22′) f is reduced from Z according to the angle θ (as $\sin \theta$) and the wavelength, λ, by the factor $\sin(\theta)/\lambda$. Shorter wavelengths exacerbate the phase differences at any angle. Tables of atomic scattering factors for the elements are available, and Figure 3.7 is a plot of the atomic scattering factor for Cu and Al.

3.3.3 Coherent Scattering from a Unit Cell

Scattering from a unit cell is a straightforward problem at this point. All we need to do is to keep track of the scattering from the atoms in the unit cell at the unique positions in the cell, and the phase differences that occur due to the relationship of one atom in the cell to the others. We formulate the problem, again, with the help of Figure 3.3 where monochromatic X-ray radiation (rays 1, 2, 3) is incident at θ on three parallel

Figure 3.6 X-ray scattering from an atom that has several electrons. The incident radiation travels from the left, scatters from two electrons (*small filled circles*), and is measured at planes 11′ and 22′.
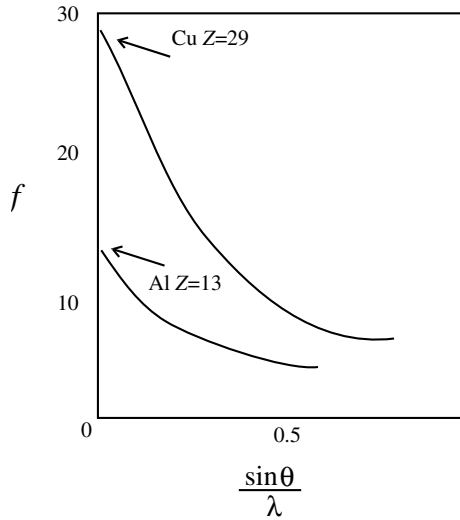
Figure 3.7   Atomic scattering factors *f* for Cu and Al as a function of the ratio of angle to wavelength.

planes with atoms at positions labeled *O*, *E*, and *B*. Coherently scattered radiation is also measured at θ and shown as 1′, 2′, and 3′. The 2-D unit cell dimensions are *a* and *c*. First we calculate the path difference between 1 and 3 as δ_{13}. The path difference δ_{13} is given in Figure 3.3 as *ABC* and will result in constructive interference according to Bragg's law when

$$\delta_{13} = ABC = 2d_{001}\sin(\theta) = \lambda \tag{3.18}$$

Notice that the distance between the bottom and top planes is the *c* lattice parameter and so can be indicated as d_{001}. Now consider the effect of the plane with atom *E* that is half way between *O* and *B* has on this situation. *E* is *x* away from *O* and *x* away from *B*. If one notices the similar triangles in Figure 3.3 (*ODE* similar to *OAB* and *OEF* similar to *OBC*), then we obtain the following

$$\frac{EF}{BC} = \frac{OE}{OB}, \quad \frac{DE}{AB} = \frac{OE}{OB} = \frac{DE}{BC} \tag{3.19}$$

When used in *ABC*, the equation above yields

$$\mathrm{DEF} = 2BC\frac{OE}{OB} = ABC\frac{OE}{OB} \tag{3.20}$$

With *ABC* = λ as given above, we obtain

$$\delta_{12} = \mathrm{DEF} = \frac{OE}{OB}\lambda = \left[\frac{x}{\left(\dfrac{c}{l}\right)}\right]\lambda \tag{3.21}$$

Now it is useful to convert the calculated path differences to angular phase differences. This is done simply, using the relationship that one wavelength (λ) traverses 360°. From the conversion 360° = 2π radians, the phase difference ϕ is then given as

$$\phi = \left(\frac{\delta}{\lambda}\right)2\pi \tag{3.22}$$

where δ/λ is the fraction of 2π radians yielding the angular phase difference. The phase difference between O and E is then readily calculated:

$$\phi_{12} = \left(\frac{\delta_{12}}{\lambda}\right)2\pi = \left(\frac{lx}{c}\right)2\pi \tag{3.23}$$

In 3-D, let $u = x/a$, $v = y/b$, and $w = z/c$, which yields

$$\phi_{12} = 2\pi lw \tag{3.24}$$

where the u, v, w are the coordinates of the atoms (recall the major translations $\frac{1}{2}, \frac{1}{2}, \frac{1}{2}$, . . . , from Chapter 2). This can be generalized for the hkl reflections, that is, diffraction from the hkl plane (hkl):

$$\phi = 2\pi(hu + kv + lw) \tag{3.25}$$

This relationship for the phase difference ϕ is the phase difference for a point with the positions x/a, y/b, z/c, or as above, u, v, w, and the origin for the h, k, l reflection.

   We now can proceed to quantify the phase difference, ϕ, for any occupied atomic position by any atom with scattering ability, f. To complete this problem, we need to add together all the ϕ's for all the unique atomic positions in a given unit cell. We return to the use of the unique lattice positions from which scattering will also be different for the different unique positions.

   One efficient way to accomplish the task of simultaneously adding together two quantities, amplitude and phase, is to use complex arithmetic. The added feature for the use of complex notation is the ability to use exponentials, and this will be shown below. We first consider that periodic waves can have different phases (ϕ_a and ϕ_b) and amplitudes (a and b), as is shown for two waves in Figure 3.8. Note that the waves need to have the
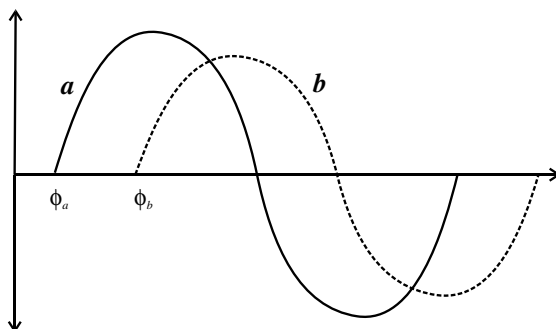
Figure 3.8   Two waves, *a* and *b*, each with different amplitude and phase.
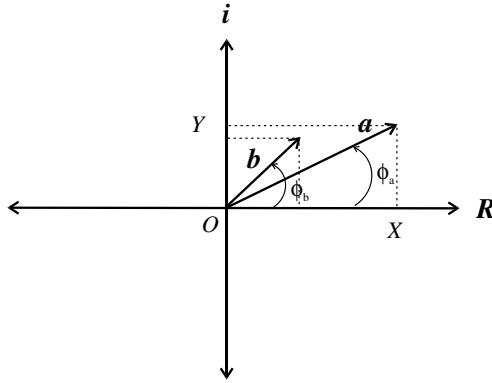
Figure 3.9 Waves *a* and *b* from Figure 3.8 can be represented as complex numbers with different projections on the real *R* and imaginary *i* axes.

same wavelengths. Figure 3.9 shows these two waves displayed on the complex plane as vectors **a** and **b** with a length and direction. For wave **a** in Figure 3.9, the projection on the real axis is *OX* and on the imaginary axis is *OY*. The components are given as

$$OX = a\cos(\phi_a) \quad \text{and} \quad OY = ia\sin(\phi_a) \tag{3.26}$$

and the resultant is

$$a\cos(\phi_a) + ia\sin(\phi_a)$$

It can be readily shown that for the substitution of power series for cos, sin, and e^x we can obtain for the resultant wave:

$$ae^{i\phi_a} = a\cos(\phi_a) + ia\sin(\phi_a) \tag{3.27}$$

which derives from the identity $e^{i\phi} = \cos\phi + i\sin\phi$.

Using this expression, we now substitute for a and ϕ from the calculations above:

$$ae^{i\phi} = f\, e^{2\pi i(hu+kv+lw)} \tag{3.28}$$

For a unit cell we need to consider scattering from all the different atomic sites. When done, this yields the so-called structure factor, *F*, as the sum over all the *n* sites:

$$F_{hkl} = \sum_n f_n e^{2\pi i(hu_n + kv_n + lw_n)} \tag{3.29}$$

The intensity is calculated in the usual way from the electric field as $I = \mathbf{EE}^*$ or FF*

3.3.4 Structure Factor Calculations

The simplest problem to address is a primitive cubic. The first step in the calculation of *F* is to determine the unique atom coordinates. For a primitive cubic, $(u, v, w) = (0, 0, 0)$.

If we assume only one kind of atom (with therefore one kind of f), then the structure factor above from equation (3.29) is given as

$$F_{hkl} = \sum_n f_n e^{2\pi i(h\cdot 0 + k\cdot 0 + l\cdot 0)} = fe^0 = 1 \tag{3.30}$$

for all values of (h, k, l). Then we can proceed to calculate the intensity $I = F^2 = f^2$, which is the same for all hkl, all (hkl) planes diffract and with equal intensity.

To this primitive cubic cell we add a base center atom at the coordinates $(\frac{1}{2}, \frac{1}{2}, 0)$. Thus, for this case there are two uvw's at $0, 0, 0$ and $\frac{1}{2}, \frac{1}{2}, 0$. For this situation F from equation (3.29) given as

$$\begin{aligned} F_{hkl} &= fe^{2\pi i(h\cdot 0 + k\cdot 0 + l\cdot 0)} + fe^{2\pi i(1/2h + 1/2k)} \\ &= f + fe^{\pi i(h+k)} = f\left(1 + e^{\pi i(h+k)}\right) \end{aligned} \tag{3.31}$$

To interpret this result, we simply use the result

$$e^{n\pi i} = (-1)^n \tag{3.32}$$

This relation teaches that for n even $e^{(\text{even } \pi i)} = 1$ and for n odd $e^{(\text{odd } \pi i)} = -1$.

Returning to the problem, we have for the sum $(h + k)$ even, $F = 2f$ and $F^2 = 4f^2$, but for $(h + k)$ odd $F = f(1 - 1) = 0$. The value for l does not matter because it was always multiplied by 0 and removed from F. It is useful to make a table of allowed plane indexes with the selection rules developed from F (this way only planes with $h + k$ even will yield an intensity and all intensities will be the same). All other planes yield $I = 0$. In order to achieve $h + k$ even, h and k need to be either even or odd but not mixed: $11x$, $22x$, $33x$, $44x$, $55x$, $35x$, $24x$, but not $12x$, $23x$, $34x$. Here x is any integer and, of course, 0 is neutral.

Now we consider a BCC with uvw of $0, 0, 0$ and $\frac{1}{2}, \frac{1}{2}, \frac{1}{2}$, and this yields

$$\begin{aligned} F &= fe^{2\pi i 0} + fe^{2\pi i(1/2h + 1/2k + 1/2l)} \\ &= f\left[1 + e^{\pi i(h+k+l)}\right] \end{aligned} \tag{3.33}$$

Then we deduce that $F = 2f$ for $h + k + l$ even, and $F = 0$ for $h + k + l$ odd. This result does not allow the (100) to diffract. Recall that this was the example we considered when we put a plane in between a diffracting set of planes at $d/2$.

The last example illustrated here is the FCC where unique atom positions are at the following (u, v, w): (000), $(\frac{1}{2}\frac{1}{2}0)$, $(\frac{1}{2}0\frac{1}{2})$, $(0\frac{1}{2}\frac{1}{2})$. This yields F with four terms:

$$\begin{aligned} F &= fe^{2\pi i 0} + fe^{2\pi i(h/2 + k/2)} + fe^{2\pi i(h/2 + l/2)} + fe^{2\pi i(k/2 + l/2)} \\ &= f\left[1 + e^{\pi i(h+k)} + e^{\pi i(h+l)} + e^{\pi i(k+l)}\right] \end{aligned} \tag{3.34}$$

If hkl are all odd or all even, then the sum of any two is even and $F = 4f$ and $F^2 = 16^2$ (i.e., the sum of two odd or two even integers is an even integer). If hkl are mixed, then there are two odd or two even. When taken in pairs they will yield three sums, two of which will be between an even and an odd. The result will be two odd sums and one even. Thus two -1's added to two $+1$'s will yield $F = 0$ as

Table 3.1   Results of structure factor calculations

$h^2 + k^2 + l^2$	PC	BCC	FCC	DC
1	(100)	—	—	—
2	(110)	(110)	—	—
3	(111)	—	(111)	(111)
4	(200)	(200)	(200)	—
5	(210)	—	—	—
6	(211)	(211)	—	—
7	—	—	—	—
8	(220)	(220)	(220)	(220)
9	(300), (221)	—	—	—
10	(310)	(310)	—	—
11	(311)	—	(311)	(311)

$$F = f\left[1 + e^{\text{even }\pi i} + e^{\text{odd }\pi i} + e^{\text{odd }\pi i}\right] = f[1 + 1 - 1 + -1] = 0 \tag{3.35}$$

Table 3.1 summarizes the structure factor calculations for the PC, BCC, FCC, and DC cubic structures for low-index planes.

3.4   RECIPROCAL SPACE, RESP

It is always advantageous in science to define coordinates that match closely the system being studied. For example, when studying a system with spherical symmetry, the applicable formulas are almost always simplified with the use of spherical coordinates rather than Cartesian coordinates. Familiar problems in classical and quantum mechanics involve a transform of coordinates to reduce the algebra and arrive at the simplest formulas. Reciprocal distances are the appropriate units to use to connect the diffraction physics discussed above with the experimental results (the experimental diffraction patterns shown in Figure 3.1).

3.4.1   Why Reciprocal Space?

From the discussion of crystal structures in Chapter 2, we arrived at the conclusion that a specific structure has lattice parameters associated with the unit cell size and shape, and interplanar spacings, or so-called d spacings. Up to this point in Chapter 3 we have discussed how emr and matter waves, in particular, X rays, interact with the arrangement of atoms to yield diffraction events as reinforced intensities at specific angular locations, and with specific intensities relative to the incident radiation. The fundamental relationship uniting planar indexes, planar spacings, and the measurable diffraction angles is Bragg's law. When rearranged appropriately, this relationship reveals the natural coordinates, which allow the simplest interpretation of a diffraction experiment. Specifically, it is seen that the measured Bragg angles are proportional to $1/d$, so rewriting Bragg's law (equation 3.5), we obtain the following formula in terms of reciprocal d spacings:

$$\sin(\theta_{hkl}) = n\lambda \cdot \frac{1}{2d_{hkl}} \tag{3.36}$$

Furthermore from equation (3.6) for a cubic structure, $1/d$ is proportional to $1/a$ as

$$\sin(\theta_{hkl}) = \frac{n\lambda}{2} \cdot \frac{\left(h^2 + k^2 + l^2\right)^{1/2}}{a_0} \tag{3.37}$$

Thus the measured Bragg angles, θ, are found to be directly related to reciprocal distances in the lattice. While for the case of simple structures this fact may not seem so impressive, for complex structures and many other ideas in solid state physics, notably electronic energy band structure (to be discussed in Chapter 9), reciprocal space rather than real space provides the most suitable coordinates with which to describe interactions of emr and electrons with a crystalline material. This case will be demonstrated throughout this text, and the reasons are those simply stated above: that the superposition of waves depends on the reciprocal size of the unit cell. Whether the waves are massless waves, such as X-ray photons or matter waves derived from particles with mass such as electrons, is largely immaterial to describing the diffraction effects that often dominate the information carrying ability of the interaction.

While reciprocal space (RESP) is tremendously important for the understanding and simplification of myriad problems of physics, this new space is not intuitive as is real space. Hence most people studying RESP for the first time and even for many times thereafter cannot really "feel" the new coordinate system. While somewhat uncomfortable, this should not result in despair. Much is to be gained from simply remembering the definitions and learning how to manipulate and calculate using the ideas of RESP. As familiarity increases, so will ones ability to "feel" this concept. For those of us who have spent a goodly fraction of their lifetime studying various aspects of materials physics, this discomfort with new ideas is commonplace. However, it is to be tolerated with patience, and the realization that when huge amounts of information are virtually perfectly understood, simply by choosing the natural coordinates, then the power and relevance of the ideas and the impact on science is also understood.
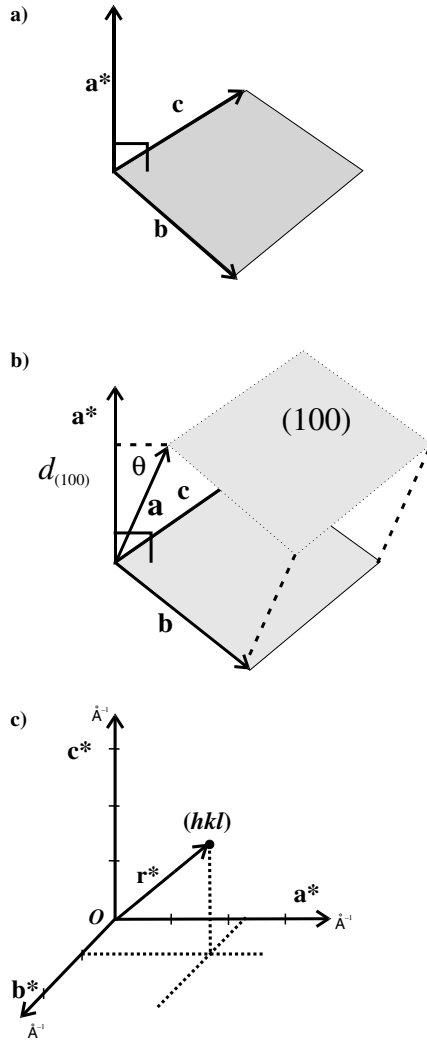
3.4.2 Definition of RESP

First a reciprocal lattice, REL, is defined by reciprocal lattice vectors $\mathbf{a}^*$, $\mathbf{b}^*$, and $\mathbf{c}^*$ using

$$\mathbf{a}^* = \frac{(\mathbf{b} \times \mathbf{c})}{\mathbf{a} \cdot (\mathbf{b} \times \mathbf{c})}, \quad \mathbf{b}^* = \frac{(\mathbf{c} \times \mathbf{a})}{\mathbf{b} \cdot (\mathbf{c} \times \mathbf{a})}, \quad \mathbf{c}^* = \frac{(\mathbf{a} \times \mathbf{b})}{\mathbf{c} \cdot (\mathbf{a} \times \mathbf{b})} \tag{3.38}$$

Figure 3.10a shows that the magnitude of the vector $(\mathbf{b} \times \mathbf{c})$ is the area of the plane (shaded) defined by the real space vectors $\mathbf{b}$ and $\mathbf{c}$, and the direction is normal to the bc plane. $\mathbf{a} \cdot (\mathbf{b} \times \mathbf{c})$ is the volume of the solid defined by $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{c}$. So the area of the base of a 3-D figure divided into the volume yields the height. This height is perpendicular to the basal area and has a magnitude of $1/\mathbf{a}^*$ and for a cubic system:

$$\mathbf{a}^* \perp \mathbf{b} \text{ and } \mathbf{c}, \quad \mathbf{b}^* \perp \mathbf{c} \text{ and } \mathbf{a}, \quad \mathbf{c}^* \perp \mathbf{a} \text{ and } \mathbf{b},$$

The solid figure defined by $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{c}$ has a basal area $(\mathbf{b} \times \mathbf{c})$, and the volume $\mathbf{a} \cdot (\mathbf{b} \times \mathbf{c})$ can be thought of as a unit cell bounded on top and bottom by planes with indexes (hkl) and with area $(\mathbf{b} \times \mathbf{c})$. The distance between the planes is the interplanar spacing, d_{hkl},

Figure 3.10 Reciprocal space (RESP) representations: (*a*) Definition of **a***; (*b*) **a*** in a non-orthogonal system; (*c*) reciprocal lattice vector **r*** in RESP.

which is the height of the 3-D figure and is the perpendicular distance from the bottom to the top plane. Now, as was determined above, this height has the direction of **a***,

$$|\mathbf{a}^*| = \frac{1}{a\cos\theta} = \frac{1}{d_{100}} \tag{3.39}$$

For orthogonal vectors **a**, **b**, and **c**, the height also has the magnitude of **a** or $1/\mathbf{a}^*$. This geometry is clarified further by considering a triclinic cell in Figure 3.10*b* where the top plane is the (100) plane, and d_{100} or $1/\mathbf{a}^*$ is shown at an angle θ to **a**. The volume of this

triclinic cell is the area of the base ($\mathbf{b} \times \mathbf{c}$), dotted by the height which is d_{100} or $|\mathbf{a}|\cos\theta$ or $1/|\mathbf{a}*|$. Some other relationships are

$$|\mathbf{a}*| = \frac{1}{d_{100}}, \quad \mathbf{a}* \cdot \mathbf{a} = 1$$

but $\mathbf{a}* \perp \mathbf{b}$ and $\mathbf{c}$

thus $\mathbf{a}* \cdot \mathbf{b} = 0, \quad \mathbf{a}* \cdot \mathbf{c} = 0$

Since $\cos 90° = 0$,

$$\mathbf{c}* = \frac{1}{d_{001}}, \quad \mathbf{b}* = \frac{1}{d_{010}}$$

3.4.3   Definition of a Reciprocal Lattice Vector

It is useful to define a reciprocal lattice vector, $\mathbf{r}*$, as follows:

$$\mathbf{r}_{hkl}^* = h\mathbf{a}* + k\mathbf{b}* + l\mathbf{c}* \tag{3.40}$$

$\mathbf{r}*$ is drawn from the origin of RESP to the hkl point, which is a plane in real space (hkl) as is shown in Figure 3.10c. With the help of Figure 3.11 it can be shown that

$$\mathbf{r}* \perp (hkl) \quad \text{and} \quad |\mathbf{r}*| = \frac{1}{d_{hkl}} \tag{3.41}$$

From Figure 3.11 the plane ABC is defined in real space with indexes (hkl). So the following relationships obtain among the defining vectors:

$$OA + AB = OB \quad \text{and} \quad OA + AC = OC \tag{3.42}$$

If $\mathbf{r}*$ is $\perp$ to plane ABC, then $\mathbf{r}*$ is also $\perp$ to all the lines in the plane, AB, AC, and so on. We can express the various vectors in terms of fractional intercepts as

$$AB = OB - OA = \frac{\mathbf{b}}{k} - \frac{\mathbf{a}}{h} \quad \text{and} \quad AC = OC - OA = \frac{\mathbf{c}}{l} - \frac{\mathbf{a}}{h} \tag{3.43}$$
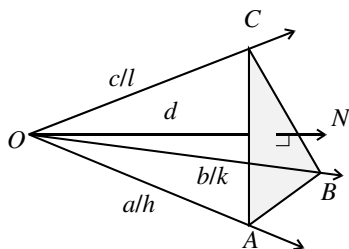

Figure 3.11   Normal N to plane ABC with fractional intercepts.

Then we can form the scalar (dot) products of **r*** with lines in the plane, in order to test the orthogonality:

$$\mathbf{r}^* \cdot AB = (h\mathbf{a}^* + k\mathbf{b}^* + l\mathbf{c}^*) \cdot \left(\frac{\mathbf{b}}{k} - \frac{\mathbf{a}}{h} \right) = -1 + 1 + 0 = 0 \tag{3.44}$$

$$\mathbf{r}^* \cdot AC = (h\mathbf{a}^* + k\mathbf{b}^* + l\mathbf{c}^*) \cdot \left(\frac{\mathbf{c}}{l} - \frac{\mathbf{a}}{h} \right) = -1 + 1 + 0 = 0 \tag{3.45}$$

Thus **r*** is $\perp$ to two lines in the ABC plane; hence **r*** must be $\perp$ to the plane ABC or (hkl).

With d_{hkl} as the distance from the origin O to the plane ABC, we observe that d is the projection of $\mathbf{a}/h$ on the ABC plane normal **N**. We have the following relationships:

$$d_{hkl} = \left(\frac{\mathbf{a}}{h} \right) \cdot \mathbf{N} = \left(\frac{\mathbf{a}}{h} \right) \cdot \left(\frac{\mathbf{r}^*}{|\mathbf{r}^*|} \right) = \left| \frac{1}{\mathbf{r}^*} \right| \tag{3.46}$$

From these equalities we can see clearly the importance of **r*** as the vector from the origin of reciprocal space to an hkl point with magnitude $1/d$; hence it is related to the interplanar spacings. Furthermore the hkl point is a plane in real space.

An example for the cubic system is the distance formula equation (3.6):

$$\frac{1}{d^2} = \frac{\left(h^2 + k^2 + l^2 \right)}{a_0^2} \tag{3.6}$$

It was derived in Chapter 2 using direction cosines. Although this method is satisfactory for the cubic system, it becomes unwieldy for more complicated crystal systems. Since $|\mathbf{r}^*| = 1/d$, we can use the expression

$$\mathbf{r}^* \cdot \mathbf{r}^* = |\mathbf{r}^*||\mathbf{r}^*|\cos\theta \tag{3.47}$$

and substitute equation (3.40):

$$\mathbf{r}^* = h\mathbf{a}^* + k\mathbf{b}^* + l\mathbf{c}^* \tag{3.40}$$

The result is

$$(h\mathbf{a}^* + k\mathbf{b}^* + l\mathbf{c}^*) \cdot (h\mathbf{a}^* + k\mathbf{b}^* + l\mathbf{c}^*) = h^2\mathbf{a}^{*2} + k^2\mathbf{b}^{*2} + l^2\mathbf{c}^{*2}$$

$$= \left(h^2 + k^2 + l^2 \right) a^{*2} = \frac{\left(h^2 + k^2 + l^2 \right)}{a_0^2} = \frac{1}{d^2} \tag{3.48}$$

if $\mathbf{a} = \mathbf{b} = \mathbf{c}$ and thus $\mathbf{a}^* = \mathbf{b}^* = \mathbf{c}^*$, and for a cubic system $\mathbf{a}^* = 1/\mathbf{a}$. Thus the use of reciprocal space simplifies the derivation of the d spacing formulas. For the other crystal systems the axes may be unequal and not orthogonal.

The directions for the various **r*** express the relative planar angles, as the angles between **r***. With this knowledge one can derive interplanar angle formulas. For example, for planes $(h_1k_1l_1)$ and $(h_2k_2l_2)$ the interplanar angle can be calculated as follows:

$$\mathbf{r}_1^* \cdot \mathbf{r}_2^* = \left|\mathbf{r}_1^*\right|\left|\mathbf{r}_2^*\right|\cos(\theta) \tag{3.49}$$

where θ is the angle between the planes. We know that

$$\mathbf{r}_1^* = h_1\mathbf{a}^* + k_1\mathbf{b}^* + l_1\mathbf{c}^* \quad \text{and} \quad \mathbf{r}_2^* = h_2\mathbf{a}^* + k_2\mathbf{b}^* + l_2\mathbf{c}^*$$

and from above that $\mathbf{a}^* \cdot \mathbf{a}^* = 1$, $\mathbf{a}^* \cdot \mathbf{b}^* = 0$, and $\mathbf{a}^* \cdot \mathbf{c}^* = 0$. The magnitude for the vector $\mathbf{r}^*$

$$|\mathbf{r}^*| = \sqrt{h^2 + k^2 + l^2} \tag{3.50}$$

The interplanar angle formula for the cubic system is written

$$\cos\theta = \frac{\mathbf{r}_1^* \cdot \mathbf{r}_2^*}{\left|\mathbf{r}_1^*\right| \cdot \left|\mathbf{r}_2^*\right|} = \frac{h_1h_2 + k_1k_2 + l_1l_2}{\sqrt{h_1^2 + k_1^2 + l_1^2}\sqrt{h_2^2 + k_2^2 + l_2^2}} \tag{3.51}$$

It is useful to construct a REL in RESP. From the definitions above several points are useful to guide the construction:

1. The symmetry of the REL must be the same as that of the real lattice.
2. Planes in real space are points in RESP.
3. Distances in RESP are reciprocal of the real space distances for orthogonal systems.
4. $\mathbf{r}_{hkl}^*$ is the vector from the origin of RESP to the plane (point in RESP) and equals $1/d_{hkl}$.

We commence the construction by arbitrarily assigning the origin of RESP as O in Figure 3.12. We also restrict the construction to a cubic real lattice with $a_0 = 2$ Å. Thus the axes in Figure 3.12 are in units of Å^{-1}. For the (100), the real space distance is 2 Å while $a^* = 0.5$ Å^{-1}. This distance is the same for the (010) and (001) in a cubic lattice. The three planes are indicated as points on the REL in RESP. Note that for the (200) in real space the distance is half that of the (100) or 1 Å from the d spacing formula above and $1/d$ is 1 Å^{-1}, and the (200), (020), and (002) planes are shown. For the (300) plane, $d = 0.67$ Å and $1/d = 1.5$ Å^{-1}, and two of these planes are shown. The calculations are summarized in Table 3.2 using the previous formula for $1/d^2$ from Table 2.4 and recalling equation (3.41) that $|\mathbf{r}_{hkl}^*| = 1/d_{hkl}$. From these planes it is easily seen that as the Miller indexes increase, the planes (points in RESP) increase in distance from O for RESP. This is exactly the opposite of real space. A number of other planes are also shown in the REL. In summary, the REL has the same symmetry as the real space lattice, but the distances are reversed.

3.4.3 The Ewald Construction

The RESP and REL ideas presented above enable the representation of diffraction conditions to be simplified. An added piece of simplification to the concepts and implications of the RESP is an elegant geometrical representation of diffraction by Ewald. The significance of this construction is that it goes beyond simply helping one understand or summarize diffraction. Its geometric representation enables one to construct and derive
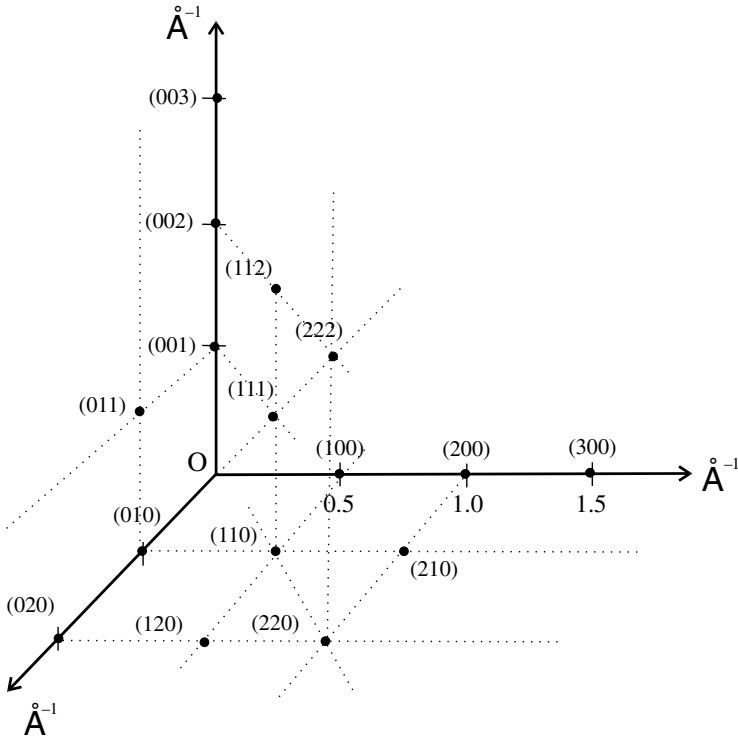
Figure 3.12   A map of reciprocal space (RESP) showing selected planes as points in RESP.

Table 3.2   Cubic REL, $a_0 = 0.2$ nm

hkl	$1/d^2$ nm^{-2}	$|r^*| = 1/d$ nm^{-1} (Å^{-1})
(100)	25	5 (0.5)
(200)	100	10 (1)
(300)	225	15 (1.5)
(400)	400	20 (2.0)
(110)	50	7.07 (0.7)
(220)	200	14.1 (1.4)
(330)	450	21.2 (2.1)
(440)	800	28.3 (2.8)
(210)	125	11.2 (1.1)
(111)	75	8.7 (0.9)
(112)	150	12.2 (1.2)
(222)	300	17.3 (1.7)

all the experimental diffraction methods directly and simply in terms of a geometric construction.

Essentially this construction considers the geometrical consequences of emr or matter waves incident with magnitude $1/\lambda$ and direction $\mathbf{S_0}$ (a unit vector) as the incident vector

$\mathbf{S_0}/\lambda$, interacting with a fixed REL. Figure 3.13 shows the Ewald construction using a 2-D array of REL points. These points in RESP are of course planes in real space. We can consider any REL point as the origin for the construction and label this point O (at the left of the REL). For any wavelength λ, we construct a circle (in 2-D) with radius $1/\lambda$, that touches O (in 3-D this construction yields a sphere). With the center of the circle labeled C, we label the line CO, as the vector $\mathbf{S_0}/\lambda$ pointing from C to O, meaning it originates at C and terminates on O. Any direction can be chosen for $\mathbf{S_0}$. In reality, and as discussed below, this direction is experimentally determined by the direction of the incident emr beam from the machine that produces the beam of X rays that impinge on a sample. From O draw a vector $\mathbf{OP}$ to any other point on the REL that lies exactly on the circle. Notice that $\mathbf{OP}$ to point P_{hkl} is a REL vector, $\mathbf{r}^*_{hkl}$ (by definition). The diffracted beam, $\mathbf{S}/\lambda$ completes the construction when drawn from C to P. It should be noted that $\mathbf{S}/\lambda$ has the same magnitude as $\mathbf{S_0}/\lambda$, since both are radii of the same circle, but these vectors have different directions. Now we will see what information is summarized by this construction.

From this construction there are several relationships that are useful to write down:

$$\mathbf{OP} = \mathbf{r}^*_{hkl} = \frac{\mathbf{S} - \mathbf{S}_0}{\lambda} \tag{3.52}$$

$$\sin(\theta) = \frac{\frac{1}{2}|\mathbf{r}^*|}{1/\lambda} \tag{3.53}$$

Recall from equation (3.46) that $\mathbf{r}^*_{hkl} = 1/d_{hkl}$. Substituting this into equation (3.53), we obtain Bragg's law:

$$\lambda = 2d\sin(\theta) \tag{3.5}$$

With point P_{hkl} on the perimeter of the circle with radius $1/\lambda$, Bragg's law is satisfied. This means that the (hkl) plane represented by P_{hkl} diffracts, yielding a diffracted
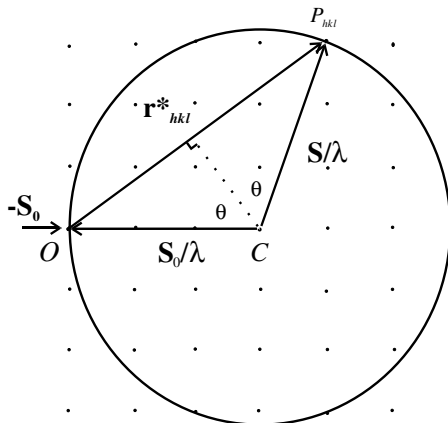

Figure 3.13 Ewald construction (circle) in 2-D RESP.

beam at 2θ relative to the incident beam. In general, all the points that lie on the Ewald circle will diffract at different angles relative to the incident beam. Because this construction can be made, with the severe restriction that an **r*** exists on the circle, λ is appropriate to the REL and so will effect diffraction from certain REL points at the appropriate direction(s). It may be the case that with a particular choice of direction and wavelength for the incident beam, there are no REL points that lie on the circle. The points P_{hkl} and directions 2θ are identified in the construction after the vector $\mathbf{S_0}/\lambda$ is identified. The Ewald construction in 2-D yields a diffraction circle, and for a usual diffraction problem in 3-D, a diffraction sphere is obtained that is usually called an Ewald sphere.

3.5 DIFFRACTION TECHNIQUES

The great power of the RESP representation along with the Ewald construction is manifest in the simplicity with which all the experimental diffraction techniques are represented. The experimental methods are directly distinguished by using the RESP representation and considering the different ways REL points, namely the planes in real space, are brought onto the surface of the Ewald sphere (i.e., brought into diffracting conditions). Among the parameters that affect the Ewald sphere are the wavelength, λ, which determines the radius of the sphere 1/λ, and the direction of the incident radiation or more generally the orientation of the Ewald sphere relative to the REL. Below only several prominent X-ray diffraction techniques are presented from the point of view of how each technique manipulates the RESP and Ewald sphere to enable diffraction. No attempt is made to treat the experimental details.

3.5.1 Rotating Crystal Method

This method is used in a number of manifestations for determining entire crystal structures, in particular, the shape and size of unit cells and atom positions. This method also nicely illustrates the point made above about the relationship of the size and orientation of the Ewald sphere relative to the REL. The technique uses monochromatic X rays and requires a single crystal of material. As shown in Figure 3.14 the crystal in the figure, as represented by its REL, is being rotated about an axis. Prior to rotation one point of the REL touches the surface of the Ewald sphere. When the REL rotates other planes (points in RESP) eventually touch the surface of the Ewald sphere and gives rise to diffraction spots with diffraction vectors $\mathbf{S}/\lambda$ on the surface of a cone. It is typical to rotate around several axes to produce a map of REL and deduce symmetry and structure from the symmetry relationships among the spots and relative intensities.

3.5.2 Powder Method

The powder method also uses monochromatic X rays, but in this method the sample is finely ground. Imagine that one starts with a single crystal, and as this is ground to a finer and finer powder, the crystals get smaller and smaller. In a pinch of fine powder one has a large number of possible crystal orientations (not really all the orientations, which would be ideal, but a huge number in a fine powder). Hence all these orientations of the REL of the material can be simultaneously painted by the incident X-ray beam of several mm in diameter without the need for rotation. This experimental situation,
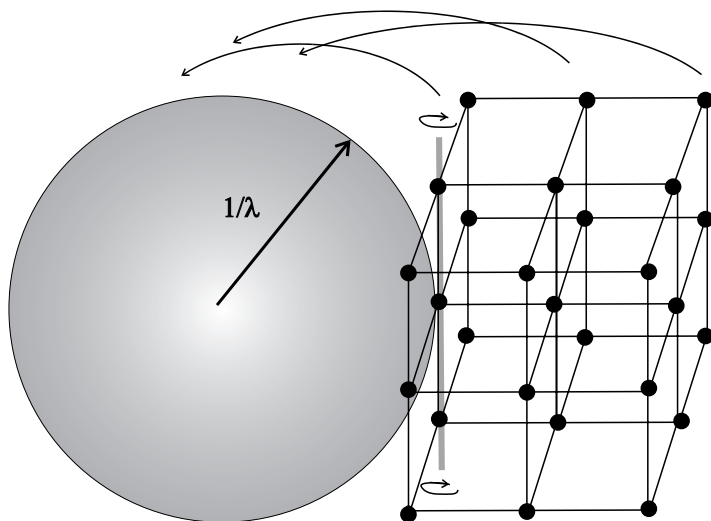
Figure 3.14 Rotating crystal method in which a materials reciprocal lattice (REL) is rotated so that REL points intersect the Ewald sphere (shaded) of radius $1/\lambda$.
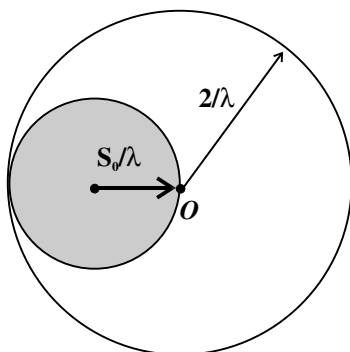
Figure 3.15 Powder method where an Ewald sphere (shaded) of radius $1/\lambda$ is rotated about O to produce diffraction from various orientations of crystals in the powder.

shown in Figure 3.15, can be thought of as having the REL fixed and S_0/λ rotating about the origin O. This way a new sphere of radius $2/\lambda$ is generated, called the limiting sphere, that contains all the possible spheres of reflection and thus defines all possible hkl's for λ to diffract. This method produces what are called cones of diffraction. The reason that it produces cones is that not only are all the Bragg angles θ produced by the finely ground crystallites in the incident emr beam at once, but different rotations about the incident beam ϕ are also present in crystallites at each θ. This results in a cone of diffraction. If a film strip is placed to intercept the cones and record the diffracted radiation, arcs, called Debye arcs, are seen on the film. Figure 3.1a shows a representation of a film strip used in one kind of powder diffraction technique with the Debye arcs. From the positions of the arcs, hkl indexes are obtained as well as d spacings, from which the identity of the
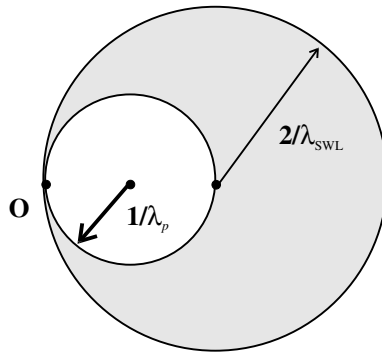
Figure 3.16 Laue method where radiation with wavelength from the shortest limit λ_{SWL} to the adsorption edge of the emulsion λ_p define those Ewald spheres in the shaded region that yield diffraction.

powder can be deduced by comparison with available libraries of powder patterns. The circles are, in reality, holes in the film through which an incident X-ray beam enters and leaves the apparatus.

3.5.3 Laue Method

In the Laue method "white" radiation (many λ's) is used with single crystal samples. This method is typically used to determine the orientation of a specific face of a crystal relative to the incident radiation. As we will see many times in materials science, many of the properties of materials are determined by crystallography. Thus it is important to know which crystallographic orientation on which single crystal property is determined.

Each of the many λ's that are used in the technique defines a sphere in which there exists a family of Ewald spheres having radii from $1/\lambda_p$, where λ_p is the wavelength that is absorbed by the photographic emulsion used to record the diffraction (the K edge for Ag assuming that a Ag photographic emulsion is used to record the diffraction) to $1/\lambda_{SWL}$ for the shortest wavelength in the X-ray spectrum (the short wavelength limit, SWL), and this situation is depicted in Figure 3.16. In essence, the spheres with radii in the shaded region represent potential diffraction spheres. The available wavelengths select planes from the REL with which to diffract for a given orientation of the REL at O. From the diffraction spot pattern, the orientation of the REL is deduced. This method is typically used to obtain the crystal orientation. For example, a Si crystal boule is pulled from the melt yielding a sausage-shaped Si single crystal. It is necessary to slice the boule into precisely oriented Si wafers for microelectronics processing. After a first cut is made, the Laue pattern from the flat boule surface is used to determine the precise sawing angle for all the slices. Figure 3.1*b* shows the results of the Laue method using a single crystal. Typically an experienced experimenter can accurately guess the crystal orientation simply from the symmetry of the REL on the film.

3.6 WAVE VECTOR REPRESENTATION

The wave vector, or **k** space, representation of reciprocal space is important for electron energy band structures (to be introduced in Chapter 9). However, the RESP development
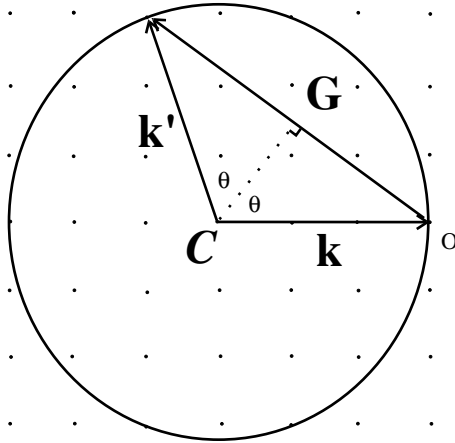
Figure 3.17 **k** space representation of RESP where **G**, **k**, and **k′** are analogous to **r***, **S**$_0$/λ, and **S**/λ, respectively.

presented above sets the stage for this representation. **k** space is obtained by simply expanding RESP by the factor 2π. To generate this space, we define a new vector, **k**, called the wave vector,

$$\mathbf{k} = \frac{2\pi}{\lambda} \qquad \left(\text{remember } \mathbf{S}, \mathbf{S}_0 = \frac{1}{\lambda}\right) \tag{3.54}$$

Next we define, for an orthogonal system,

$$\mathbf{a}^* = \frac{2\pi}{\mathbf{a}}, \quad \mathbf{b}^* = \frac{2\pi}{\mathbf{b}}, \quad \mathbf{c}^* = \frac{2\pi}{\mathbf{c}} \tag{3.55}$$

In RESP **r*** = *h***a*** + *k***b*** + *l***c***, but now with the REL's expanded by 2π, the new REL vector is labeled **G**, the diffraction vector in **k** space. The situation analogous to the Ewald diffraction sphere is shown in **k** space in Figure 3.17 where

$$\mathbf{k}' - \mathbf{k} = \mathbf{G} \tag{3.56}$$

$$\Delta\mathbf{k} = \mathbf{G} \tag{3.57}$$

Squaring both sides yields

$$\mathbf{k}'^2 = \mathbf{k}^2 + 2\mathbf{k}\mathbf{G} + \mathbf{G}^2 \tag{3.58}$$

The condition for diffraction is when $2\mathbf{k}\mathbf{G} + \mathbf{G}^2 = 0$, which means that **k** = **k′** and that **G** lies at a point of the REL. The solution of this quadratic is **k** = ±**G**/2. Thus the wave vector that bisects a REL vector **G** will be diffracted with **a*** = (2π/*a*)**x**, where **x** is a unit vector. Then, for a linear array of regularly spaced lattice points in the REL, the diffraction condition is given as
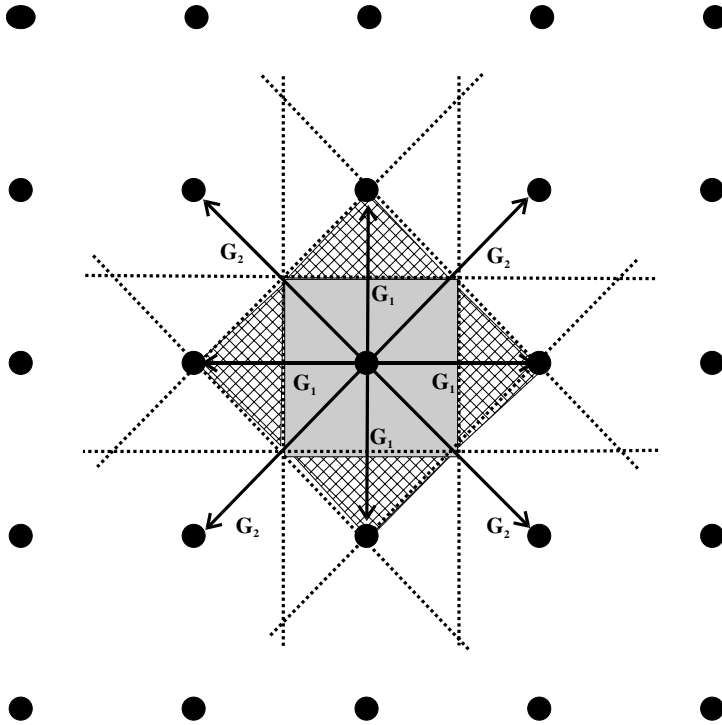
Figure 3.18 Brillouin zones in **k** space defined by the bisectors of the **G**$_i$ vectors.

$$\mathbf{G} = n\mathbf{a}^* = \left(\frac{2\pi}{a}\right)n\mathbf{x} \qquad (3.59)$$

$$\mathbf{k} = \pm\frac{n\pi}{\mathbf{a}}, \qquad n = 1, 2, \ldots \qquad (3.60)$$

Electrons propagating in a crystal lattice often have deBroglie wavelengths commensurate with the subnanometric and nanometric crystal dimensions. Thus some of the electron waves in a material with the appropriate wavelengths and directions in reciprocal space will diffract. The electrons of certain energy and direction that diffract yield intensities outside of the crystal, and it is as if these electrons are prevented from propagating. So there are energy regions and directions in which electrons can propagate, called allowed "energy bands" and energy regions where electrons cannot propagate. The electrons that cannot propagate are effectively diffracted out of the crystal, and the associated energies give rise to energy "band gaps." The band gap regions of electron energy are said to have no allowed energy states to and from which electrons can flow. (More will be said about allowed and disallowed electron energies in Chapter 9.) The **k** space representation can be used to represent these allowed zones, called Brillouin zones after the scientist who made this idea popular. The construction of Brillouin zones is similar to the construction of a Wigner-Seitz unit cell (see Figure 2.13), and it is

depicted in Figure 3.18. First a 2-D crystal lattice is represented by the 2-D REL points. An origin is chosen that is often, but not necessarily, a REL point. From this point **G** vectors are drawn first to the nearest neighbor planes and labeled G_1, and then to succeeding planes, or REL points, and labeled G_2, G_3, and so on. Now, from the discussion above, diffraction takes place at **G**/2. So we bisect $G_1/2$ and extend the bisectors so that they intersect and from an enclosed region (shaded region) around the origin. This region is called the first Brillouin zone, and an electron can propagate up to the region's edge whereupon it is diffracted. Notice that this region is also a primitive cell containing one lattice point in RESP. The first Brillouin zone in RESP is directly comparable to the Wigner-Seitz primitive cell in real space. Succeeding Brillouin zones are obtained by bisecting succeeding **G** vectors and considering the area enclosed less the area of the preceding Brillouin zone. Interesting shapes are obtained for 3-D RELs, and this important idea will be revisited when the subject of electron band structure is covered in Chapter 9.

RELATED READING

E. D. Cullity. 1956. *Elements of X-ray Diffraction.* Addison Wesley, Reading, MA. All editions of this book contain voluminous structure information in readable text form, and in appendixes the X-ray diffraction techniques are discussed at length.

D. McKie and C. McKie. 1986. *Essential of Crystallography.* Blackwell Scientific Publications, Cambridge, MA. A thorough treatment of crystallography and of diffraction and diffraction techniques.

J. M. Schultz. 1992. *Diffraction for Materials Scientists.* Prentice Hall, Englewood Cliffs, NJ. A more advanced treatment of diffraction theory and practice.

EXERCISES

1. For Al predict the diffraction angles (2θ) for the first three planes for $\lambda = 0.1542$ nm.

2. For FCC Cu with atomic radius of 0.2552 nm:
   (a) Calculate a_0
   (b) Given X rays of $\lambda = 0.152$ nm, and measured 2θ values for two Cu samples of 42°46′ and 41°55′. Which sample is pure?

3. From the following information determine whether the structure is BCC or FCC. X ray $\lambda = 0.171$ nm, 2θ values are 60°00′ and 70°34′.

4. Construct a 2-D real space lattice with an assumed a_0 and that shows five planes. Then construct the REL showing the same five planes and **r***. Discuss differences in directions and spacings in real and reciprocal spaces, and relate your response with Chapter 2, exercise 4.

5. You are assigned to build an X-ray monochromator to cover the energy range of about 10^5–10^4 eV. Design it, suggesting materials (justify) and geometry (sketch). Emphasize principles.

6. Prove that the (111) diffracts for the FCC but not the BCC.

7. For a simple or primitive cubic (PC), structure calculate the indexes for the 10 lowest index planes that diffract.

8. Calculate the first five lowest index planes that diffract for the BCC and FCC. Discuss why the planes are different for the PC, BCC, and FCC.

9. Calculate the five lowest index planes for the diamond cubic, DC, structure. This structure has 8 atoms per unit cell located at:
   0 0 0 1/2 1/2 0 1/2 0 1/2 0 1/2 1/2 1/4 1/4 1/4 3/4 3/4 1/4 3/4 1/4 3/4
   1/4 3/4 3/4
   and sketch the DC unit cell.

10. For the FCC metal with $a_0 = 0.51603$ nm (Ce):
    (a) Make list of the diffracting planes.
    (b) Calculate the diffraction angles for CuK_α radiation $\lambda = 0.15418$ nm.

11. Using a sketch of an Ewald construction for a 2-D square lattice in RESP and with a wavelength λ that touches at least one point of the REL, explain what experimental changes you would need to make to get another point that you choose to diffract.

12. Determine (calculate) whether the (111) diffracts in a monatomic BCC solid. Then show how this calculation would be modified if the solid were not monatomic.

13. Discuss what structural information you can and cannot obtain from Bragg's law.

14. From Chapter 2, Exercise 16, where you discussed the structural difference(s) between a compound and an alloy, now explain the differences you would expect from diffraction.

4

DEFECTS IN SOLIDS

4.1 INTRODUCTION

In the discussions in Chapters 2 and 3 the emphasis was on perfect crystalline solids. At the outset of Chapter 2 a brief and cursory comparison of crystalline and noncrystalline solids was presented. As the study of materials science proceeds throughout this text, more references to and discussions about both glassy and amorphous noncrystalline solids will be made. Noncrystalline solids offer literally a limitless variety of structures, and many of them are dependent on specific preparation conditions. On the other hand, crystalline solids, in particular, single crystals, are well described using the 14 Bravais lattices. Thus it is usually considered useful in the study of materials to commence with the structurally simplest materials—single crystals—and then proceed toward greater structural complexity. Before getting to noncrystalline materials, another idea is crucial to enable understanding of the structural nature of both real crystalline and noncrystalline materials. This idea is about defects, that defects occur in all real materials to some degree and can be evoked to explain the difference between crystalline and noncrystalline solids (where in the latter case the defect level is very high). In addition defects are inherent to many physical and chemical properties of solids. Among the important properties that are largely controlled by defects, and discussed in this text, are the following:

- Resistivity, ρ, through current carrier scattering from defects.
- Conductivity, σ, in semiconductors due to substitutional lattice defects.
- Deformation and strength due to dislocations.
- Nucleation of phases and site specific chemical reactions at surface defects.
- Diffusion via defect mechanisms.

There are many kinds of defects. One way to systematize the study of defects is by a categorization according to the dimension of the defect:

- 0-D Point defects: impurities, vacancies, interstitials.
- 1-D Line defects: dislocations.
- 2-D Planar defects: interfacial defects such as surfaces, grain boundaries, and phase boundaries.
- 3-D Bulk defects: voids, cracks, pores.

These kinds of defects are defined and discussed below with particular attention to 0-D point defects and 1-D line defects. But before we begin the discussion, we need to lay the important physical groundwork.

4.2   WHY DO DEFECTS FORM?

For any defect to form of the type mentioned above, a chemical bond needs to be broken. Thus, defects raise the free energy of a perfect crystal, and they may not form spontaneously. If one considers that under near equilibrium conditions the solid will attain the lowest energy configuration, that is, the equilibrium crystal structure for crystalline materials or a network structure for noncrystalline materials, then why do virtually all materials display a substantial number of defects, particularly point defects? Often for most solids the number of defects is large enough to measurably alter (raise) the enthalpy of the solid. Other observations that explain the nature of defects include the following: solids that display low defect concentrations are difficult to preserve, virtually any processing of perfect crystals will increase the number of defects, different kinds of defects are associated with different properties, under a given process some kinds of defects will increase while other will disappear. How are all these observations reconciled? A good approach to this question is to turn to the field of thermodynamics. So before proceeding to the heart of the defects issues, we will briefly review several relevant principles of thermodynamics. However, although the thermodynamics of the different kinds of defects can yield an understanding of defects in crystalline solids that can lead to a quantitative description of many important properties, the area of defects in solids remains an open and fertile area of materials research.

4.2.1   Review of Some Thermodynamics

4.2.1.1   First Law of Thermodynamics   Several relationships we develop here will be used later. We first consider the simple system of an ideal gas at constant temperature ($dT = 0$) that undergoes an isothermal expansion from V_1 to V_2, where $V_2 > V_1$. We can depend on the First Law for this situation, which asserts the conservation of energy. It is written in terms of the change of the internal energy E for a system:

$$dE = dq + dw = 0 \tag{4.1}$$

where q is heat into (or out of) the system and w is work done on or by the system. The change in work (done by the system) is

$$dw = -p_{ext}dV \tag{4.2}$$

for a constant external pressure, p_{ext}. Work is then the part of the energy E that is concerted to a task (one does work). Strictly speaking, work is defined in terms of force, F, and displacement dx as

$$w = \int \mathbf{F} \cdot dx \tag{4.3}$$

Equations (4.2) and (4.3) are concordant, since $p = \mathbf{F}/A$ with A the area. However, there is another component of E, namely the heat, q, that is not focused to perform a task. Heat is the component of the energy found in the random motion of the atoms/molecules in a material. So q is that part of E that cannot be fully harnessed to directly do work. A part of q is then irrecoverable. It should also be understood that the temperature T is an indicator of the direction of flow of heat. Heat flows from high to low temperature. The scale for T is arbitrary and chosen for convenience.

Now assuming the applicability of the ideal gas law

$$p = \frac{nRT}{V} \tag{4.4}$$

and using the differential form of First Law as written above, we obtain for a reversible process

$$dq_{rev} = \frac{nRTdV}{V} \tag{4.5}$$

Upon integration of equation (4.5) from state 1 to 2, the result to be used below is

$$\Delta q_{rev} = nRT \ln\left(\frac{V_2}{V_1}\right) \tag{4.6}$$

4.2.1.2 Second Law of Thermodynamics The Second Law deals with the availability of energy once a system absorbs (or emits) energy. It is observed that all the energy within a system is not later available. This idea was mentioned above in relation to the heat, q. The thermodynamic state function that quantifies the unavailable part of the energy is called the entropy, and is symbolized by S. The use of S and more importantly for a change ΔS or dS can yield information about the direction of reactions in which energy is exchanged. We commence with a few statements about ΔS:

$$\Delta S_{tot} = \Delta S_{sys} + \Delta S_{sur} > 0 \tag{4.7}$$

for a spontaneous process where the subscripts tot, sys, and sur refer to total, system, and surroundings, respectively. The infinitesimal change (as indicated using the subscript rev for reversible) in entropy for an isolated system can be obtained using

$$dS_{sys} = \frac{dq_{rev}}{T} \tag{4.8}$$

To obtain dS_{sur}, one needs to consider the specific process conditions. For example, at constant p, $\Delta H = q_P$. Thus dS_{tot} is written as

$$dS_{\text{tot}} = dS_{sys} - \left(\frac{dH}{T}\right) \tag{4.9}$$

where the $-dH$ indicates that enthalpy flows out of the system to the surroundings. The point of reference is taken in the system. Then this expression can be simplified to

$$TdS_{\text{tot}} = TdS_{sys} - dH \tag{4.10}$$

And now by defining $dG = -TdS_{\text{tot}}$ and removing the d's, we obtain the formula that defines G, the Gibbs free energy:

$$G = H - TS \tag{4.11}$$

It should be noted that this new thermodynamic state function called the Gibbs free energy is nothing more (or less) than the negative of the sum of the entropy for the system and surroundings at constant pressure. Since it is written in terms of H rather than q, it is quite useful because values for H are relatively available in Tables.

We need a useful expression for our present purposes. So we integrate equation (4.8), using (4.5), to obtain ΔS_{sys}:

$$\Delta S_{sys} = \int_1^2 \frac{dq_{\text{rev}}}{T} = nR\ln\left(\frac{V_2}{V_1}\right) \tag{4.12}$$

The change in Gibbs free energy, ΔG is often used to determine the spontaneity of a chemical process at constant pressure (i.e., in an open vessel on earth):

$$\Delta G = \Delta H - T\Delta S \tag{4.13}$$

where the enthalpy $H = E + pV$ is related to the internal energy and T is the absolute temperature. The Helmholtz free energy, $A = E - TS$, is used similarly for the less common constant volume processes. For ΔG, the process is spontaneous with ΔG being negative (or $T\Delta S$ positive), which can be obtained readily with negative ΔH and positive ΔS_{sys}.

4.2.1.3 *Notion of State*   We want to apply the thermodynamic reasoning developed above to the problem of defects, in particular, the entropy of defects. So it is important to have a clear idea of the meaning of a thermodynamic state. It is part of our experience that the expansion of an ideal gas (or any gas) from V_1 to V_2, as was discussed above, is spontaneous. That this occurs spontaneously is summarized by the condition that at $\Delta H = 0$, $\Delta S_{\text{tot}} > 0$ for the case of $V_2 > V_1$, as is given by equations (4.12) and (4.9).

We can visualize the expansion as two equal volume chambers A and B separated by a valve, as shown in Figure 4.1. When the valve is opened, the gas initially in chamber A will fill all the available volume (at a reduction in p). We let $V_1 = A$ and $V_2 = A + B$, and then V_1 will change to V_2. The reverse process of the gas particles returning into V_1 is not observed so long as the number of gas particles is large. While we have no difficulty in asserting our experience that this is indeed the case, we may have some difficulty in explaining why this is so.

Since each gas particle velocity vector is equally likely even in the reverse direction, it is obvious that only single gas particle is equally likely to be in A or B, if $V_A = V_B$. If V_B is larger than V_A, then we can intuitively conclude that our one particle of interest
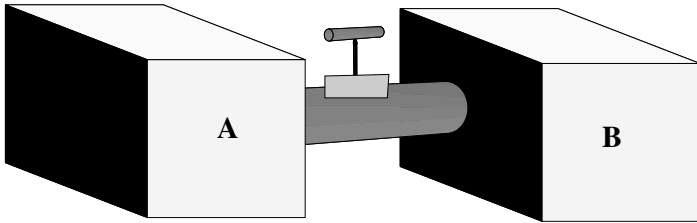
Figure 4.1 Two equal volumes *A* and *B* separated by a valve.

will more likely be found in the larger volume. However we cannot exclude it from being in V_A, especially if both volumes were close in size. Thus our experience and intuition indicate that the probability of finding the particle is related to the size of the container. One can express this experience more definitively in terms of allowed particle states. Simply, in the larger volume there are more available and allowed particle states. In our simple unrestricted example, we can consider an allowed state as the minimum volume necessary to contain a gas particle, and it is an available state if unoccupied. Obviously there are more of these volume states in the larger volume. Extending this idea of volume states to many particles, we consider that if there are more than one gas particle that we wish to follow then there is a greater probability for the particles being in the different volumes than in the same volume. This is so, because there is simply more volume available, since $V_2 > V_1$. Furthermore a particle residing in one volume lowers the probability for the second particle to be in that volume, because the presence of the first particle reduces the number of available states for the second particle. Then the number of states in V_2 is greater than the number in V_1. Thus there are more places (volume states) in V_2 for particles than in V_1. If the probability for each state as defined is equivalent, then the probability is higher for particles to occupy V_2 than V_1.

4.2.1.4 *Boltzmann Relationship*   It is important to remember that ΔS_{tot} is a measure of the chaotic component of the energy in a system and also a measure of the randomness of a system. When no heat can flow ($\Delta T = 0$) a system can still have randomness (relative to a previous state or to another system). Previously we considered ordered (crystalline) and relatively disordered pieces of material. Boltzmann recognized that a material has a configurational entropy or ordering regardless of the flow of heat. The Boltzmann relationship endeavors to quantify the randomness by relating thermodynamical and statistical variables and it can be written for a change as

$$\Delta S_{tot} = k \ln \Omega \tag{4.14}$$

In this equation Ω is the ratio of the probability of the final to initial states and given as

$$\Omega = \frac{w_f}{w_i} \tag{4.15}$$

where w_f is related to the number of ways of forming the final state and w_i is the number of ways of forming the initial state. S_{tot} refers to the total entropy, which is the sum of $S_{sys} + S_{sur}$. If the T were the same for the system and surroundings, then no q would flow

to the surroundings and $dS_{sur} = 0$. We take this isothermal case and apply the Boltzmann relationship to the system discussed above comprised of V_1 and V_2 and some number of gas particles. When the valve is opened the particles arrange themselves in the larger volume with more available states. There are more ways to arrange the particles in the V_2 state by virtue of the more available states. Then Ω is greater than unity, and ΔS_{tot} is positive ($\Delta S_{sys} = +$ and $\Delta S_{sur} = 0$). One can see from the Gibbs free energy relationship that for all other variables being the same, the higher entropy associated with the system with higher probability yields a lower free energy (ΔS appears as $-T\Delta S$), and at higher temperatures the entropy is even more effective at reducing ΔG.

We now have some tools with which to attempt to understand why some kinds of defects always exist.

4.3 POINT DEFECTS

Three different kinds of common point defects are shown in Figure 4.2: vacancies, interstitials, and substitionals. Vacancies and substitutional point defects refer to the normal lattice positions. The substitutional defect is a different atom on a lattice site normally occupied by a normal lattice atom. An interstitial point defect is an atom in the interstices in between normal lattice sites, and it can be a self-interstitial, namely the interstitial is a normal lattice atom displaced from its normal position, or it can be a foreign atom. These defects occur in numbers dictated by the amount of energy necessary to produce the defect. Below it will be argued that the configurational entropy attained by creating a defect lattice helps to offset the enthalpy of formation of the defects. The
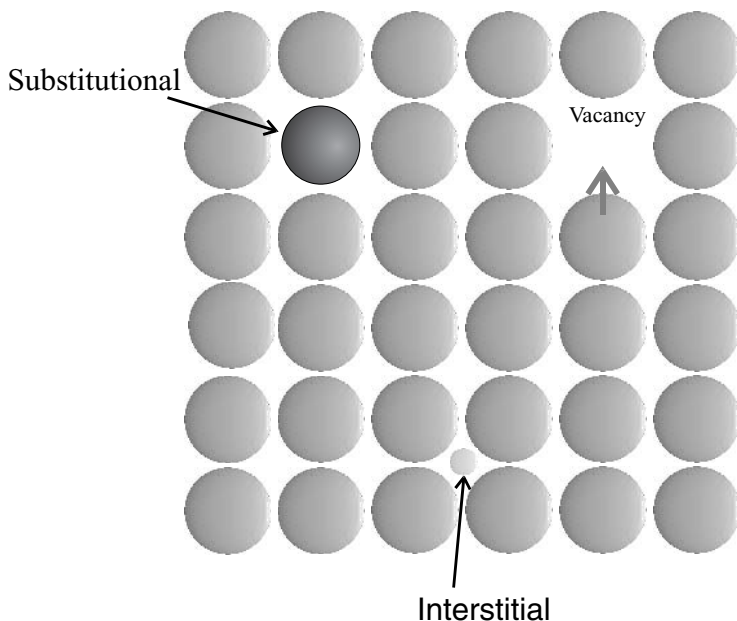
Figure 4.2 2-D square lattice displaying several point defects: Substitutional impurity, interstitial impurity, and vacancy. Vacancy diffusion is also shown (*arrow*).

enthalpy of formation for an interstitial often is a strong function of the size and charge of the interstitial to be formed. There is also the possibility of forming charged point defects in ionic crystals, and these are known as Frenkel and Schottky defects. The Frenkel defect is an interstitial of a charged atom that, when formed, creates two regions of different polarity. It is often referred to as an interstitial pair defect. The Schottky defect is also a pair defect, but it is the absence of both ions. Overall charge neutrality must be maintained for the formation of charged defects. Often the motion of atoms in a solid takes place by atoms migrating into vacancies or interstices that will leave a vacant position behind for another atom to migrate into. Self-diffusion via vacancies is shown in Figure 4.2 with the arrow to indicate the possible direction of migration of an atom that leaves behind a vacancy. The atom moves one way and the vacancy the opposite. The diffusion via point defects will be treated again in Chapter 5 on diffusion.

4.4 THE STATISTICS OF POINT DEFECTS

We compare the entropy for a crystal that has some number of defect lattice sites (either interstitials or vacancies) to that for a perfect single crystal using the tools developed above. Interstitials are simply an atom residing in between lattice sites such as in the tetrahedral or octahedral interstices discussed earlier, and vacancies are simply atoms missing from lattice sites. For example, in order to form a self-interstitial or a vacancy from a perfect lattice, some bonds need to be broken and possibly rearrangement occurs. Thus energy is required for each defect added to the lattice. Furthermore there is only one possible configuration for a perfect single crystal, and for a simple monatomic solid it is with all the lattice points occupied or associated with atoms. Therefore $w_i = 1$ for a perfect crystal as the initial condition. Then from equation (4.15) $\Omega = w_f$ and $\Delta S = k \ln w_f$ from equation (4.14). Now we need to calculate w_f for a material that has some number of defects present. It is clear that the number of defects will determine the number of ways the defects can be arranged on the lattice. Also ΔS will increase with the number of ways to arrange the system. The influence of S on the resulting free energy is offset or counterbalanced by the energy required to create the defects. For interstitials and vacancies, respectively, we can write

$$\Delta E = n_i \cdot \Delta E_i \quad \text{and} \quad \Delta E = n_v \cdot \Delta E_v \tag{4.16}$$

where the n's are the numbers of the defects and the E's are the creation/formation energies per defect. Because of the small volume changes associated with solid materials E and H are often used interchangeably for solids. Thus equation (4.16) teaches that ΔE or ΔH increases linearly with the number of defects. In this case where there is energy required to form the defects, this energy needs to flow to the system and thus $\Delta S_{sur} \neq 0$. Then we can use the expression for ΔG, which sums the surroundings and system entropies. We are now ready to calculate ΔS for any number of defects. First we calculate w_f, then Ω and ΔS, and finally the number of point defects at any temperature.

We consider that the defect is a vacancy of a lattice site, although the same kind of argument is appropriate for any type of lattice defect. For one vacancy on a lattice of N sites: $w_f = N$. For the next vacancy: $w_f = N - 1$, because only $N - 1$ sites remain available. Now to generalize, we can write

$w_f = N$ for the first, $w_f = N - 1$ for the second, $w_f = N - 2$ for the third, etc.

We also include the assumption that all the sites and all the vacancies are respectively indistinguishable. This gives rise to a denominator 1, 2, 3, to account for the indistinguishable defects and yields for w_f:

$$w_f = \frac{N(N-1)(N-2)\dots}{1 \cdot 2 \cdot 3 \dots} \tag{4.17}$$

We proceed to generalize for n_v vacancies. The $N-1$ term for the second vacancy is formulated by subtracting from the number of lattice sites N the number of vacancies remaining, $(n_v - 1)$ or $(2 - 1)$, to be added to the lattice; that is, $N - 1$ as obtained from $N - (n_v - 1)$, yielding $N - (2 - 1)$. This result can be generalized as

$$w_f = \frac{N(N-1)(N-2)(N-3)\dots(N-n_v+1)}{1 \cdot 3 \cdot 4 \dots n_v} \tag{4.18}$$

For example, for $n_v = 5$,

$$w_f = \frac{N(N-1)(N-2)(N-3)(N-4)}{5!} \tag{4.19}$$

where $N-4$ comes from $N-(5-1)$ with $n_v = 5$. The numerator does not go all the way to $N!$. It stops, for n_v at $N - n_v + 1$. To force the numerator to be $N!$, we multiply both numerator and denominator by $(N - n_v)!$ This yields

$$w_f = \frac{N!}{n_v!(N-n_v)!} \tag{4.20}$$

Now with $\Delta S = k \ln(w_f/1)$ we obtain

$$\Delta G = \Delta E - kT \ln\left(\frac{w_f}{1}\right) = n_v \Delta E_v - kT \ln\left(\frac{N!}{n_v!(N-n_v)!}\right) \tag{4.21}$$

Equilibrium is when ΔG is a minimum or $\partial(\Delta G)/\partial n_v = 0$:

$$\frac{\partial(\Delta G)}{\partial n_v} = 0 = \Delta E_v - kT \frac{\partial(\ln N! - \ln n_v! - \ln(N-n_v)!)}{\partial n_v} \tag{4.22}$$

We proceed to evaluate the derivative of each term on the right-hand side of this expression.

First term:

$$\frac{\partial(\ln N!)}{\partial n_v} = 0 \tag{4.23}$$

since N is not a function of n_v.

Second term:

$$\frac{\partial \ln n_v!}{\partial n_v} = \ln n_v \quad \text{from the relationship} \quad \frac{d \ln X!}{dX} \approx \ln X \tag{4.24}$$

Third term:

$$\frac{\partial \ln(N - n_v)!}{\partial n_v} = \frac{\partial[(N - n_v)\ln(N - n_v) - (N - n_v)]}{\partial n_v} \tag{4.25}$$

which is obtained using Sterling's formula:

$$\ln X! = X \ln X - X \tag{4.26}$$

Upon differentiation the first term on the right side yields two terms:

$$\frac{\partial[(N - n_v)\ln(N - n_v)]}{\partial n_v} = -\frac{(N - n_v)}{(N - n_v)} - \ln(N - n_v) \tag{4.27}$$

For the second term on the right of (4.26) it is

$$\frac{\partial[-(N - n_v)]}{\partial n_v} = 1 \tag{4.28}$$

Upon combining the results we obtain

$$-\frac{N - n_v}{N - n_v} - \ln(N - n_v) + 1 + \ln n_v = -\frac{\Delta E_v}{kT} \tag{4.29}$$

Rearranging and simplifying yields

$$\ln n_v - \ln(N - n_v) + 1 - 1 = -\frac{\Delta E_v}{kT} \tag{4.30}$$

It further yields

$$\ln\left[\frac{n_v}{N - n_v}\right] = -\frac{\Delta E_v}{kT}$$
$$\text{and for } N > n_v \quad \frac{n_v}{N} = e^{-(\Delta E_v/kT)} \tag{4.31}$$

where n_v/N is the concentration of vacancies (i.e., the number of vacancies n_v divided by N, the total number of sites). A similar development for interstitials would yield a similar final result where only the subscripts would be different. Thus it is concluded that for any (positive) value for the energy required to produce the point defect, higher temperatures will produce more point defects. This is attributable to the greater stability of the system derived from the TS term in the free energy equation (4.13).

The derivation above is specific to the case of creating defects and how the system configuration or the configurational entropy impacts the calculation of the number of defects existent at any T. This notion of point defect creation can also be readily envisioned as a so-called two-state problem. We consider two allowed states, 1 and 2, and that the change in the occupation of one of the states, δn, is proportional to the energy differences between the states, δE. In particular, the more energy it takes to populate the higher energy (state 2), the smaller is the chance for the change of state to occur, then we use the negative sign, $-\delta E$, and we can write this situation for the relative population of a state as

$$\frac{\delta n}{n} \propto -\delta E \tag{4.32}$$

If we convert the δ to d and integrate the resulting differerntial equation from state 1 to state 2 using a constant C to change the proportionality to an equality, we have

$$\int_1^2 \frac{dn}{n} = -C \int_1^2 dE \tag{4.33}$$

We obtain the result as follows:

$$\ln n_2 - \ln n_1 = -C\Delta E$$

and with $C = 1/kT$,

$$\frac{n_1}{n_2} = e^{-\Delta E/kT} \tag{4.34}$$

This is essentially the same result as equation (4.32) for the creation of vacancies. The process is called a temperature-activated process in which n_2 is the number of lattice sites that, when vacated, produce n_1 vacancies. A vacancy in this material requiring ΔE_v for production from an occupied site yields

$$\frac{n_v}{N} = A e^{-\Delta E_v/kT} \tag{4.35}$$

where A is representative of a specific application.

Equation (4.35) is a form reminiscent of the so-called Arrhenius activation energy that is used to obtain or characterize rate processes. Now we see that this exponential form for energy is the result of considering the statistics of populating various states that require different energies. The exponential expression $e^{-E/kT}$ is commonplace in the physical sciences and is often referred to as the Boltzmann factor.

Let us turn to the implications of the calculations made above. For metals the enthalpy necessary to create a vacancy is 1 to several eV. From equation (4.31) it can be demonstrated that at any T there exists a finite number of vacancies in any crystal. This means that the vacancies form spontaneously to minimize the free energy of the system. This situation is represented in Figure 4.3, where it is shown that as the number of vacancies increases, the energy associated with the bond breaking increases linearly. In contrast,
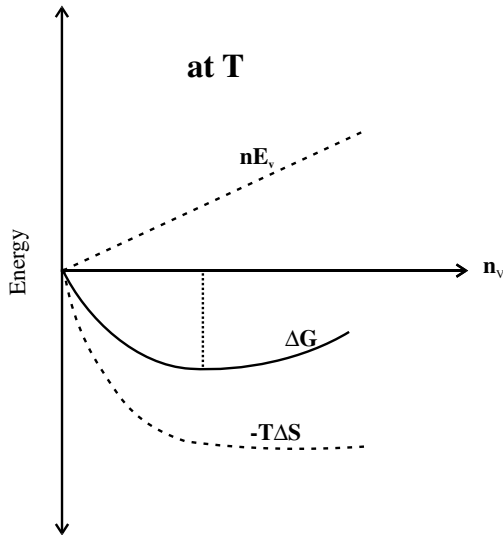
Figure 4.3 Thermodynamic function (*E*, *G*, *S*) variation with the number of vacancies at a constant temperature.

the energy component associated with the entropy decreases rapidly. Since the free energy change ΔG is the sum of these two components, a point of inflection must occur. At this point the equilibrium number of vacancies at the specified temperature is obtained (the dotted line).

4.5   LINE DEFECTS—DISLOCATIONS

Crystalline solids that have been subjected to tensile forces often display groups of parallel lines, called slip lines. The slip lines occur as the applied tension exceeds a threshold for the material, but before fracture. Figure 4.4*a* shows a tensile force $\mathbf{F}_T$ applied perpendicularly to the face of a single crystal sample. The tensile force's direction stretches the material parallel to the direction of the applied force, and this results in the elongated sample shown in Figure 4.4*b*. Figure 4.4*b* includes a close-up sketch of slip lines, or bands. These macroscopic lines have a microscopic origin, which the diagram of Figure 4.4*b* helps elucidate. Note that the applied force is resolved into force components that exist on any plane in the crystal. It is usual to resolve the applied force into normal and parallel force components on a specific plane.

With $\mathbf{F}_T$ as the applied force component normal to the top plane of the sample, and $\mathbf{F}_P$ the force resolved on the planes parallel to the slip bands, $\mathbf{F}_P$ is written as

$$\mathbf{F}_P = \mathbf{F}_T \cos\theta \tag{4.36}$$

The angle θ between the normal to the top plane and the normal to the slip plane as is shown in the figure. Thus $\mathbf{F}_T$ and $\mathbf{F}_P$ are both tensile force components, with the former perpendicular to the top plane and the latter to the slip plane. It is also possible to resolve
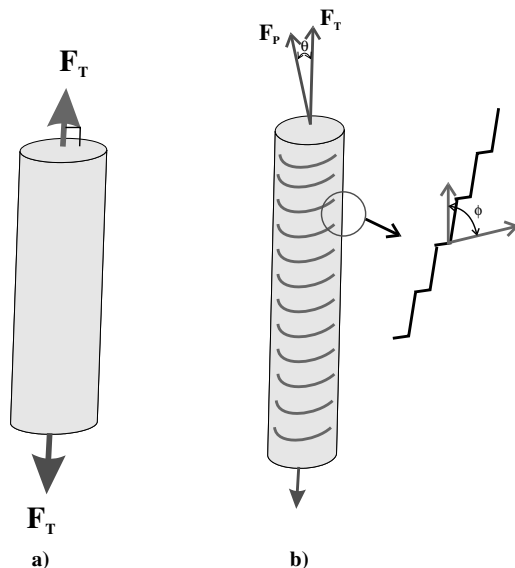
Figure 4.4 (a) Application of tensile forces to a crystal; (b) rise of slip bands due to the applied forces, which can be resolved onto specific planes and directions.

the applied force $\mathbf{F}_T$ in the slip plane (rather than perpendicular to it) in any direction in the slip plane. This component of force is called the shear force component $\mathbf{F}_S$ and is given as

$$\mathbf{F}_S = \mathbf{F}_T \cos\phi \qquad (4.37)$$

with ϕ being the angle between $\mathbf{F}_T$ and a particular direction in the plane. $\mathbf{F}_S$, the shear component of the applied force, is the force that appears to be a likely candidate to give rise to the slip lines, since this force exists in the direction of the planar displacement, namely the slip. Actually it is more appropriate to discuss slip in terms of the shear stress, τ, which is the resolved shear force, $\mathbf{F}_S$, per unit area on a plane. In Chapter 7, τ will be shown to be a maximum when both θ and ϕ are 45° to the applied force. However, slip lines are not always found only at 45° to the applied forces. This strongly suggests that the specific crystallography influences the slip line formation. In fact it was discovered that certain planes, called slip planes, and certain directions in the planes, called slip directions, were almost always implicated in the slip phenomena. The combination of slip plane and direction is called a slip system, and this combination is specific within the crystal structures. Table 8.1 displays the slip systems for cubic structures that will be used later for elucidating mechanical properties. The slip direction is found to have close packing in the slip plane, which is a plane already of high (usually highest) atom density.

From these facts a picture of the physics of slip emerges. Consider a 3-D lattice for a monatomic solid where each atom is held to the next by chemical bonds that are represented mechanically by springs (see a 2-D representation in Figure 7.2 where mechanical properties are discussed). Further we imagine that the springs with the highest

tension are those to the closest neighbors (representing the strongest bonds). With the application of an external force to the solid, the springs will distend. With more springs of highest tension on the planes with more atoms and in the directions of densest packing, it is these planes and directions that first "feel" or bear the distension. Another way to view this is that the shortest tightest springs will bear the applied force before the longer and looser ones do so.

When a sufficiently large shear force exists on a slip plane, one plane will yield in the slip direction relative to the other. The atomistic result of this motion, and the resulting deformation, is called a dislocation. The result of motion of one plane relative to another yields a dislocation of less than an atomic spacing in the direction of the relative motion. When exactly a full lattice displacement is achieved, registry will again be achieved across the dislocation line. Thus the maximum disregistry is at a half planar spacing. When a large number of planes participate all with fractional lattice spacing deformation, then the sum is a macroscopic displacement that results in the step-like total deformation shown in Figure 4.4.

It is observed (and we will discuss later in Chapters 7 and 8) that the actual shear force that is required for the dislocation type of deformation is about $G/1000$, where G is the bulk modulus (defined and discussed in Chapter 7). Even without proper background definitions, it is interesting to compare this measured value of force of around $G/1000$ to create a dislocation with the theoretical forces required for deformation of a material without defects, as these are of the order of $G/30$ or a factor of about 30 larger. The difference lies in the presence, multiplication, and movement of dislocations in a crystalline solid. We reserve this discussion to Chapter 8, which is devoted to mechanical properties. Here we define more precisely what a dislocation is and the kinds of dislocations. However, before we proceed to that subject, we should reflect on the important fact that it is only by the input of considerable energy or work (force through a displacement, see equation 4.3) that dislocations are formed. This is in contrast to the point defects, such as vacancies, that form spontaneously. It is concluded that the energy input required to form a line defect is 5 to 10 times larger and the configurational entropy gained with the formation of line defects is small. Thus line defects are less likely to form spontaneously.

4.5.1   Edge Dislocations

Imagine a block of crystalline material, like that in shown in Figure 4.5a, that is cut halfway with a knife having a cutting direction normal to one face of the solid. The cut is made in between two planes of atoms, thereby slicing through the bonds. One-half of the sliced cube is held in place (the left side in Figure 4.5a), and the other side (the right side) is compressed, as shown by the arrow, keeping the rear edges together. At some point in the compression of the right half an extra plane is created. The additional plane in the compressed half will appear as a displacement of one-half the atomic spacing in the direction of the applied compressive force. With the top and bottom halves reattached (i.e., the bonds made across the cut or shear plane), a maximum displacement of half an atomic spacing occurs, and to either side there is less displacement laterally where the distorted bonds relax. Figure 4.5b shows a view of this result normal to the direction of the displacement (i.e., on edge) and with the compression symmetric from both sides. The result is the same as if an extra plane of atoms were inserted into (only) the bottom half of the crystal. This kind of dislocation is called an edge dislocation and it is defined by the fact that the dislocation line (the black filled circle at the center of the ⊥, the
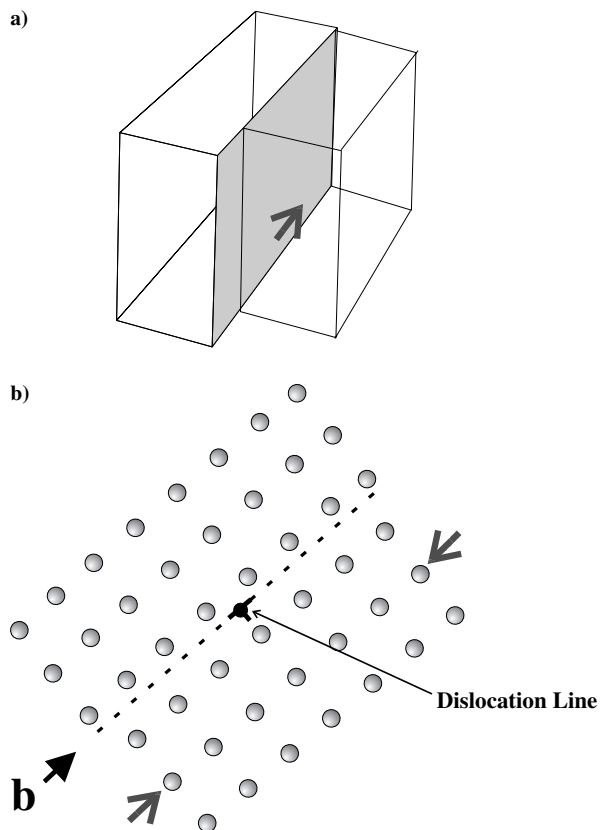
Figure 4.5   (*a*) Compressive force (*arrow*) exerted on part of a crystal leading to displacement of planes; (*b*) formation of an edge dislocation. The dislocation line is normal to the direction of the force, and the Burgers vector **b** indicates the magnitude and direction of the dislocation.

symbol to indicate an edge dislocation) is perpendicular to the direction of the shear. The dislocation line runs in and out of the plane of the paper under the extra half plane. This kind of dislocation is symbolized by: ⊥ or ⊤, for the extra half plane in the bottom or top part of the crystal, respectively. It is easy to imagine that if somehow ⊥ were to approach ⊤, the dislocations in the top and bottom of the crystal would annihilate, and perfect registry would be restored.

4.5.2   Screw Dislocations

To form the edge dislocation, we compressed the right half of the crystal relative to the left in Figure 4.5*a*. The forces were applied normal to the dislocation line. If we, instead, had simply laterally pushed the top half of a cut crystal (as was cut in the case above) to one side as is shown in Figure 4.6*a*, the result would be called a screw dislocation. This type of dislocation results when the shearing force is applied in the same direction (parallel) as the dislocation line. The outer part of the solid is deformed more than the

a)
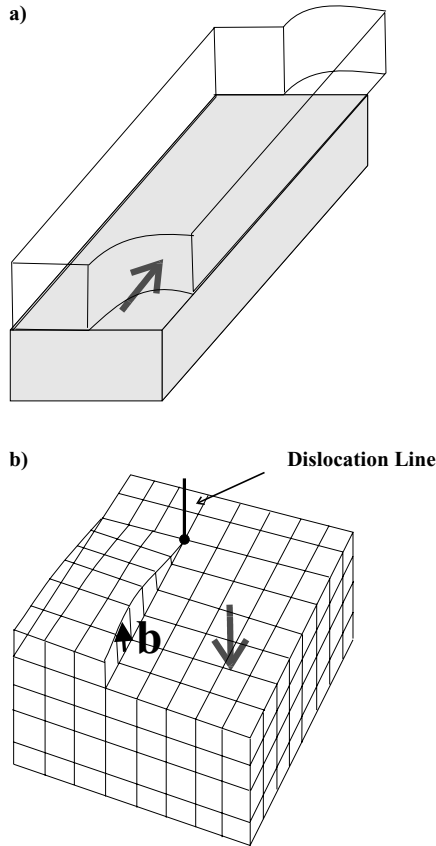
b)   **Dislocation Line**

Figure 4.6   (*a*) Shear force (*arrow*) applied to part of a crystal and resulting displacement of planes paral-
lel to the force, called a screw dislocation; (*b*) displacement intersecting a surface with Burgers vector **b** par-
allel to the dislocation line.

dislocation core. A spiral appears around the part of the crystal that is not cut nor
deformed, as in the view shown in Figure 4.6*b* (also displayed in larger view in Figure
8.4*a*). In this figure the dislocation line is normal to the crystal surface but parallel to
the direction of the applied force that caused the dislocation.

The pure edge or screw dislocations are then the result of specific directional forces
applied relative to the slip system in a given lattice. It is unlikely that such forces will
occur precisely in such a way to produce only an edge or only a screw dislocation, except
in our imaginations. Thus, in reality, a combination of edge and screw characteristics will
be displayed in most dislocations as the dislocation line snakes through the deformed
solid. However, a short segment of a dislocation might be accurately described as a pure
edge or screw type dislocation.

What is now needed is a quantitative descriptor for dislocations. This descriptor is
called the Burgers vector because it describes both the magnitude and direction of a dis-
location line.

4.5.3 Burger's Vector and the Burger Circuit

For both edge and screw type dislocations, the dislocation is defined quantitatively by a vector **b** called "Burger's" vector. As a vector it expresses both the magnitude and direction of the dislocation. Note that it is labeled in both Figures 4.5 and 4.6. (In Chapter 8 **b** will be used to calculate the energy of a dislocation.) The magnitude and direction of **b** is obtained from a Burger circuit, which is now described.

Figure 4.7 illustrates the Burger circuit for both edge (Figure 4.7*a*) and screw dislocations (Figure 4.7*b*). To form a Burger circuit, one first selects a starting place in the undeformed region of the crystal (labeled *S* in Figure 4.7). Then one proceeds from lattice position to lattice position around the dislocation in a clockwise manner, as indicated by the arrow. The vertical and horizontal lattice positions necessary to go halfway around the dislocation are separately tabulated. From the halfway position a return to the starting position is attempted using the tabulated number of vertical and horizontal jumps to get to the halfway point. If a dislocation exists within the circuit, it will not be possible to return to the starting position using these tabulated steps. This is, of course, due
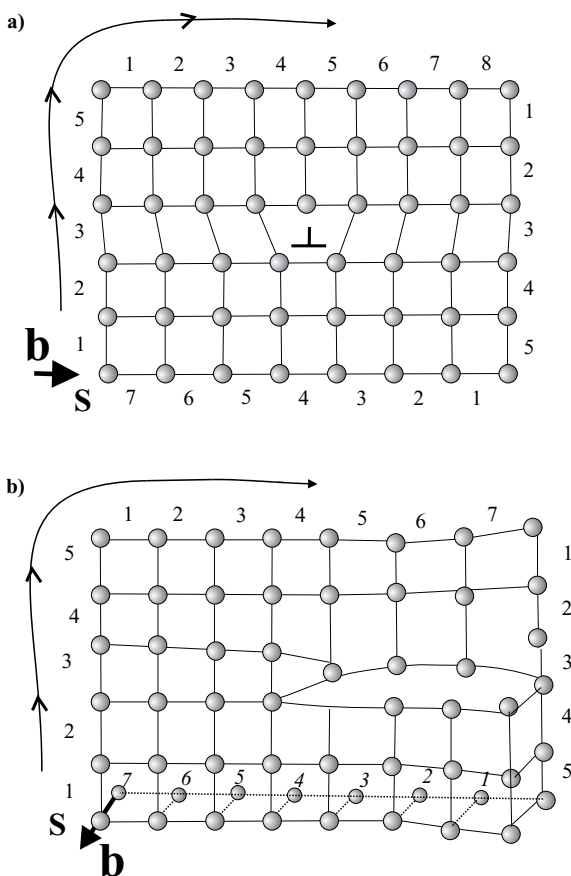
Figure 4.7 Burger's circuits for (*a*) edge dislocation and (*b*) screw dislocation. The start point *S* for the circuit, the clockwise direction, and the Burger vectors **b** are shown.

to the existence of a dislocation, which is a displacement. Nevertheless, the tabulated jumps are executed to leave a displacement from the starting position. Then the last piece of the circuit or the difference necessary to return to the starting point is noted, and this segment is **b** in both magnitude and direction.

In Figure 4.7*a* for an edge dislocation we arbitrarily start in the lower left side of the crystal. We can count an arbitrary number of lattice positions, say, 5 toward the top and 8 across to the right in a clockwise manner. The number of lattice positions is determined so as to encircle the dislocation. Having accomplished this and being halfway around the edge dislocation, we proceed to return by 5 jumps downward from the halfway point and then 8 back to the left to complete the circuit. However, 8 jumps would pass the start position. Then we form a vector **b** pointing from 8 to 7 to complete the circuit. The Burger vector is this last segment necessary to complete the circuit back to *S*.

In Figure 4.7*b* we proceed in the same manner for a screw dislocation. We commence the Burger circuit in the lower left and proceed clockwise upward, first 5 jumps and then 7 to the right to the halfway point. Then we return with 5 jumps down and then 7 to the left. However, these final 7 to the left are on a plane behind the plane on which we started. To indicate this on the figure, italics are used for the jump numbers. The seventh jump to the left would be one position directly behind the start position *S*. So, to complete the circuit, we need a jump perpendicular to the plane of the paper to the start position, and this vector forms **b** for the screw dislocation depicted.

4.5.4   Dislocation Motion

The motion of a line defect or dislocation is an important concept in defining the mechanical properties of materials, in particular, plasticity. (This context will be discussed in Chapter 8.) Here we consider the low-resistance (low-energy) path by which dislocations are observed to move. We might imagine the simultaneous bond breaking of all bonds around the extra half-plane that forms an edge dislocation. Then the extra plane can be removed and transported to the edge of the crystal and reattached. This is tantamount to major surgery involving simultaneous multiple-bond breaking. It is not, however, Nature's path of choice. An edge dislocation can actually move relatively effortlessly in a lattice (as in plastic deformation, which is discussed in Chapter 8, Figure 8.2). The atoms need only be displaced a small amount from their equilibrium position as a shear stress τ is applied. This way atoms on either side of the shear plane first go out of alignment and then into alignment with atoms that were not in alignment prior to the deformation. Only a small number of atoms then needs bonding rearrangements, and the extra half-plane moves one lattice position at a time as the stress continues. Finally the extra half-plane moves to the surface, where it results in an atomic step. The motion of many dislocations in that plane will increase the size of the step.

4.6   PLANAR DEFECTS

4.6.1   Grain Boundaries

In previous chapters crystalline, polycrystalline, and amorphous materials were discussed. Polycrystalline materials are made up of single crystals that are bonded together and have little or no crystallographic relationship to one another. The bonding region is called a grain boundary. Grain boundaries are classified according to the magnitude of the misorientation that occurs at the boundaries. Figure 4.8 compares a perfect crystal
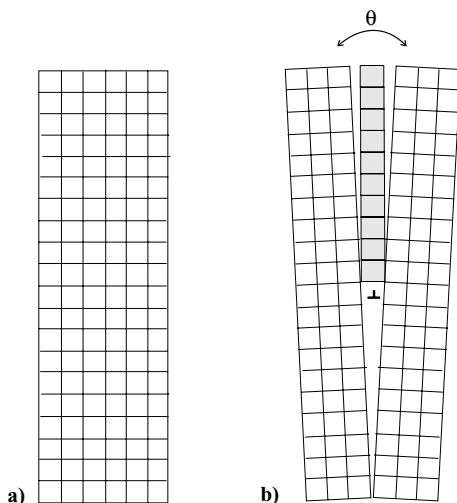
Figure 4.8 (*a*) A perfect crystal and (*b*) a low-angle grain boundary that resembles an array of edge dislocations.

(atoms are represented as squares) with one that has a low-angle grain boundary. Imagine that Figure 4.8*a* is a perfect crystal with atoms at each square. All the atoms are in registry in the perfect crystal. In Figure 4.8*b* the left half of the crystal is tilted relative to the right half, and to bond the halves together, another row of atoms is necessary; it is depicted as shaded squares. Each side is called a single crystal, since each is a perfect albeit small crystal. The angle θ is a measure of the misregistry with small angles resulting in "low-angle grain boundaries" and large angles resulting in what are called "high-angle grain boundaries." For low-angle grain boundaries as shown in Figure 4.8*b*, the extra row of atoms results in an edge dislocation. One can imagine the grain boundary extended over many more atoms than shown in the illustration and at the grain boundary an array of edge dislocations. The energy of the misorientation varies with θ. It always costs energy to produce a surface, and the production of small-grained material with large surface-to-volume ratios is not as stable as larger grained materials. Therefore the tendency is for the grain size to increase when the circumstances permit. One simple way to achieve grain growth, and hence reduce the grain boundary area, is to anneal at elevated temperatures. The high temperature enhances atomic mobility (as will be discussed in Chapter 5), which enables atoms to "find" and settle into lower energy configurations. It is also possible to have large-angle grain boundaries where the misregistry angle θ is comparatively large. In such cases the grain boundary region is more complex and cannot be characterized as simply as an array of edge dislocations. In fact discontinuities can arise.

4.6.2 Twin Boundaries

A twin boundary (sometimes called an interface boundary) is where in a given single crystal (often in a grain of a polycrystalline aggregate) one part of the crystal is a mirror
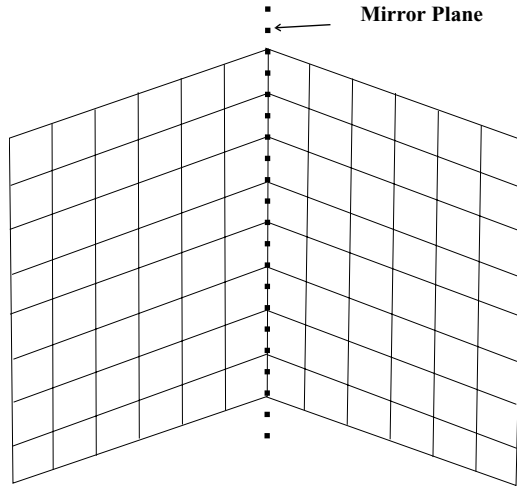
Figure 4.9   A twin boundary with the mirror symmetry plane shown.

image of the other adjacent part. Figure 4.9 shows the twin plane or boundary as a mirror plane. These defects are often the result of applied stress to an already formed grain structure.

4.7   THREE-DIMENSIONAL DEFECTS

Among 3-D defects are macroscopic irregularities such as voided pockets, cracks, and pore structure whose individual natures are evident from the names. Virtually all real materials have these kinds of defects, and some reference to each will be given throughout this text as appropriate. The fact that little space is devoted to 3-D defects should not diminish their importance. For example, cracks and crack propagation in glasses dominate the fracture mechanism. Pores dominate materials transport in many materials. Voids can dominate strength. Indeed, because the definitions are almost self-evident and the structures complex, few unifying generalizations can be made.

RELATED READING

C. R. Barrett, W. D. Nix, and A. S. Tetelman. 1973. *The Principles of Engineering Materials*. Prentice Hall, Englewood Cliffs, NJ. A readable elementary text for a first course in materials science.

D. Hull. 1975. *Introduction to Dislocations*. Pergamon Press, New York. A readable and well-illustrated treatment of dislocations.

P. A. Thornton and V. J. Colangelo. 1985. *Fundamental of Engineering Materials*. Prentice Hall, Englewood Cliffs, NJ. A readable elementary text for a first course in materials science.

EXERCISES

1. The room temperature density for Cu an FCC metal is given as 8.94 g/cm³. At 1000°C the density was found to be 8.92 g/cm³. Calculate how many sites are vacant at 1000°C assuming that the density changes are due to vacancies.

2. FCC Al has a_0 = 0.4050 nm. The measured density is 2.698 g/cm³. Calculate the number of vacancies in the Al.

3. A metal has an enthalpy for formation for a vacancy of 2 eV. Calculate the number of vacancies that can be quenched at 500°C, and then calculate the temperature increase necessary to increase the number of vacancies by a factor of 10 from the 500°C value.

4. Show with sketches what happens when an edge dislocation in the top half of a crystal meets an edge dislocation in the bottom half: $\top + \bot$.

5. Sketch the relationship of the applied force and the direction of the Burgers vector for both edge and screw dislocations in a lattice.

6. Discuss why point defects form spontaneously but line defects do not.

DIFFUSION IN SOLIDS

5.1 INTRODUCTION TO DIFFUSION EQUATIONS

Diffusion problems and diffusion-like problems pervade science and engineering. Basically these problems consider a flux or flow in response to a spatial gradient called a force, namely the driving force for the flow. Some examples in terms of a one dimensional gradient are

$$J_h = \frac{-\kappa dT}{dx} \qquad \text{Fourier's law of heat flow}$$

$$J_e = \frac{-\sigma dV}{dx} \qquad \text{Ohm's law}$$

$$J_m = \frac{-cdP}{dx} \qquad \text{Poiseuille's law} \tag{5.1}$$

$$J_D = \frac{-DdC}{dx} \qquad \text{Fick's first law}$$

Fourier's law is for a flow of heat J_h in response to (and proportional to) the temperature gradient dT/dx with a heat transfer coefficient κ being the constant of proportionality between the flux and driving force. Likewise Ohm's law is for a flow of electrons J_e (an electric current flux) that is proportional to the potential gradient dV/dx with σ being the conductivity. The next example is Poiseuille flow, with J_m being a mass flux in response to a pressure gradient. The subject of this chapter is another mass flux called diffusion with its flux, J_D, that is proportional to a concentration gradient. In these equations the negative sign indicates a flux down the gradient, and of course, all the fluxes

can be written in 3-D. All of these laws have much in common in that they ignore atomic structure and assume a continuum of matter, and all are intuitive, and therefore agree with experience. For example, we expect an electrical current to flow in response to a potential applied to a conductor, and for want of more information, a reasonable guess is that the relationship is linear. Often such laws that correspond to experience are termed "phenomenological" or "thermodynamic" laws. Also, like the laws of thermodynamics, these laws apply to large numbers of atoms/molecules or other objects, and thus like thermodynamics are understood on a statistical basis. That is, the laws describe statistical behavior of a large number of objects.

These laws and the phenomena governed can operate individually and/or simultaneously. In the latter case there exists coupling. For example, if there exists a spatial temperature gradient, dT/dx, not only will heat flow but mass may flow due to a "thermal migration" cross term. Possibly with mass and heat flows, additional species can migrate due to gradients other than the chemical gradient of the species in question. A theoretical understanding of simultaneous fluxes was developed by Onsager commencing with the principle of microscopic reversibility. The flux of the ith species can be given as the sum of all the contributions to this flow:

$$J_i = \sum L_{ik} X_k \tag{5.2}$$

For several fluxes we can write

$$\begin{aligned} J_1 &= L_{11}X_1 + L_{12}X_2 + L_{13}X_3 \\ J_2 &= L_{21}X_1 + L_{22}X_2 + L_{23}X_3 \\ J_3 &= L_{31}X_1 + L_{32}X_2 + L_{33}X_3 \end{aligned} \tag{5.3}$$

where X's are the driving forces. For example, we can assume that J_1 is a heat flux. Then the primary driving force is X_1 which is dT/dx, and $L_{11} = \kappa$, the heat flow coefficient. The other fluxes may be diffusional fluxes of each of the various components. For example, J_2 can have $X_2 = dC_2/dx$, and J_3 can have $X_3 = dC_3/dx$. The interesting major contribution of Onsager is that there is a relationship among the L's:

$$L_{ik} = L_{ki} \tag{5.4}$$

Here we do not consider this important field of nonequilibrium thermodynamics further. However, the useful message is that there are many important physical phenomena that are governed by similar formulas. In this chapter we specifically consider only solutions to the Fickian diffusional formulas, but we can recognize the similarities in mathematical form of other phenomena and consequently the similarities in the form of resulting solutions. A large number of applications exist for the diffusion equations from mass flow to all of the formulas above for different fluxes (heat, charge, etc.). Even quantum mechanics, which makes much use of the Schrödinger equation (see Chapter 9) and which is constructed of a first temporal derivative and second spatial derivatives, has the form of Fick's second law. Each of these laws has similar mathematical form. But they not only govern vastly different physics, the boundary conditions for the solutions are different. Hence the solutions are different. Nevertheless, there are similarities in the fundamental physics that underly the phenomenological equations, mainly that fluxes are in response to forces.

In this chapter three aspects of the subject of mass flow by diffusion in solids are considered: physical models and ideas, mathematical aspects, and applications such as nucleation and phase changes.

5.2 ATOMISTIC THEORY OF DIFFUSION: FICK'S LAWS AND A THEORY FOR THE DIFFUSSION CONSTRUCT *D*

Consider two adjacent planes at X and $X + \Delta X$ shown in Figure 5.1*a*, that are a distance ΔX apart. (Later we will consider this separation to be a lattice spacing or lattice parameter, a_0.) There are other parallel planes some ΔX apart and others farther away, but we now only consider these two. We proceed by defining the number density of atoms N = #/area and set the initial number densities as N_x on the plane X and $N_{x+\Delta X}$ on the plane $X + \Delta X$. Only jumps of atoms to the left or right in units of ΔX (or a_0) are permitted (1-D). This means that the atoms must jump to an empty state on the adjacent plane. We assume that sufficient empty states are available, and a jump does not depend on a previous jump to leave an empty state. A jump frequency Γ is defined by units #jumps/sec·atom. Then the number of atoms jumping from plane X in the time interval δt is

$$N_x \cdot \Gamma \cdot \delta t \, [(\#\,\text{atoms/area}) \cdot (\#\,\text{jumps/sec·atom}) \cdot (\text{time})]$$

With only left or right jumps of distance ΔX, the number of jumps to the right from X to $X + \Delta X$ will be given as

$$\#\mathrm{R} = \frac{1}{2} N_x \cdot \Gamma \cdot \delta t \qquad (5.5)$$
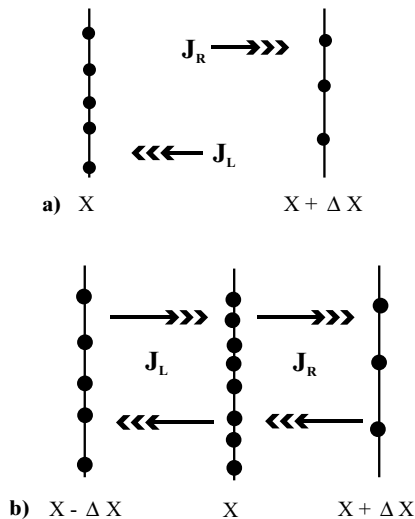
Figure 5.1 (*a*) Fluxes to and from two adjacent planes for steady state diffusion; (*b*) fluxes to the right and left from and to a center plane for non–steady state diffusion.

A similar argument applies for the atoms jumping from plane $X + \Delta X$ to plane X, or to the left:

$$\# L = \frac{1}{2} N_{x+\Delta x} \cdot \Gamma \cdot \delta t \tag{5.6}$$

The other half of the jumps occur (to a plane to the left of X and right of $X + \Delta X$) with equal frequency, but the result of those jumps do not appear on the two planes that are presently being considered. For our purpose the other half of the jumps are unobserved. Notice in equations (5.5) and (5.6) that for atoms, planes and jumps are all the same; the number of jumps in a given time interval is only a function of the number of atoms per plane.

The net flux, J, is obtained from the difference in the fluxes to the right and left, and given as

$$J = \frac{1}{2} \Gamma (N_x - N_{x+\Delta x}) \tag{5.7}$$

To make equation (5.7) conform to the usual form for Fick's laws, we proceed to convert the number of atoms on a plane (N) to concentrations (C). With concentration as the number per volume or $C = \#/V$, it follows that $C \cdot \Delta X = N$, the number per area. Then we obtain

$$J = \frac{1}{2} \Gamma (C_x \cdot \Delta X - C_{x+\Delta x} \cdot \Delta X) \tag{5.8}$$

ΔX is introduced into the numerator and denominator of equation (5.8) by multiplying by $\Delta X / \Delta X$, to obtain

$$J = \frac{1}{2} \Gamma \cdot \Delta X^2 \cdot \left\{ \frac{\Delta C}{\Delta X} \right\} \tag{5.9}$$

Converting to small changes (Δ to d) and setting $\Delta X = a_0$, the lattice spacing, obtains the following formula:

$$J = \frac{1}{2} \Gamma a_0^2 \left\{ \frac{dC}{dX} \right\} \tag{5.10}$$

Since $D = \frac{1}{2} \Gamma a_0^2$, we have what looks like Fick's first law:

$$J = \frac{-Ddc}{dx} \tag{5.11}$$

D is defined above, and the minus sign is added to indicate the flux down the gradient. In 3-D with $\frac{1}{3}$ of the flux along each coordinate,

$$D = \left(\frac{1}{6} \right) a_0^2 \Gamma \tag{5.12}$$

The simple way we used to obtain Fick's first law reveals the dynamics underlying net diffusive flux. It is basically random jumps along with an imbalance in N's on planes X and $X + \Delta X$. In other words, the plane with more N's will have the greatest number of jumps if other potential variables, X (or a_0) and Γ, are equal. We can summarize the physics of Fickian diffusion as that process of mass flow that depends on the concentration gradient and randomness. Strictly speaking, we could have used the thermodynamic chemical activity instead of concentration. However, these are nearly equivalent in simple cases. If, on the other hand, the jumps were somehow biased, then the net flux calculated using equation (5.7) would be in error. Later we will briefly discuss convectional mass flow where the randomness assumption is lifted.

To obtain a feel for the magnitude of Γ, it is instructive to consider some real numbers. For diffusion of carbon in αFe, D is about $10^{-6}\,\text{cm}^2/\text{s}$ at 900°C (this is a rather large D for diffusion in solids). We can assume that a_0 is about $1\,\text{Å}$ or $10^{-8}\,\text{cm}$. From $\Gamma = 2D/a_0^2$, $\Gamma = 2 \cdot 10^{-6}/10^{-16}$ or $10^{10}\,\text{s}^{-1}$. Thus C changes position or jumps 10 billion times per second! Consider that an atomic vibration is about $10^{13}\,\text{s}^{-1}$. So there are 10^3 vibrations for every jump. For many metals D is around $10^{-8}\,\text{cm}^2/\text{s}$ near the melting point, and this yields Γ of about $10^8\,\text{s}^{-1}$. Thus only one in 10^5 vibrations results in a jump. So while jump frequencies appear to be high numbers, the actual jumping of atoms to adjacent sites is an infrequent occurrence on the atomic scale.

Fick's first law is useful when concentrations are established at all points in a system. However, it is often the case that the C's are a function of time in a process. For example, at the beginning of the diffusion of Ni into pure Cu, there is no Ni in the Cu; the Ni concentration in Cu builds up over time. The same treatment as done above can be pursued to find the time evolution of concentration on a given plane ($C(t)$ on X). Consider Figure 5.1b where three planes are considered. The number arriving at X from the two adjacent planes is

$$N_{\text{Arrive}} = \frac{\Gamma}{2} N_{x-\Delta x} + \frac{\Gamma}{2} N_{x+\Delta x} \tag{5.13}$$

The number leaving X is

$$N_{\text{leaving}} = \Gamma N_x \tag{5.14}$$

The total change is obtained from the difference. Since half of the jumps are from X to the left and half are to the right, the net change at X is given as

$$\frac{\delta N_x}{\delta t} = \frac{\Gamma}{2}[(N_{x-\Delta x} - N_x) + (N_{x+\Delta x} - N_x)] \tag{5.15}$$

Here the left-hand term in the square brackets is for the change from the left (L), and the right-hand term is for the change from the right (R). However, if $L \neq R$, then a steady state does not obtain, and $dC/dT \neq 0$, and N (or C) is a function of time. Now convert N's to C's as before, using

$$N = C \cdot \Delta x \quad \text{and} \quad \frac{dN}{dt} = \frac{dC}{dt} \cdot \Delta x \tag{5.16}$$

The following result is obtained:

$$\frac{\delta C}{\delta t} \cdot \Delta x = \frac{\Gamma}{2}\Delta x[(C_{x-\Delta x} - C_x) + (C_{x+\Delta x} - C_x)] \tag{5.17}$$

The expression in square brackets is essentially ΔC on the X plane. To put this expression in appropriate form, we divide through by Δx and then multiply the right-hand side by $\Delta x/\Delta x$ to obtain

$$\frac{\delta C}{\delta t} = \frac{\Gamma}{2}\Delta x^2 \frac{(\Delta C)}{\Delta x^2} = \frac{\Gamma}{2}\Delta x^2 \frac{d}{dx}\left(\frac{\Delta C}{\Delta x}\right) \tag{5.18}$$

Upon converting equation (5.18) to derivatives, we have Fick's second law:

$$\frac{dC}{dt} = \frac{\Gamma}{2}\Delta x^2 \frac{d^2 C}{dx^2} \tag{5.19}$$

with $D = \frac{1}{2}\Gamma a_0^2$ for $\Delta x = a_0$, as before.

While the method to obtain Fick's second law is appealing in that it follows the method to obtain the Fick'e first law, it is useful to take a step back and consider what is happening. In looking again at Figure 5.1b, we see two flux problems rather than one as was considered in Figure 5.1a for Fick's first law. That is, we see one problem yielding a net flux, say, J_L by at planes $X - \Delta X$ and X and a second problem yielding a net flux J_R at planes $X + \Delta X$ and X. Since the planes $X - \Delta X$ and $X + \Delta X$ can arbitrarily have different N's (again assuming equal Γ's and random jumps), then the two net fluxes are not equal, $J_L \neq J_R$ in general. For this reason the number of atoms on X per unit area, N_x, will be a function of time, dN_x/dt, and specifically dN_x/dt will be determined by the change in J or $(J_R - J_L)$ or dJ for small changes. Also we know that

$$\frac{dN}{dt} = dx\frac{dC}{dt} \tag{5.20}$$

because $C = N/V$ and the volume $V = Adx$. Putting this together, we have

$$\frac{dN}{dt} = dx\frac{dC}{dt} = -dJ$$

then

$$\frac{dC}{dt} = -\frac{d}{dx}J = -\frac{d}{dx}\left(\frac{-DdC}{dx}\right) = D\frac{d^2 C}{dx^2} \tag{5.21}$$

which is Fick's second law again. The formula above for dN/dt has a negative sign because the fluxes with a minus sign from the $X - \Delta X$ and $X + \Delta X$ planes yield a net gain or $+dN/dt$. The message from this procedure is that the change in C with t, or dC/dt, is given by the change in the flux with x or dJ/dx.

5.3   RANDOM WALK PROBLEM

From the preceding discussions, we know that diffusion is the time integral of all the random jumps. Now imagine the jumps to be occurring at a rapid statistical rate (as was estimated above), and every once in awhile a snapshot of the atoms is taken in order to determine their relative positions. The diffusion distance is the final place of an observed atom after a specified time. With the random jumps in virtually all directions and with a huge number of jumps, is seems impossible to analytically determine the displacement of atoms. The problem seems intractable at first. However, the very randomness and large numbers enables a solution with far-ranging implications. The problem is a classical physics problem called the random walk problem.

It is usual for a random walk problem, as applied to diffusion, to include a large number of random jumps, with each jump being unbiased by the previous jumps, and to aim at computing how far from the start position a particular atom can be expected to be after a time interval. Figure 5.2 shows a composite of a number of snapshots for a particular atom where $\mathbf{r}_i$ are the vectors that correspond to each of the i jumps that this atom makes. There is a defined start, S, and finish, F, point. The final displacement from S to F, designated by the vector $\mathbf{R}_n$ can be obtained by vector addition:

$$\mathbf{R}_n = \mathbf{r}_1 + \mathbf{r}_2 + \mathbf{r}_3 + \ldots = \sum_{i=1}^{n} \mathbf{r}_i \tag{5.22}$$

In order to find the magnitude of the vector $\mathbf{R}_n$, the dot or scalar product of $\mathbf{R}_n$ with itself ($\theta = 0°$) is determined:

$$\begin{aligned}
R_n^2 &= \mathbf{R}_n \cdot \mathbf{R}_n \\
&= \mathbf{r}_1 \cdot \mathbf{r}_1 + \mathbf{r}_1 \cdot \mathbf{r}_2 + \ldots + \mathbf{r}_1 \cdot \mathbf{r}_n \\
&\quad \mathbf{r}_2 \cdot \mathbf{r}_1 + \mathbf{r}_2 \cdot \mathbf{r}_2 + \ldots + \mathbf{r}_2 \cdot \mathbf{r}_n \\
&\quad \mathbf{r}_3 \cdot \mathbf{r}_1 + \ldots \mathbf{r}_3 \cdot \mathbf{r}_3 + \ldots + \mathbf{r}_3 \cdot \mathbf{r}_n \\
&\quad \ldots \mathbf{r}_n \cdot \mathbf{r}_n \\
&= R_n^2 \cos \theta
\end{aligned} \tag{5.23}$$
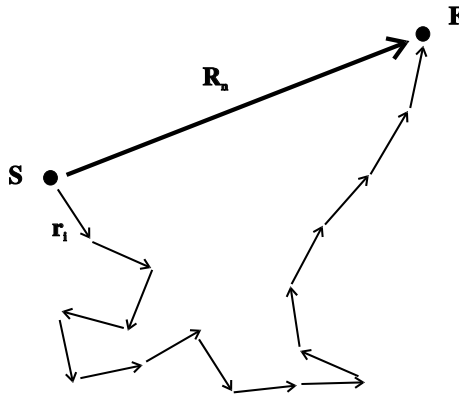
Figure 5.2   Random Jumps from the start position S to the finish F are given by the vectors $\mathbf{r}_i$. The resultant of the jumps is $\mathbf{R}_n$.

This array of products can be simplified and written as a series of sums, where the first is a sum of diagonal terms

$$\sum_i^n \mathbf{r}_i \cdot \mathbf{r}_i \tag{5.24}$$

and the second consists of semidiagonal terms as the sums of $\mathbf{r}_i \cdot \mathbf{r}_{i+1}$ and $\mathbf{r}_{i+1} \cdot \mathbf{r}_i$. The semidiagonal terms are equal and can be combined, and there are $n - 1$ of these semi-diagonal terms:

$$2\sum_{i=1}^{n-1} \mathbf{r}_i \cdot \mathbf{r}_{i+1} \tag{5.25}$$

The next and succeeding terms are developed as above:

$$2\sum_{i=1}^{n-2} \mathbf{r}_i \cdot \mathbf{r}_{i+2} + \ldots \tag{5.26}$$

In summary, this yields

$$R_n^2 = \sum_{i=1}^n \mathbf{r}_i^2 + 2\sum_{j=1}^{n-1}\sum_{i=1}^{n-j} \mathbf{r}_i \cdot \mathbf{r}_{i+j} \tag{5.27}$$

This can be put into more usable form as

$$R_n^2 = \sum_{i=1}^n \mathbf{r}_i^2 + 2\sum_{j=1}^{n-1}\sum_{i=1}^{n-j} |\mathbf{r}_i||\mathbf{r}_{i+j}| \cos \theta_{i,i+j} \tag{5.28}$$

where θ is the angle between the adjacent vectors (of course, $\theta = 0$ when the i's are the same and thus the cosine term is unity for the first term on the right of equation 5.28). The manipulations above have made no restrictions upon the randomness of successive jumps, the jump distances, the angles, or anything else. The first assumption is to consider crystalline materials and, in particular, cubic crystals where each jump distance is identical (i.e., all $\mathbf{r}$'s are the same) then R_n^2 is given as

$$\begin{aligned} R_n^2 &= n\mathbf{r}^2 + 2\mathbf{r}^2\sum_{j=1}^{n-1}\sum_{i=1}^{n-j} \cos \theta_{i,i+j} \\ &= n\mathbf{r}^2\left(1 + \frac{2}{n}\sum_{j=1}^{n-1}\sum_{i=1}^{n-j} \cos \theta_{i,i+j}\right) \end{aligned} \tag{5.29}$$

where the identity of the $\mathbf{r}$'s is dropped. This equation gives R_n^2 for one particle after n jumps. The average value is obtained from many particles each with n jumps. The value of nr^2 remains unchanged, but the double sum terms are averaged with the result:

$$\overline{R_n^2} = nr^2\left(1 + \frac{2}{n}\overline{\sum_{j=1}^{n-1}\sum_{i=1}^{n-j} \cos \theta_{i,i+j}}\right) \tag{5.30}$$

Now comes the important part. If we simply assume that there is randomness in the jumps so that each jump direction is independent of the preceding jump and all jumps are equally probable negative and positive jumps, then the + and − values of cosine terms will add to 0, and the final result is obtained:

$$\overline{R_n^2} = nr^2 \quad \text{or}$$
$$\sqrt{\overline{R_n^2}} = \sqrt{n}\, r \tag{5.31}$$

From equation (5.31) we learn that the root mean square displacement given on the left side of the equation is proportional to n. For amorphous materials the jump distance is irregular, and thus the average value $\bar{r}$ can be used for **r** so that

$$\overline{R_n^2} = n\bar{r}^2 \tag{5.32}$$

5.3.1 Random Walk Calculations

Start with some unit cell where atomic distances are of the order of 10^{-8} cm or 1 Å (10^{-10} m). Consider a diffusivity of about $D = 10^{-6}$ cm²/s that approximately corresponds to C in αFe at 900°C, as was used above. Now, equation (5.31) yields $R_n = 10^{-5}$ m for 1 second using $n = 10^{10}$ s⁻¹ (see section 5.2 where $\Pi = 2D/a_0^2$) and $r = 10^{-10}$ m. In 3 hours or about 10^4 seconds, n is about 10^{14} yielding $R_n = 10^{-3}$ m as the random walk distance. However, if we consider the straight line distance obtained by placing all the r's end to end, or the total distance traveled, we obtain (10^{-10} m · 10^{10} s⁻¹ · 10^4 s) 10^4 m or about 6 miles compared to the net distance of 10^{-3} m.

5.3.2 Relation of *D* to Random Walk

With the random walk result in terms of the lattice parameter,

$$R_n^2 = nr^2 = na_0^2 \tag{5.33}$$

Using the results in Section 5.2 where $\Gamma = \#/t \cdot \text{atom} = n/t$, we have $D = \frac{1}{2}a_0^2\Gamma$ and

$$Dt = \frac{1}{2}a_0^2\Gamma t = \frac{1}{2}a_0^2 n \quad \left(\text{or} \left(\frac{1}{6}\right)na_0^2 \text{ in 3-D} \right) \tag{5.34}$$

or $2Dt = na_0^2$ per direction. In 3-D multiply by $\frac{1}{3}$, because there are $\frac{1}{3}$ jumps in any one direction. Finally we obtain an expression for D in terms of fundamental parameters:

$$6Dt = na_0^2 = R_n^2 \tag{5.35}$$

and

$$D = \left(\frac{1}{6}\right)\Gamma a_0^2 \tag{5.36}$$
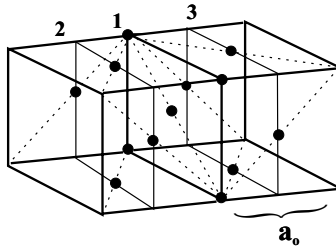
Figure 5.3 FCC lattice showing 12 nearest neighbors for an atom at the center of plane 1.

5.3.3 Self-Diffusion Vacancy Mechanism in a FCC Crystal

In Figure 5.3 we see two joined FCC unit cells (without all the positions occupied for clarity) where the nearest neighbors and planes are identified (planes labeled 1, 2, 3). We consider the self-diffusion problem where a tracer (an isotope) atom is used to monitor the migration. Γ is taken as the average number of jumps per second per atom, as above. With n_1 as the number of tracer atoms on plane 1 then $n_1 \cdot \Gamma \cdot \delta t$ is the number of tracer that will jump from plane 1 in a time interval δt. $\Gamma \cdot \delta t$ is proportional to the number of nearest neighbor sites multiplied by the probability of nearest neighbor sites being vacant, p_v, and the probability that a tracer atom will jump to a particular vacant site in δt (i.e., $\omega \cdot \delta t$ where ω is the frequency for the vibration). All of this is summarized as

$$\Gamma \cdot \delta t = 12 \cdot p_v \cdot \omega \cdot \delta t \tag{5.37}$$

where the number 12 comes from the number of nearest neighbors for an FCC, as seen in Figure 5.3. For this calculation consider the atom at the center of plane 1, and note that there are 12 nearest neighbors with 4 each on planes 1, 2, and 3. For a flux from plane 1 to plane 2, J_{12} only 4 of 12 provide potential nearest neighbor sites, so

$$J_{12} = 4 \cdot n_1 \cdot p_{v2} \cdot \omega_{12} \tag{5.38}$$

Likewise for the reverse flux,

$$J_{21} = 4 \cdot n_2 \cdot p_{v1} \cdot \omega_{21} \tag{5.39}$$

In a pure metal where the adjacent sites are indistinguishable,

$$\omega_{21} = \omega_{12} \tag{5.40}$$

$$p_{v1} = p_{v2} \tag{5.41}$$

and $n_1 = x \cdot C_1$, which is the number per area $= C_1 \cdot a_0/2$. As a result

$$J = 4 \cdot \frac{a_0}{2} \cdot p_v \cdot \omega \cdot (C_1 - C_2) \tag{5.42}$$

and with $(C_1 - C_2) = -(a_0/2)dC/dX$, we obtain

$$J = -a_0^2 \cdot p_v \cdot \omega \cdot \frac{dC}{dX} \qquad (5.43)$$

$$p_v = N_v = \text{\# of vacancies} \qquad (5.44)$$

then $D = a_0^2 \cdot N_v \cdot \omega$. Once again, we have D in terms of fundamental parameters. For the case of interstitial diffusion a similar development yields $D = \gamma \cdot a_0^2 \cdot \omega$, where γ is a geometrical constant.

5.3.4   Activation Energy for Diffusion

Experimentally it is found that D follows an Arrhenius expression as

$$D = D_0 e^{-E_D/kT} \qquad (5.45)$$

To see that this functionality is observed for diffusion and for many physical and chemical processes, consider, again, a two-state situation as was treated in Chapter 4 for the prediction of the number of vacancies in a perfect crystal. The present situation involves the migration of an atom from one position to a vacant position. For this it requires an energy input of E_D. (This example can also be treated as a vacancy moving in the opposite direction.)

As is usual in virtually all processes, for a diffusing species to move to a new site (whether substitutional or interstitial), some energy input is required (for bond breaking, for squeezing through interstices and over saddle points, etc.). Thus this situation can be cast as the two-state problem in Chapter 4 (e.g., see the development of equation 4.34) where a Boltzmann-like energy factor determines the occupancy of the states based on the energy differences among states.

5.4   OTHER MASS TRANSPORT MECHANISMS

5.4.1   Permeability versus Diffusion

Figure 5.4a illustrates the common situation of two interfaces with solid state diffusion occurring between the interfaces. At interface 1, say, a gas solid interface, although this is not necessary, a gas dissolves in the solid, meaning the gas crosses boundary 1 and dissolves. Following this interface reaction, the gas dissolved in the solid diffuses by one mechanism or another to interface 2 where it exits into vacuum. If we tag a particular gas species and follow it from the gas on the left to the vacuum on the right and measure its flux, we call this net observed flux for the three series processes a permeation flux, and the overall process permeation. For such a series process, the slowest of the three steps in the series process determines the permeation rate. It may be diffusion or one of the interface reactions at 1 or 2 (solubilization or desolubolization, respectively). The permeation flux or rate is often the experimentally determined rate. As such it is often the rate that is erroneously ascribed to diffusion, particularly, where it is believed that diffusion is occurring somewhere in the system. One useful, but not exclusive, test to determine whether the measured flux is really a diffusion flux, and not limited by some other
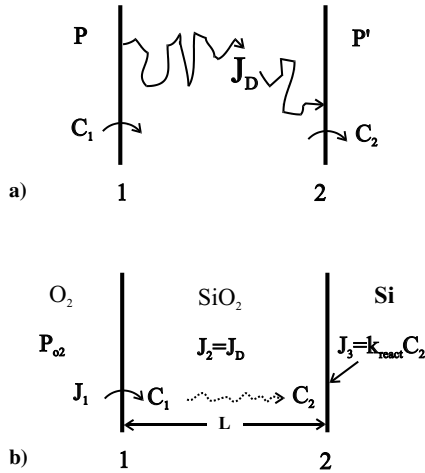
Figure 5.4 (*a*) Three series processes showing dissolution at plane 1, diffusion from plane 1 to plane 2, and removal at plane 2; (*b*) an actual example of three processes in series for the oxidation of Si in O_2.

process, is to determine if the measured flux is proportional to $1/L$, where L is the path length in the solid, namely the distance between interfaces 1 and 2. This test is obtained from Fick's first law where the flux J is proportional to $\Delta C/\Delta x$ and Δx is the path length. If the test is positive, the flux is likely a diffusive flux.

In summary, permeability includes the reactions at 1 and 2 and diffusion in the case above, and in general, all the processes from input to output. Permeability is a measure of the total transport of matter through a system.

A straightforward example of three processes in series, where one is diffusion, is the thermal oxidation of silicon in oxygen gas. In this process a cleaned silicon surface, typically a single crystal surface, is exposed to oxygen at temperatures above 500°C. Almost instantly a SiO_2 film forms, and from this point in time forward the oxide grows. To follow the film formation kinetics, one observes the increase in the thickness L of SiO_2 as a function of time, dL/dt. As shown in Figure 5.4b, the first process in the series process scheme is the dissolution of O_2 in the SiO_2 at the gas-solid O_2–SiO_2 interface. This process is typically fast and results in a concentration of O_2 in the SiO_2 of C_1 at the outer interface. The flux corresponding to this process is

$$J_1 = k_{sol}p_{ox} \tag{5.46}$$

This process is essentially Henry's law. The dissolution of O_2 in SiO_2 is proportional to the pressure of O_2 with the Henry law constant k_{sol}. The second process is the diffusion of O_2 through the SiO_2 of thickness L to the SiO_2–Si interface, resulting in a smaller concentration of O_2, C_2. In the steady state the flux corresponding to this process is given as

$$J_2 = \frac{D_{ox}dC}{dL} = \frac{D_{ox}(C_1 - C_2)}{L} \tag{5.47}$$

The third process in the series scheme is the reaction of the O_2 at concentration C_2 with Si at the inner SiO_2–Si interface. This process continually removes O_2 and prevents accumulation. The flux corresponding to this process is expressed by a first-order reaction between O_2 with Si:

$$J_3 = k_{react}C_2 \tag{5.48}$$

A first-order rate expression with k_{react} the first-order rate constant was chosen initially for simplicity to describe the reaction of the O_2 arriving at the Si surface to react with a constant concentration of Si at the crystal surface. Later research confirmed the assumption of first order in O_2. This three flux scheme can be simplified by considering that J_1, a gas phase flux, is much faster than the other two fluxes that occur in the solid state. With processes in series the slowest process determines the observed rate. Therefore, because J_1 is much faster than the other fluxes, we need not consider it further. Fluxes J_2 and J_3 must be equal after an initial transient. To see this, consider the consequences of using the formulas above for J_2 and J_3. If the fluxes were not equal, then either J_2 or J_3 would be the larger flux. If J_2 is larger, then the supply flux of oxidant to the SiO_2–Si interface (that depends on the difference $C_1 - C_2$) would increase C_2 at the interface. The increase in C_2 would in turn increase J_3, which would reduce C_2 and regulate the process. Likewise, if J_3 were larger than J_2, then J_3 would reduce C_2, and in so doing, increase the difference $C_1 - C_2$ and thereby increase J_2, again regulating the series steps. Thus this self-regulating set of series fluxes must be equal, and hence a steady state obtains after some initial transient during which the fluxes equalize. Now with $J_2 = J_3$, C_2 can be obtained as follows:

$$k_{react}C_2 = \frac{D_{ox}(C_1 - C_2)}{L} \tag{5.49}$$

and

$$C_2 = \frac{D_{ox}C_1}{k_{react}L + D_{ox}} \tag{5.50}$$

A rate expression can be formed using this value for C_2 in the relationship:

$$J = J_2 = J_3 = \Omega dL/dt = k_{react}C_2 \tag{5.51}$$

with C_2 substituted from above, and Ω is the conversion from moles of oxygen gas in J_2 to moles of oxygen in solid SiO_2, and Ω has the value of about $2.3 \times 10^{22}\,cm^{-3}$. This value is obtained from the density of SiO_2 of $2.2\,g/cm^3$ divided by $60\,g/mol$ for the molecular weight of SiO_2, all multiplied by Avagadro's number. A rate equation can be written with these results as follows:

$$\Omega \frac{dL}{dt} = \frac{k_{react}D_{ox}C_1}{k_{react}L + D_{ox}} \tag{5.52}$$

Upon integration this yields a linear-parabolic dependence of thickness L on time of oxidation t as follows:

$$t + \text{const} = AL^2 + BL \tag{5.53}$$
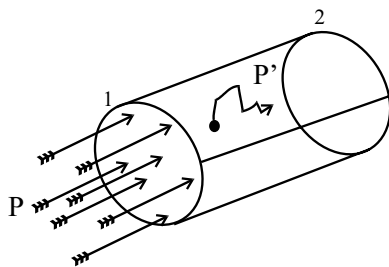
Figure 5.5 Pressure difference across a pipe causes convective gas flow from plane 1 to plane 2. The motion of a gas particle (*filled circle*) in the pipe is biased by the convective flow (*arrows*).

where $A = \Omega/2D_{ox}C_1$ and $B = \Omega/k_{react}C_1$. A and B are called the parabolic and linear rate constants, respectively. Diffusion through D_{ox} enters into the parabolic constant while the linear constant depends upon the interface reaction through k_{react}. This linear-parabolic formula has been verified for the thermal oxidation of silicon. This real example illustrates how mass transport, in terms of diffusion, couples into the permeation problem, and it illustrates that often the overall permeation process is observed in which diffusion is only a part of the process.

5.4.2 Convection versus Diffusion

Figure 5.5 shows a cylindrical system (a pipe) with two boundary planes 1 and 2. Diffusion can be thought of as occurring from either plane as a random jumping from an occupied state, say, on plane 1, with equal probability in all directions and observed as a flux when the species arrives at an unoccupied state on plane 2. The diffusional flux is written as J_{12}, and as discussed above, it depends on N_1 and N_2 or, better, on the concentration gradient from 1 to 2. At low total gas pressures the randomness required for diffusional transport will be maintained. However, we can also imagine a large pressure drop Δp ($= p - p'$) from planes 1 to 2, and high total pressures of a species. The pressure drop from planes 1 to 2 gives rise to a concerted flow of gas as is indicated by the arrows in Figure 5.5. Then we can imagine that a particle on plane 1 is repeatedly bumped by the particles falling down the $\Delta p/\Delta x$ gradient (where Δx is the distance from 1 to 2) from the left at 1 toward 2. Repeated collisions with other particles moving in this direction will eventually impart a velocity component to the originally randomly moving particle that undergoes collisions. Thus the motion of the particle will no longer be random, but rather strongly influenced by the surrounding flowing particles. It is as if a wind were blowing from 1 to 2 that carries randomly moving particles along. This kind of motion is called convective transport. Convective transport occurs at high p and large Δp, while diffusional transport occurs at low p and Δp so as to maintain randomness.

5.5 MATHEMATICS OF DIFFUSION

In this section we solve Fick's laws for several important cases. This is in contrast to the approach taken above where Fick's laws were obtained from first principles. Both approaches are important, since by the atomistic approach we can understand the nature

of diffusion and by the phenomenological approach we can understand the dynamical process of diffusion.

5.5.1 Steady State Diffusion—Fick's First Law

The phenomenological diffusion flux equation shown at the beginning of this chapter as equation (5.1), known as Fick's first law, can be applied directly to situations where there exists a steady state condition in concentration. This means that at every point in the medium under study, the concentrations are not changing in time. This condition does not mean that the concentration (or better chemical activity) is the same at every point. Indeed, if that were the case, then the driving force for diffusion, the concentration gradient, would be absent ($dC/dx = 0$). The steady state is expressed as

$$\frac{dC}{dt} = 0 \tag{5.54}$$

with the x axis parallel to the gradient in the concentration, a flux J is produced in response to the force, dC/dx:

$$J \propto \frac{-dC}{dx} \tag{5.55}$$

It is intuitive that the flux is down the concentration gradient in the direction of decreasing C; hence a negative sign is appropriate to indicate this. With a constant of proportionality, a phenomenological coefficient, D, we obtain Fick's first law:

$$J = \frac{-DdC}{dx} \tag{5.56}$$

It is useful to examine the units for diffusion problems:

$$J\left(\text{mass}/\text{displacement}^2 \cdot \text{time}\right) = \frac{-D\left(\text{displacement}^2/\text{time}\right)dC}{dx\left(\text{mass}/\text{displacement}^4\right)}$$

In usual units for mass, displacement, and time,

$$J\left(\text{mol}/\text{cm}^2\text{s}\right) = -\frac{D\left(\text{cm}^2/\text{s}\right)\cdot dC}{dx\left(\text{mol}/\text{cm}^4\right)}$$

It is found that the diffusion coefficient, D, often called the diffusivity varies with direction in a crystal lattice. This is entirely sensible because the derivations above of Fick's laws are based on the random jumping from and to atomic positions or interstices, both of which have crystallographic bias. Thus D is a tensor of rank 2 (it has three variations) corresponding to the threefold coordinate system:

$$\begin{vmatrix} D_a & 0 & 0 \\ 0 & D_a & 0 \\ 0 & 0 & D_a \end{vmatrix} \text{ and } \begin{vmatrix} D_a & 0 & 0 \\ 0 & D_b & 0 \\ 0 & 0 & D_c \end{vmatrix} \tag{5.57}$$

for cubic and orthorhombic systems, respectively. The appropriate flux equations are

$$J_x = \frac{-D_{11}dC}{dx} + \frac{-D_{12}dC}{dy} + \frac{-D_{13}dC}{dz}$$

$$J_y = \frac{-D_{21}dC}{dx} + \frac{-D_{22}dc}{dy} + \frac{-D_{23}dC}{dz} \quad (5.58)$$

$$J_z = \frac{-D_{31}dC}{dx} + \frac{-D_{32}dC}{dy} + \frac{-D_{33}dC}{dz}$$

However, in most cases an average D is used in a single flux equation.

An Example of Fick's First Law—Steady State Diffusion. The carburizing of Fe provides that produces steel is a classic example of a steady state situation. An iron pipe with radius r, wall thickness dr, and length l is shown in Figure 5.6. To carburize the pipe to make steel, one can flow CH_4 (a carburizing or carbon containing gas) down the pipe with the pipe at temperature T and with a gas that can react with and/or carry the C away from the outer surface of the pipe. The C from the CH_4 will commence to diffuse into the Fe. When the concentration of C, [C], in the pipe does not change with time as given by equation (5.54), a steady state in C is given as

$$\frac{d[C]}{dt} = 0 \quad (5.59)$$

Then the amount of C, q divided by the time is constant, q/t, is constant. This steady state condition occurs after a transient when C builds up from an initial zero concentration. The flux is written as

$$J_C = \frac{q}{A \cdot t} \quad (5.60)$$

With area of a section of pipe with length l given as $A = 2\pi rl$, the flux is

$$J_C = \frac{q}{2\pi rl \cdot t} \quad (5.61)$$
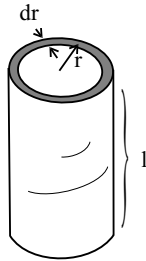
Figure 5.6 Iron pipe of length *l* radius *r* and wall thickness *dr* caburized by flowing a carbon containing gas down the inside of the heated pipe.

Under steady state, by Fick's first law, we have

$$\frac{q}{2\pi r l \cdot t} = \frac{-DdC}{dr} \tag{5.62}$$

This can be readily solved for q as

$$q = \frac{-D \cdot 2\pi r l t \cdot \mathrm{dC}}{dr} = \frac{-D \cdot 2\pi l t d\mathrm{C}}{dlnr} \tag{5.63}$$

Experiments are done where q is measured (the total amount of C) as well as C as a function of r with t and l known. One way to obtain C as a function of depth r is to literally shave the pipe on a lathe, keeping track of the depth into the pipe wall and chemically analyze the shavings for C. A plot of C versus lnr yields a straight line with slope = $-q/D \cdot 2\pi l t$ from which D can be readily extracted. This analysis assumes a steady state and that D is not a function of concentration.

Earlier in Section 5.4.1 the example of permeability discussed was the oxidation of Si with gaseous O_2 where a steady state approximation among the series processes led to the equation

$$t + \mathrm{const} = AL^2 + BL \tag{5.53}$$

where $A = \Omega/2D_{ox}C_1$ and $B = \Omega/k_{react}C_1$. This problem is similar to the Fe carburization problem above in that there is a supply of diffusant on one side of a solid and a mechanism to remove diffusant on the other side of the solid, although these methods are different in the two problems. The series of three processes can lead to a steady state, and Fick's first law can apply to the diffusion step. For large SiO_2 film thicknesses the parabolic term with L^2 will be much larger than the linear term, $L^2 >> L$. Consequently the overall Si oxidation process will be dominated by the diffusion term that is the slow process for large thicknesses and long times. The result is that thickness and time are related by a parabolic relationship and the interface reactions can be ignored. On the other hand, for small SiO_2 film thicknesses the linear term can dominate. This example of Si oxidation is an important one in microelectronics, since Si is typically oxidized to form an electronically useful surface to build computer chips.

5.5.2 Non–Steady State Diffusion—Fick's Second Law

Many important problems in diffusion cannot be solved using the steady state assumption that $d[C]/dt = 0$. For example, in the preceding problem on the carburizing of Fe, one could inquire about the distribution of carbon before steady state is achieved at the beginning of the experiment, when the concentration of C is changing in time. Thus an expression for $d[C]/dt$ is required. We obtained the result using Fick's first law and mass balance, namely by keeping track of arrivals and departures on a given reference plane. This result was Fick's second law equation (5.21), which is often termed a continuity equation:

$$\frac{dC}{dt} = D\frac{d^2C}{dx^2} \tag{5.21}$$

We will illustrate below that the solutions to this kind of equation are dependent on the boundary conditions, or the conditions of the diffusion experiment.

5.5.2.1   *Solutions to Fick's Second Law*   We will consider several useful solutions to Fick's second law that are distinguished by initial and boundary conditions. As was done above, in all solutions we consider D to be a constant and not a function of concentration. We use the first, "thin film solution" to compute the diffusion of a fixed amount of material into another (or the same) material. Next, rather than starting with a fixed amount of diffusant, we consider an infinite source of diffusant, at least for short diffusion times. Last, we consider the case where surface concentration is held to zero or any fixed value. The surface concentration can be held to zero with the use of a reactant that reacts with diffusant coming to the surface and thereby removes it from the problem, as was done for the Fe carburization and the Si oxidation cases above. For simplicity, on the solutions we will at first assume that the initial concentration of diffusant is zero. Later we return to consider more complex solutions where this is not the case. These solutions are all short time solutions. The short time solutions are important in microelectronics where, for example, a dopant is desired in a localized region, and also for determining D's for various species. The final solution we consider is a general form for a long time solution. The limit of the long time solutions is complete homogenization, where $dC/dx = 0$. The long time solutions are particularly important in metallurgy where uniform distributions of selected alloy components are desired.

There are many possible technologically and scientifically important diffusion situations represented by Fick's second law, each of which has different boundary conditions and consequently different mathematical solutions. The treatment in this section is intended to cover only the more interesting cases and provide a brief outline of the mathematical methods. For more examples the reader is encouraged to consult the reading materials cited at the end of this chapter.

Thin Film Solution.   For this analysis Figure 5.7 shows a thin film of one material coated onto the end of a bar of another or the same material. If the materials are the same, the process is termed self-diffusion. The total amount of diffusant, α, is fixed at the number of atoms originally in the deposited film. The supply of diffusant becomes depleted as the diffusion processes continues. A large literature on the subject of self-diffusion in metals exists for the thin film solution. In our example we consider a radio tracer of a metal A, A^*, coated on one surface of a bar of pure A, and with a second bar of pure metal A welded to the free surface of the coating. This is nothing more than a thin film of A^* sandwiched in between two bars of pure A. This diffusion couple is heated at constant temperature T for a time t. After some diffusion time, the distribution of the concentration A^*, $C_{A^*}(x)$, is considered a Gaussian distribution whose tails of A^*
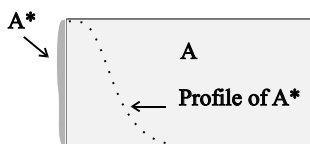
Figure 5.7   Bar of metal A coated with a thin film of an isotope of A, A^*. Heating the bar produces a Gaussian diffusion profile.

penetrate into the bars shown in Figure 5.7 (on one side of the sandwich) as the dashed curve. In Figure 5.7 the center of the tracer thin film is the origin for the problem at $x = 0$. After more time the material spreads into the bar and the peak in C_{A*} at $x = 0$, $C_{A*}(0)$ falls (the Gaussian flattens), but since the amount of material is constant, the area under the $C_{A*}(x)$ Gaussian is constant. Next, for this experiment, we need to set up the boundary conditions for the solution. The boundary conditions are the all important factors in correctly solving Fick's second law equation for a specific case.

First the initial and boundary conditions for the specific situation must be cast in analytical form:

1. Before the diffusion experiment starts, at $t = 0$, the concentration of diffusant, A^*, is 0 within the bar of A, so $C_{A*} = 0$ except at $x = 0$.
2. At $x = 0$, there is no gradient since there is a uniform layer of A^*, so $dC_{A*}/dx = 0$.
3. As the experiment ensues and after a short time the concentration of A^* remains at zero away from the surface of the bar. In other words, the time for the experiments is very short relative to the time needed for homogenization to occur, or even too short for A^* to reach the end of the bar. Thus $C_{A*}(\infty, t) = 0$.
4. The amount of diffusing material, A^*, is fixed. There is no source beyond the number of atoms originally deposited upon the bar, and this is expressed as

$$\int_{-\infty}^{\infty} C_{A*} = \alpha \tag{5.64}$$

The solution of Fick's second law for this situation is a Gaussian of the form

$$C(x, t) = \frac{\alpha}{\sqrt{\pi Dt}} e^{-x^2/4Dt} \tag{5.65}$$

This solution (shown to be the correct solution below) can be used to find D by taking the logarithm of both sides and then differentiating with respect to x^2 to yield the following:

$$\frac{d(\ln C_{A*})}{dx^2} = \frac{-1}{4Dt} \tag{5.66}$$

which can be visualized as a plot of $\ln C_{A*}$ versus x^2. The slope of the resulting straight line yields D.

If in the example above a radio tracer deposited on one end of the bar is of the same or different material, first there will be a fixed amount of diffusant (α), and then all of the diffusant will travel in the direction $x > 0$ with 0 defined as the plane of the surface of the bar. In the example depicted in Figure 5.7 there is no possibility of solid state diffusion in the $x < 0$ direction because there is no solid on that side of the deposit. However, suppose that after deposition of the material onto the bar, another bar is welded so that a symmetrical sandwich is created with diffusant in the middle of the two equal bars, as was discussed above. In this case half of α (or A^*) diffuses in $x > 0$ and half in $x < 0$ direction. The solution for either symmetrical half of the diffusion couple is

$$C(x, t) = \frac{\alpha}{2\sqrt{\pi Dt}} e^{-x^2/4Dt} \tag{5.67}$$

With only half of the couple it is as if the other half of the diffusing material is reflected to the present half, effectively doubling the amount of diffusant, and consequently the 2 in the denominator disappears.

Mathematical Interlude. Equation (5.67) (or equation 5.65) was stated to be the solution for the thin film case. However, in order to show that $C(x, t)$ above is a solution to Fick's second law, we need to do some math. In particular, we need to do the following:

1. Differentiate once with respect to t and twice with respect to x. Then these two terms should differ by D as given by Fick's second law.
2. Test the limits: (a) at $|x| > 0$, C approaches zero, $C \to 0$, as $t \to 0$; (b) $x = 0$, $C \to \infty$ as $t \to 0$.

The first derivative of $C(x, t)$ in equation (5.67) is

$$\frac{dC}{dt} = \frac{\alpha}{2}(\pi Dt)^{-1/2} \cdot e^{-x^2/4Dt} \cdot \frac{x^2}{4Dt^2} + e^{-x^2/4Dt} \cdot \frac{\alpha\pi D}{2} \cdot \left(-\frac{1}{2}\right)(\pi Dt)^{-3/2}$$

$$= \frac{\alpha x^2}{8Dt^2} \cdot \frac{e^{x^2/4Dt}}{\left(\pi Dt^{1/2}\right)} - \frac{\alpha\pi D}{4} \cdot \frac{e^{x^2/4Dt}}{\left(\pi Dt\right)^{3/2}} \tag{5.68}$$

Now the derivative of $C(x, t)$ is taken twice with respect to x:

$$\frac{dC}{dx} = \frac{\alpha}{2}(\pi Dt)^{1/2} \cdot e^{-x^2/4Dt} \cdot \frac{2x}{4Dt} = \frac{-\alpha x}{4Dt} \cdot \frac{e^{-x^2/4Dt}}{(\pi Dt)^{1/2}} \tag{5.69}$$

$$\frac{d^2C}{dx^2} = \frac{-\alpha x}{4Dt} \cdot \frac{e^{-x^2/4Dt}}{(\pi Dt)^{1/2}} \cdot \frac{-x}{2Dt} + e^{-x^2/4Dt} \cdot \frac{-\alpha}{4Dt(\pi Dt)^{1/2}}$$

$$= \frac{-\alpha x^2}{8D^2t^2} \cdot \frac{e^{-x^2/4Dt}}{(\pi Dt)^{1/2}} - \frac{-\alpha}{4Dt(\pi Dt)^{1/2}} \cdot e^{-x^2/4Dt} \tag{5.70}$$

Now a comparison of equations (5.68) and (5.70) shows that the difference between these two final expressions for dC/dt and d^2C/dx^2 is D.

While this agreement is necessary, it is not sufficient. In addition the consistency of initial conditions with the solution also needs to be tested. First at $x > 0$, $C \to 0$ as $t \to 0$. Simply inserting these in the solution $C(x, t)$ above yields the indeterminate form: $C = \infty/0$. To proceed, we make the following substitutions:

$$A = \frac{\alpha}{(4BD)^{1/2}}, \quad B = \frac{x^2}{4D}, \quad \text{then } C = \frac{Ae^{-B/t}}{t^2} \tag{5.71}$$

We next take the derivative of numerator and denominator, and apply the limits according to L'Hospital's rule. This is as follows:

$$\frac{At^{-3/2}}{e^{B/t}} = \frac{0}{\infty} \tag{5.72}$$

which remains indeterminate. To continue, we make some further substitutions:

$$z = \frac{B}{t}, \quad \text{then } t^2 = B^2 z^{-2} \tag{5.73}$$

and

$$C = \frac{Ae^{-z}}{B^2 z^{-2}} = \frac{Az^2}{B^2 e^z} \tag{5.74}$$

Taking the derivative yields

$$\frac{1}{z^2 e^z} \tag{5.75}$$

which yields 0 in the limit as $z \to \infty$ ($t \to 0$). This is the appropriate limit. Also $x = 0$, for $c \to \infty$ as $t \to 0$ needs to be checked. This is done as

$$\lim C(x,t) \text{ as } t \to 0 = \lim \left(\frac{\alpha}{2(\pi D t)^2} \right) = \infty \tag{5.76}$$

as is required. Thus the solution is checked as appropriate and correct.

Semi-infinite Solid Solution. Figure 5.8*a* shows a kind of problem addressed with the semi-infinite solid solution. Material *A* is joined to material *B*; each is pure and devoid of the other at the beginning of the experiment (later we will relax this requirement). One way to solve this problem is to think of each side of the diffusion couple as made
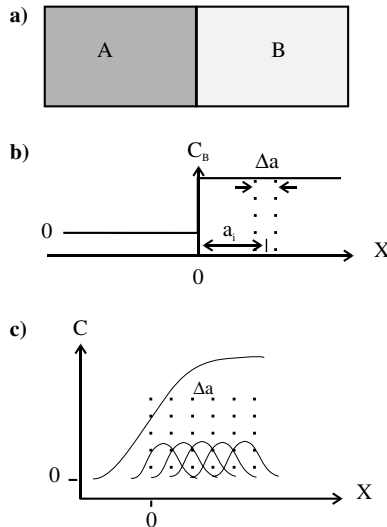
Figure 5.8   (*a*) Bar of metal *A* joined to metal *B* at 0; (*b*) thin slabs imagined at each side of the couple for which the thin film solution can be used; (*c*) the Gaussian in each slab during diffusion.

up of i small slices of width Δa and each of these slices are a_i away from the origin as is shown in Figure 5.8b. As A decreases on side A, B increases, and on side B as B decreases, A increases. We commence the analysis by considering events on side B where $x > 0$. One can imagine two identical bars joined where one is pure Cu and the other is pure Ni. Later we can consider more complex situations such as a pure Ni bar joined to a Cu bar initially with 10% Ni. As for the thin film case above, the first step is to write the initial and boundary conditions in analytical form:

1. At the start of the diffusion experiment, there is no mixing of A and B. $C_B = 0$ for $x < 0$ at $t = 0$ and $C_A = 0$ for $x > 0$ at $t = 0$.
2. Initially at $t = 0$, $C_B = C_{iB}$ for $x > 0$ and $C_A = C_{iA}$ for $x < 0$ at $t = 0$.

Using Figure 5.8b consider that a region in B is composed of n thin slices, each of width Δa and cross-sectional area 1, and each slice is at $x - a_i$. The use of $x - a_i$ for the position is useful because, when $x = a_i$, each of the Gaussians is centered on the slab in question. This will be clearer below. At the outset one slice in B has $C_{iB} \cdot V$ ((amount/vol)$\cdot$ vol) = $C_{iB} \cdot \Delta a \cdot 1$ of solute. For each slice the thin film solution discussed above applies; that is, in each slice a Gaussian obtains (each slice has a fixed amount of material) and changes with time in the slice where equation (5.67) is again given:

$$C(x,t) = \frac{\alpha}{2\sqrt{\pi Dt}} e^{-x^2/4Dt} \tag{5.77}$$

Figure 5.8c shows this situation where the solution would be the superposition of the Gaussians:

$$C(x,t) \approx \frac{C_{iB}}{2\sqrt{\pi Dt}} \sum_i^{i=n} \Delta a_i \cdot e^{-(x-a_i)^2/4Dt} \tag{5.78}$$

This sum can be converted into an integral by taking the limit of a large number of very thin slices: $n \to \infty$, $\Delta a_i \to 0$, yielding

$$C(x,t) = \frac{C_{iB}}{2\sqrt{\pi Dt}} \int_0^\infty e^{-(x-a_i)^2/4Dt} d\alpha \tag{5.79}$$

Now substituting $(x - a)/2\sqrt{Dt} = \eta$ we can rewrite the integral as

$$C(x,t) = \frac{C_{iB}}{\sqrt{\pi}} \int_{-\infty}^{x\sqrt{Dt}/2} e^{-\eta^2} d\eta \tag{5.80}$$

where

$$da = -2\sqrt{Dt}\, d\eta \tag{5.81}$$

with the minus sign reversing the limits. At $a = 0$,

$$\eta = \frac{x}{2\sqrt{Dt}} \tag{5.82}$$

Table 5.1   Error function values

z	$\mathrm{erf}(z)$	z	$\mathrm{erf}(z)$	z	$\mathrm{erf}(z)$
0	0	0.55	0.5633	1.3	0.9340
0.25	0.0282	0.60	0.6039	1.4	0.9523
0.05	0.0564	0.65	0.6420	1.5	0.9661
0.10	0.1125	0.70	0.6788	1.6	0.9763
0.15	0.1680	0.75	0.7112	1.7	0.9838
0.20	0.2227	0.80	0.7421	1.8	0.9891
0.25	0.2763	0.85	0.7707	1.9	0.9928
0.30	0.3286	0.90	0.7970	2.0	0.9953
0.35	0.3794	0.95	0.8209	2.2	0.9981
0.40	0.4284	1.0	0.8427	2.4	0.9993
0.45	0.4755	1.1	0.8802	2.6	0.9998
0.50	0.5205	1.2	0.9103	2.8	0.9999

and at $a = \infty$, $\eta = -\infty$. This integral is called an error function, and the values are shown in Table 5.1:

$$\mathrm{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-\eta^2} d\eta \tag{5.83}$$

The final solution in readily usable form is

$$C(x, t) = \frac{C_{iB}}{2}\left[1 + \mathrm{erf}\left(\frac{x}{2\sqrt{Dt}}\right)\right] \tag{5.84}$$

at $x = 0$, $C(0, t) = C_{iB}/2$. This form is obtained from the split integrals:

$$C(x, t) = \frac{C_{iB}}{2}\left(\int_{\infty}^{0} \mathrm{erf}(z) + \int_{0}^{x\sqrt{Dt}/2} \mathrm{erf}(z)\right) \tag{5.85}$$

where $\mathrm{erf}(\infty) = 1$, $\mathrm{erf}(-z) = -\mathrm{erf}(z)$, $\mathrm{erf}(0) = 0$. This solution pins the concentration of B at the interface to half of the initial concentration of B, or $C_{iB}/2$.

The case that we solved was on the B side of the AB couple ($x > 0$), and the solution above tracks the decrease in B on the B side. Now using the results obtained, we can explore the A side and the change in B (increase) on that side. The A side is $x < 0$. For this purpose we slightly modify equation (5.84), the $x > 0$ solution for B in B (decreases):

$$C(x, t) = \frac{C_i}{2}\left[1 + \mathrm{erf}\left(\frac{x}{2\sqrt{Dt}}\right)\right] \tag{5.86}$$

At $x = 0$, $c(0, t) = C_i/2$; at $x = \infty$, $c(\infty, t) = C_i$ and C_i refers to the initial concentration value for B. Using this formula, we substitute $x < 0$ or $-x$ and obtain

$$C(x, t) = \frac{C_i}{2}\left[1 - \mathrm{erf}\left(\frac{x}{2\sqrt{Dt}}\right)\right] \tag{5.87}$$

since erf($-z$) = $-$erf(z). The term in square brackets [ ] in equation (5.87) is called the error function compliment, erfc, and it is often separately tabulated.

Now let us consider several other important solutions where the surface or boundary is held at either a constant $C(0, t) = C_S$ with the initial zero concentration of diffusant in the medium, $C_i = 0$, and then where $C_i \neq 0$ where there is already some solute initially in the solid. These more complicated cases are best solved with the use of the Laplace transform.

Mathematical Interlude. The Laplace transform L of a function $f(t)$ for positive values of t is defined as

$$L(p) = \int_0^\infty e^{-pt} f(t) dt \qquad (5.88)$$

where p is sufficiently large to force the integral to converge. For example, if $f(t) = 1$, then

$$L(p) = \int_0^\infty e^{-pt} dt = \frac{1}{p} \qquad (5.89)$$

using the definite integral:

$$\int_0^\infty e^{-ax} dx = \frac{1}{a} \qquad (5.90)$$

Other Solutions for Semi-infinite Solids. Consider a case where the surface concentration is fixed at the outset of the diffusion experiment. We solve equation (5.21):

$$\frac{dC}{dt} = D \frac{d^2C}{dx^2} \qquad (5.21)$$

using Laplace transforms. For $C = C_S$ at $x = 0$ and $t > 0$, and the initial conditions that $C = 0$ at $x > 0$ and $t = 0$, we multiply both sides by e^{-pt} and integrate from 0 to ∞ with respect to t:

$$\int_0^\infty e^{-pt} \frac{dC}{dt} dt - D \int_0^\infty e^{-pt} \frac{d^2C}{dx^2} = 0 \qquad (5.91)$$

If we interchange the order of integration and differentiation and assume that

$$C' = \int_0^\infty C_S e^{-pt} dt = \frac{C_S}{p} \qquad (5.92)$$

Then the second term on the left is given as d^2C'/dx^2, and the first term on the left can be integrated by parts to yield pC' as

$$\int u \, dv = uv - \int v \, du \qquad (5.93)$$

and

$$\int_0^\infty e^{-pt} \frac{dC}{dt} dt = \left[Ce^{-pt}\right]_0^\infty + p\int_0^\infty Ce^{-pt} dt = pC' \tag{5.94}$$

Then Fick's second law is transformed into an ordinary differential equation:

$$D\frac{d^2C'}{dx^2} = pC' \tag{5.95}$$

with the solution:

$$C' = \frac{C_S}{p} e^{-\sqrt{px/D}} \tag{5.96}$$

For equation (5.96) C' remains finite as x goes to ∞. From a table of Laplace transforms the function whose transform is given by this expression for C' is

$$C(x,t) = C_S \text{erfc} \frac{x}{2\sqrt{Dt}} \tag{5.97}$$

For the case where C_S is again constant but $C_i \neq 0$, by the same methods the following solution is obtained:

$$\frac{C(x,t) - C_S}{C_i - C_S} = \text{erf} \frac{x}{2\sqrt{Dt}} \tag{5.98}$$

Now for the case of $C_i = 0$ as was treated above, the same erfc formula is obtained for any C_S. For the case of $C_S = 0$, there is out diffusion of the background diffusant, C_i, with time. It is given as

$$C(x,t) = C_i \text{erf} \frac{x}{2\sqrt{Dt}} \tag{5.99}$$

Since erf $z = 0$ for $z = 0$ and erf $z = 1$ for $z = \infty$, at any position within the solid, x, $C(x, t)$ is decreasing as t increases.

5.5.2.2 Long Time Solution—Homogenization

The last case discussed is homogenization, or a long time solution. Another usual way to solve a differential equation of the form of Fick's second law is to separate variables. Any solution of this equation must involve both x and t, so we can write $C(x, t)$ in general form as

$$C(x,t) = F(x)G(t) \tag{5.100}$$

If we then use this in Fick's second law (equation 5.21), we obtain

$$\frac{1}{G}\frac{dG}{dt} = \frac{D}{F}\frac{d^2F}{dx^2} \tag{5.101}$$

which separates the variables. Both sides must be equal to the same constant, which for convenience we can denote as $-\lambda^2 D$. Now we can write and solve the separate differential equations:

$$\frac{1}{G}\frac{dG}{dt} = -\lambda^2 D$$

Solution: $G = e^{-\lambda^2 Dt}$

(5.102)

$$\frac{D}{F}\frac{d^2 F}{dx^2} = -\lambda^2 D$$

Solution: $F = (A\sin\lambda x + B\cos\lambda x)$

(5.103)

Combining these individual solutions, we obtain

$$C(x,t) = e^{-\lambda^2 Dt}(A\sin\lambda x + B\cos\lambda x)$$

(5.104)

where A and B are constants. A more general solution would be a sum of solutions:

$$C(x,t) = \sum_{m=1}^{\infty} e^{-\lambda_m^2 Dt}(A_m \sin\lambda_m x + B_m \cos\lambda_m x)$$

(5.105)

where all constants are determined by the initial and boundary conditions.

In order to use this solution, a set of specific initial and boundary conditions need to be specified and from which the constants (A_m, B_m) are obtained. We do not treat this further except to point out that the solution appears to be the superposition of the periodic waves of concentrations. For example, to understand the diffusion into a slab of material, we can imagine it as a peak in concentration at one side that decreases to a nearly homogeneous state in time for a fixed amount of diffusant or by holding the surface at a fixed concentration.

5.5.2.3 Diffusion Length

It is often useful in thinking about solid state diffusion problems to be able to estimate how far an atom will travel in a certain time or how long will it take to diffuse a certain distance. That problem is straightforward to solve if we consider the change in shape of a Gaussian with time. With the help of Figure 5.9 that shows the evolution of a Gaussian over time where $\int\alpha dx$ = constant, we calculate the distance between the plane defined at $x = 0$, $C = C_{max}$ and the plane where $C = 1/e(C_{max})$. The distance from $x = 0$ to this plane is called the diffusion length.

We start from the expression

$$C(0,t) = \frac{\alpha}{2(\pi Dt)^{1/2}}$$

(5.106)

This is the solution to equation (5.67) at the plane where $x = 0$. Clearly, C decreases as $f(1/\sqrt{t})$, but from conservation of matter $\int Cdx$ = constant. The distance to the $1/e(C_{max})$ plane increases as $\sqrt{t}$. At $x = 0$,

$$C(0,t) = \frac{\alpha}{2(\pi Dt)^{1/2}}$$

(5.106)
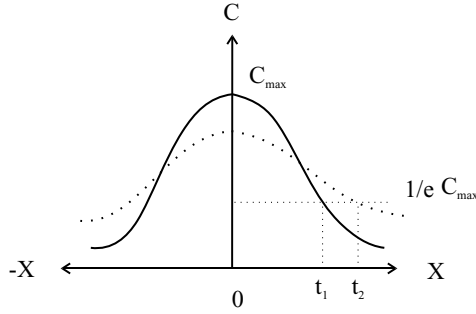
Figure 5.9 Evolution of a Gaussian profile over time (*solid* to *dashed* profiles). A concentration point on the original profile move from t_1 to t_2 in the time difference.

and

$$\ln[C(0, t)] = \ln \alpha - \ln 2 - \frac{1}{2}(\ln(\pi Dt)) \tag{5.107}$$

At x,

$$C(x, t) = \left(\frac{\alpha}{2\sqrt{\pi Dt}}\right)\exp\left(\frac{-x^2}{4Dt}\right) \tag{5.108}$$

Thus we obtain

$$\ln[C(x, t)] = \ln \alpha - \ln 2 - \frac{1}{2}(\ln(\pi Dt)) - \frac{x^2}{4Dt} \tag{5.109}$$

The conditions for the solution are

$$\text{at } x = 0, C = 1 \quad \text{and} \quad \text{at } x, C = \left(\frac{1}{e}\right)1 \tag{5.110}$$

Then using relations (5.110) and subtracting equation (5.109) from equation (5.107), we obtain

$$\ln 1 - \ln(1/e) = \frac{x^2}{4Dt} \tag{5.111}$$

and

$$0 + 1 = \frac{x^2}{4Dt} \tag{5.112}$$

which yields the final result:

$$x = 2\sqrt{Dt} \tag{5.113}$$

This formula teaches that the distance traveled is proportional to $\sqrt{t}$. So, if we want to measure (using chemical analysis) the distance B traveled into A at various times, we can plot the data as x versus $\sqrt{t}$. If diffusion is the mechanism for mass transport, a straight line will result and yield D as the slope.

RELATED READING

R. J. Borg and G. J. Dienes. 1988. *An Introduction to Solid State Diffusion*. Academic Press, San Diego. Similar to Shewmon's classic but with many modern topics and examples.

R. Ghez. 2001. *Diffusion Phenomena*. Kluwer Academic, Dordrecht. More advanced that the Shewmon book and the Brog and Dienes book but readable and with excellent insights into modern diffusion problems in science and technology.

J. Crank. 1975. *The Mathematics of Diffusion*. Clarendon Press, Oxford, England. A treasure trove of solved diffusion problems with all the relevant mathematics.

P. G. Shewmon. 1963. *Diffusion in Solids*. McGraw-Hill, New York. A classic book in diffusion for materials scientists.

EXERCISES

1. C is diffused into a tube of Fe at the rate of 3.6 g/100 h. The tube has an i.d. = 0.86 cm, an o.d. = 1.11 cm, and a length of 10 cm. The variation of C with radius is given below. Calculate D and determine if D is dependent on concentration.

r (cm)	wt% C
0.553	0.28
0.540	0.46
0.527	0.65
0.516	0.82
0.491	1.09
0.479	1.20
0.466	1.32
0.449	1.42

2. A radioactive Cu thin film was deposited onto a pure Cu bar. After an isothermal anneal for 20 hours, the radioactive Cu was determined at various depths in the Cu bar. Determine D.

Activity (counts/min-mg)	Average distance (10^{-2} cm)
5000	1
4000	2
2500	3
1500	4
500	5

3. For a diffusion of 10^{15} radiotracer Cu atoms into Cu at 800°C to a diffusion length x_D of 10^{-5} cm, using $D_0(Cu) = 0.16 \text{ cm}^2/\text{s}$ and $E_D = 2.07 \text{ eV}$ calculate the diffusion coefficient D and time t.

4. You find that the diffusion coefficient D of boron is a factor of 10 greater than that of As at 1150°C and that the solid solubility (assume equal to the surface concen-

tration) of As is a factor of 10 higher than that of boron: $D_B = 10D_{AS}$; $C_0(As) = 10C_0(B)$. For the same diffusion time, t:

(a) Which species, boron or As, has the largest diffusion length?

(b) Which has the largest number of atoms/cm^2 diffused into Si?

(c) What would lead to the greatest change in diffusion length: a 20% change in temperature, time, or surface concentration?

5. Show that

$$C(x,t) = \frac{\alpha}{2\sqrt{\pi Dt}} e^{-x^2/4Dt}$$

is the solution to Fick's second law for the thin film problem.

6. Consider an alloy with a uniform concentration of X of 0.25 wt% ($D = 1.6 \times 10^{-4}$ m/s at 950°C). If the alloy is then treated at 950°C with a gas that brings the concentration of X to 1.2 wt%, calculate how long it would take to achieve 0.8 wt% AT 0.5 mm below the surface of the alloy.

7. P from a deposit was diffused into Si at 1000°C for 2 hours, resulting in a junction depth of 0.537 microns. Estimate D.

8. (a) Atoms of Ag have migrated a distance x in time t in a solid as given below. Assume that the migration was by Fickian diffusion, estimate D.

(b) Using your calculations from (a), justify that diffusion was a good assumption for the mechanism.

x (cm)	t (s)
0	0
0.002	10^2
0.2	10^6

9. (a) A thin plastic membrane is used to separate H from a carrier gas stream. The concentration of H on one side of the membrane is constant at 0.025 mol/m^3 while on the other side it is also constant at 0.0025 mol/m^3. The membrane is 100 μm thick. Given that the flux of H is 2.25×10^{-6} mol/m^2-s, calculate D for H from these conditions.

(b) Suppose that initially the concentration of H was zero in the membrane but built up to 0.0025 mol/cm^3 at 10 μm from the surface in 1 minute, calculate D.

(c) Explain how you can test whether the process in the plastic membranes was a diffusion process or permeation.

10. A Si crystal is put in contact with In vapor, in order to diffuse In into Si so that at the depth of 10^{-3} cm the concentration of In is half of the surface concentration. For $D_{In} = 8 \times 10^{-12}$ cm^2/s at 1600 K, how long must you heat Si in contact with In vapor to perform this diffusion.

11. An atom makes 10^{10} random jumps of 10^{-8} cm each and each with identical probability. Determine how far from the starting point the atom travels.

12. Metal A is diffused into metal B at 1300°C (B also diffuses in A). D_0 and E_D for A in B is 8 cm^2/s and 80 kcal/mol, respectively. Calculate D for A in B at 1300°C, and

then find how long it will take for the concentration of A in B at 0.001 cm from the junction to be half of the surface concentration of A.

13. An isotope of a metal was coated onto a bar of the same metal that was initially free of the isotope. After heating for 1 hour at 1000°C, the isotope was found to be at a relative concentration of 4×10^{-5} at 10^{-3} cm below the surface, and 2.3×10^{-6} at 8×10^{-3} cm. Calculate D.

PHASE EQUILIBRIA

6.1 INTRODUCTION

Phase equilibria deals with the existence of phases, namely what phases exist when equilibrium occurs. For a materials scientist a knowledge of phase equilibria provides a road map to the chemical system(s), the ingredients, the states of matter with which one is dealing. It is particularly important for work on a new material or materials system to first peruse what is known about the phase equilibrium in the system under study.

6.2 THE GIBBS PHASE RULE

6.2.1 Definitions

The Gibbs phase rule (or just phase rule) establishes the relationship among phases and components and intensive variables. As is usual and crucial to understanding thermodynamics, we commence with a set of definitions.

An intensive variable is a variable that does not rely on the total amount of material present such as pressure, temperature, energy per mole, enthalpy per mole, acceleration due to gravity, composition fraction, and refractive index. This is contrasted with extensive variables that do depend on the total amount of material such as: mass, weight, volume. The number of intensive variables is designated herein as F. F is sometimes called the number of degrees of freedom, since its numerical value is the number of intensive variables that need to be specified to describe a particular equilibrium situation. A phase, P, is a homogeneous, physically distinct, and mechanically separable portion of material with a composition and structure. A component, C, is a specific kind of atom or

molecule. The Gibbs phase rule to be derived below establishes a relationship among F, C, P, and other intensive variables (typically temperature T and pressure p).

The definitions for P and C break down for small quantities, namely on an atomistic scale, as do other thermodynamic variables. As thermodynamic variables, P and C apply to large numbers of atoms or molecules. Strictly speaking, it is improper to use the phase rule when describing atomic sized collections and even nanometer sizes. Indeed, most of equilibrium thermodynamics is not reliable in dealing with small systems. However, as a first approximation it is useful, and often done, in materials science to commence describing the existent phases using the anticipated thermodynamic values for small systems based on the large system values.

A phase diagram is a plot of the intensive variables, such as free energy, pressure (p), temperature (T), and mole fraction (X), and this plot is a map of the phases that exist for the materials system. At equilibrium the most stable phase has the lowest free energy. Figure 6.1 is a phase diagram for the single component ($C = 1$), water. First, it is seen that under the excursion of the p and T intensive variables water displays three distinct phases: solid, liquid, and gas. Recall that in Chapter 2, Figure 2.1b is also a phase diagram for water but over a much more extensive p and T range. In this figure water was shown to have a variety of different phases, but all in the same solid state of matter and this is termed "polymorphism," which refers to different phases of the same composition and usually in the same state of matter. In Figure 2.1b each of the solid crystalline water phases depicted has a distinct crystal structure. Figure 6.1 is focused on that region of the water p, T phase diagram at or near $p = 1$ atm that is of interest to people on earth. Notice first that the lines on the diagram are essentially two phase boundaries at which two phases coexist at the p, T conditions specified by the exact position on the line. There is one point on the p, T diagram labeled A where three phases coexist. Needless to say, this point is called a triple point. This point is also called an invariant point because there
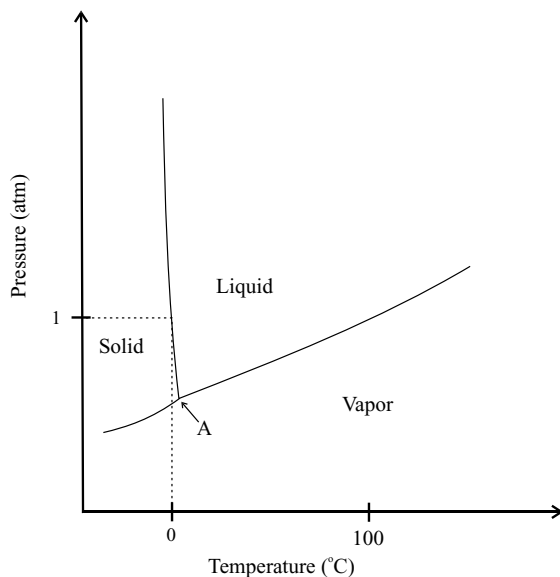
Figure 6.1   Pressure–temperature phase diagram for water.

is only one set of conditions at which the three phases coexist, $p = 0.006$ atm, $T = 0.0075°C$. Next we turn attention to another $C = 1$ phase diagram in Figure 6.2 for sulfur. Once again, all three states of matter are represented and the solid sulfur has two phases, rhombic and monoclinic sulfur. There are two triple points at A and B, and at pressures higher than D the monoclinic phase no longer exists at least as far as is shown for the conditions on this diagram. Last we look at the Fe p, T phase diagram in Figure 6.3 and see three distinct polymorphs of solid Fe and three triple points at A, B, and C. So even with single component materials a diverse phase behavior can be realized at earth ambient conditions. Before proceeding to more complicated multicomponent ($C > 1$) phase diagrams, we return to the Gibbs phase rule and derive it.

The phase diagrams used in the following discussions are only approximations for instructional use, and accurate diagrams based on current data are available in compilations and in the original literature.

6.2.2 Equilibrium Among Phases—The Phase Rule

The Gibbs phase rule relates F, P, and C and enables the determination of the minimum value of F for equilibrium. The derivation commences with the consideration of a series of phases, $P(1) \ldots P$ as shown in Figure 6.4. Each phase has components $C(1) \ldots C$. Imagine a flask containing oil and water that separates into two immiscible phases. Now mix in several components each of which has some solubility in both oil and water. If the two phases are in contact so that all the gradients are reduced to zero, then equilibrium in T, p and chemical potential, μ, will obtain. Indeed, for each component, C, the
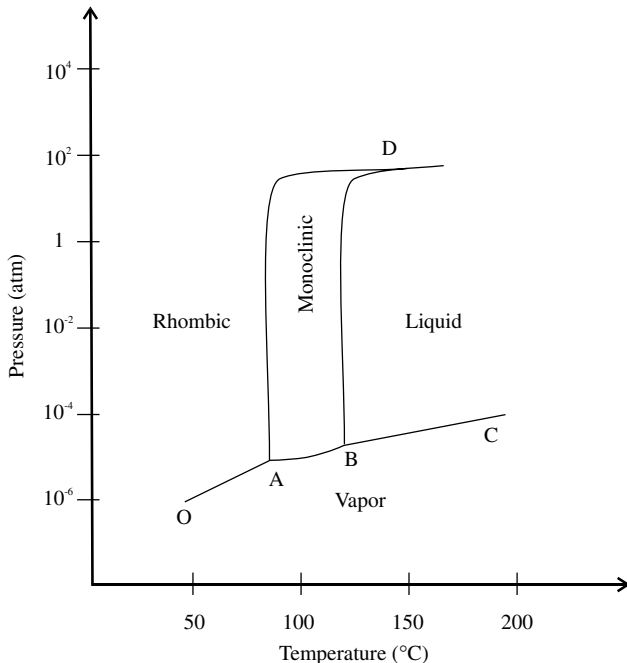
Figure 6.2   Pressure–temperature phase diagram for sulfur.

Figure 6.3 Pressure–temperature phase diagram for iron.

Figure 6.4 *P* phases each with *C* components in equilibrium.

chemical potential will be the same among the *P* phases in which the component appears. At equilibrium

$$\mu_1(1) = \mu_1(2) = \ldots = \mu_1(P)$$
$$\mu_2(1) = \mu_2(2) = \ldots = \mu_2(P)$$
$$\vdots \qquad \vdots \qquad \qquad \vdots$$
$$\mu_C(1) = \mu_C(2) = \ldots = \mu_C(P)$$

(6.1)

for each component 1, 2, . . . , *C*. Now Figure 6.4 depicts *P* phases in intimate contact with each phase having each of the *C* components. It should be understood that in each of the phases the *C* components do not have to be in the same concentration as in the

other phases. Indeed, the components will adjust in the phases according to their individual solubilities. It is required that for equilibrium the chemical potential for each component be the same in each phase. For our oil and water case there are two phases, each with an equilibrium fraction of the C components. In general, for each phase the mole fractions, X_i, sum to 1 and are given as

$$\Sigma X_i = 1 \tag{6.2}$$

In each phase because the sum of the mole fractions is unity only $C - 1$ compositional variables are required (the last is known by difference) in a phase. Thus the total number of compositional variables for all P phases is

$$P(C - 1) \tag{6.3}$$

For each C, equilibrium obtains at the phase boundaries. If we assume that all P phases are in intimate contact so that each C can equilibrate among all P phases, then there are $P - 1$ equilibria, or one for each C at each boundary. Hence the number of equilibrium relations is

$$C(P - 1) \tag{6.4}$$

which is $P - 1$ equilibria for each C. If we now include the common situation of T and p (note that we use P for phases and p for pressure, for clarity) as variables that need to be specified, the total number of intensive variables is given by the number of variables minus the number of relations among the variables:

$$F = P(C - 1) + 2 - [C(P - 1)] \tag{6.5}$$

$$F = PC - P + 2 - CP + C \tag{6.6}$$

$$F = C - P + 2 \tag{6.7}$$

If, as is often the case, we establish the equilibria on the surface of the earth in our laboratory, the pressure is fixed (at near 1 atm). Then the 2 for T and p is reduced to 1 for only T (p is specified). Likewise, if we do work on another planet where there is a different gravity, and/or perhaps work around a strong magnetic field or wherever more intensive variables require specification, then the number 2 will increase appropriately. Since we will consider that the phase equilibria are established on earth in a normal laboratory, we usually need only to specify T. Hence the form of the phase rule form commonly used is

$$F = C - P + 1 \tag{6.8}$$

6.2.3   Applications of the Phase Rule

We now consider the pressure (p) versus T phase diagram for H_2O shown in Figure 6.1 in which p is a variable and thus use $F = C - P + 2$. We see that $C = 1$, since we have only H_2O. In this figure the lines represent one degree of freedom $F = 1$. The lines indicate two phases that coexist at the specific point on a line so $P = 2$. Thus $F = 1 - 2 + 2 = 1$. Another way to view this is that to stay on a line, one needs to specify either T or p, and

the other is determined. However, off the line in the areas between lines there is a continuous range of T and p even when one is specified. In these single phase areas $C = 1$ and $P = 1$, and thus $F = 2$. Notice the intersection of lines at the point $p = 0.006$ atm and $T = 0.0075°C$. As was previously mentioned, this point (A) is called an invariant point, since it exists as a point in intensive variable space and is therefore completely specified. Unlike a line, no variation is possible that can maintain the coexistence described, namely the coexistence of three phases at a point. The invariant point has no degrees of freedom: $F = 0$, and is therefore completely specified.

For Figure 6.2 the only difference is that there are two invariant points at A and B where three phases coexist and $F = 0$. Also this diagram teaches that if one starts to heat the rhombic phase at any $p < 50$ atm, rhombic S doesn't melt! Rather, it converts to the monoclinic phase that does melt at T greater than about 130°C at p greater than about 10^{-5} atm but at lower pressures the rhombic phase sublimes. The phase diagram also suggests that if one wants to prepare the monoclinic phase at 1 atm, one route could be to heat the S until it melts, cool to the monoclinic phase, and then quench rapidly in liquid nitrogen (95 K) in order to prevent atoms of S from rearranging further. Figure 6.3 for Fe displays three different solid phases for Fe and three invariant points (A, B, C). The diagram shows that starting at 1 atm and 25°C, Fe is in a single phase region αFe ($F = 2$), which upon heating to about 910°C is on the equilibrium line that separates αFe from γ Fe ($F = 1$); converts to γFe, which in turn converts to δFe at about 1394°C; and then melts at about 1538°C. Once again, these high temperature phases of Fe can be obtained at laboratory ambient by quenching.

6.2.4   Construction of Phase Diagrams: Theory and Experiment

6.2.4.1   Theory   Since phase diagrams represent equilibrium conditions, it is possible by equilibrium thermodynamics reasoning to obtain formulas from which to calculate, or at least estimate, equilibrium phase diagrams. First consider that there are three conditions for true equilibrium. A system at equilibrium must be at thermal, chemical, and mechanical equilibria. For two phases, α and β, at equilibrium where α and β are composed of two substances (pure compounds or elements), meaning A and B are in each phase, the following conditions must obtain at equilibrium:

1. $T_\alpha = T_\beta$.
2. $\mu_\alpha(A) = \mu_\beta(A)$ and $\mu_\alpha(B) = \mu_\beta(B)$.
3. $\delta w_{\alpha-\beta} = 0$.

Conditions 1 and 2 are reasonably obvious, but condition 3 requires that reversible work not be done when a component changes between the two phases. For example, if a gas and solid phases are in dynamic equilibrium where gas is changing to solid, and vice versa, this interchange, if at equilibrium, takes place without pdV work in the gas or solid or any other kind of work.

Consider the phase equilibrium between phases of one component such as water in solid, liquid, or vapor or diamond and graphite. The variation of G with p at $dT = 0$ is explored starting with an expression for G, then H, then E as

$$G = H - TS, \quad H = E + pV, \quad E = q + w \tag{6.9}$$

$$dG = dH - TdS - SdT \tag{6.10}$$

$$dG = dE + pdV + Vdp - TdS - SdT \tag{6.11}$$

$$dG = dq + dw + pdV + Vdp - TdS - SdT \tag{6.12}$$

Now with dw indicating work done by the system $dw = -pdV$ and $dq/T = dS$, then dG becomes

$$dG = -SdT + Vdp \tag{6.13}$$

With $dT = 0$ as above,

$$dG = Vdp \tag{6.14}$$

Depending on the phases existing, the nature of this equation can take different forms. For example, for a mole of an ideal gas where $pV = RT$,

$$dG = RTd(lnp) \tag{6.15}$$

which is obtained from dp/p. However, for a relatively incompressible solid with a volume V_S,

$$dG = V_S dp \tag{6.16}$$

Now to return to our two phases α and β, we write

$$dG_\alpha = dG_\beta \tag{6.17}$$

$$dG_\alpha = -S_\alpha dT + V_\alpha dp = dG_\beta = -S_\beta dT + V_\beta dp \tag{6.18}$$

From which upon rearranging, we have

$$\Delta S dT = \Delta V dp \tag{6.19}$$

where $\Delta S = S_\beta - S_\alpha$ and $\Delta V = V_\beta - V_\alpha$. At equilibrium, with $\Delta G = 0 = \Delta H - T\Delta S$, $\Delta S = \Delta H/T$ and combining, we obtain

$$\left(\frac{\Delta H}{T} \right) dT = \Delta V dp \tag{6.20}$$

Then

$$\frac{dp}{dT} = \left(\frac{1}{\Delta V} \right) \frac{\Delta H}{T} \tag{6.21}$$

which is called the Clapeyron equation. If the change in phase is from solid to liquid, then ΔH corresponds to melting or fusion, and ΔV corresponds to difference in the specific volumes for the liquid and solid phases.

Consider a solid phase in equilibrium with a vapor phase where the specific volume of the vapor is much larger that that for the solid ($\Delta V = V_v$). Then the Clapeyron equation can be written as

$$\frac{dp}{dT} = \left(\frac{1}{V_v}\right)\frac{\Delta H_{vap}}{T} \qquad (6.22)$$

If we assume that the vapor behaves like an ideal gas, then we obtain

$$\frac{dp}{dT} = \left(\frac{p}{RT}\right)\frac{\Delta H_{vap}}{T} \qquad (6.23)$$

From equation (6.23) we obtain

$$d(lnp) = \frac{\Delta H_{vap}}{R} d\left(\frac{1}{T}\right) \qquad (6.24)$$

Using this relationship, one can then define the solid–vapor or liquid–vapor boundary in a p–T phase diagram. For the solid–liquid boundary the following equation is applicable:

$$\frac{dp}{dT} = \left(\frac{1}{\Delta V_{SL}}\right)\frac{\Delta H_{fus}}{T} \qquad (6.25)$$

For solid–solid equilibria a similar form is used with the appropriate volume and ΔH. For a more precise calculation the changes in the H values would need to be appropriately corrected using heat capacity (C_P) values. Therefore, if the thermodynamic properties are known, then phase diagrams can be calculated or at least approximated.

6.2.4.2 *Experiment* One traditional method for determining phase diagrams is to measure solidification temperatures during the slow cooling of molten mixtures of various compositions. Figure 6.5 shows the result from a variety of starting compositions of Cu and Ni. For pure Cu (far left) and pure (Ni) the horizontal plateau seen in the cooling curve (T = constant) plateau corresponds to the phase transition

$$Cu(or\ Ni)(L) \rightarrow Cu(or\ Ni)(S) \qquad (6.26)$$
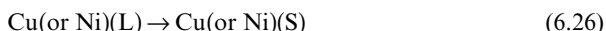
The T at which the phase transition occurs is the melting point, which for a pure crystalline material is a well-defined T at a fixed p. The time duration for the plateau, or width, corresponds to the time necessary for the completion of the phase transition. The more Cu present, the longer it takes. However, during the entire time the T remains constant, since the latent heat of fusion, ΔH_f, is released during solidification. Once the new phase is formed, the ΔH_f associated with the transition is no longer available. Thus cooling of the solid now continues but at a different rate than the liquid because of the differing heat capacities (C_p). Exactly the same physical picture obtains for pure Ni, though the plateau occurs at a different T. However, for all the mixed compositions, the region of the cooling curves between the beginning and end of the more horizontal regions is not level. The slope is also different in the three different regions of the curve:
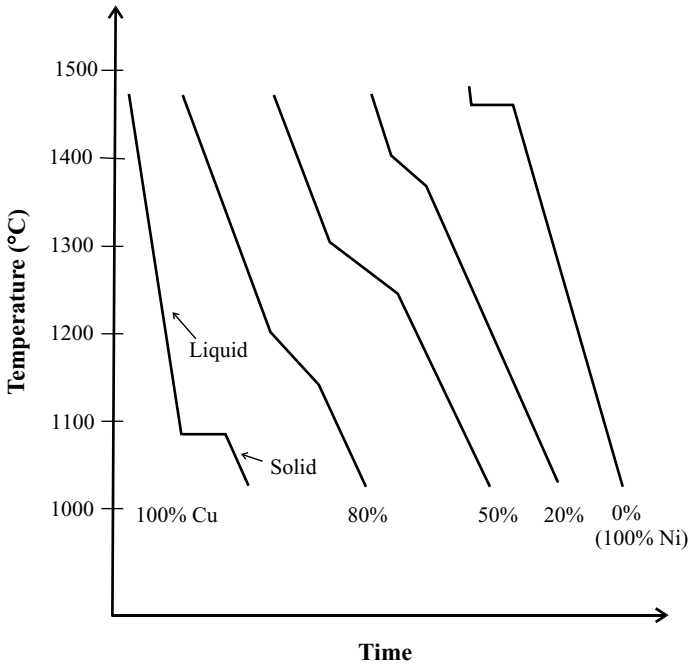
Figure 6.5 Temperature–time data for the cooling of various compositions of CuNi alloys.

before phase change, during phase change, and after phase change. In the phase change region, the mixture has been observed to be composed of two phases: liquid and solid. Figure 6.6 shows the data from Figure 6.5 replotted with the information in terms of T and composition. In any composition that is subjected to cooling, the start of the two phase region from the melt is called the "liquidus" temperature while the end of this two-phase region, where complete solidification occurs, is called the "solidus" temperature. These temperatures form the boundary for the two-phase region where both solid and liquid coexist and vary with the starting compositions. The upper boundary of the two-phase region is called the liquidus line and the lower the solidus line.

Diffraction techniques are useful in determining both the solid crystalline phases that exist and the structure of the phases. For example, a series of mixtures of Cu and Ni can be made, melted, and cooled. X-ray powder diffraction can be used to identify the presence of crystalline phases. Diffraction can also be used at elevated temperature to determine solid to solid phase transitions as was discussed above for S and Fe. Electron microscopy, particularly transmission electron microscopy (TEM) where electron diffraction is obtained, is also routinely used to identify the presence of solid state phases. Metallurgists routinely use optical microscopy in the reflection mode to observe the presence of different phases that have different reflectivities. Samples that contain grains of different phases are carefully polished to remove roughness from affecting the reflectivity. Experienced observers can even identify the phases in well-known materials systems using optical microscopy.

We now explore in more detail the wealth of information contained within phase diagrams.
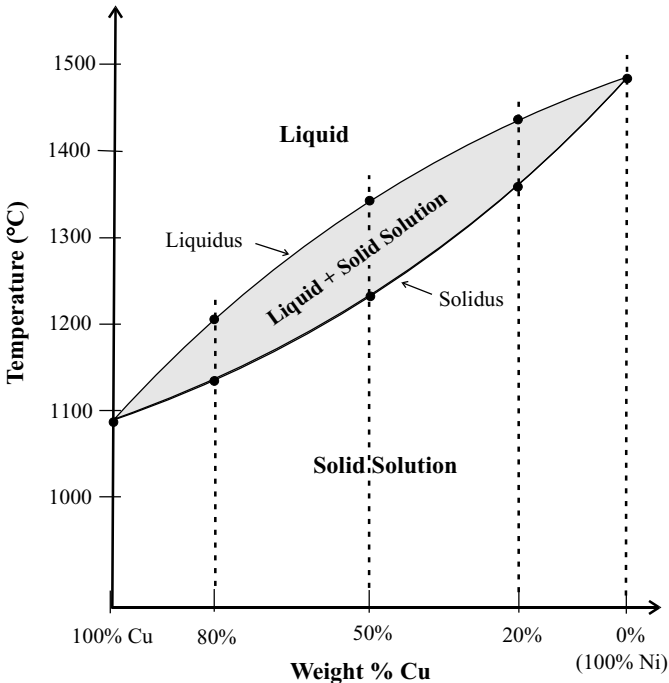
Figure 6.6   Temperature–composition (wt %) phase diagram for the Cu–Ni System.

6.2.5   The Tie Line Principle

A tie line is an isothermal line through a two-phase region. Tie lines are shown as the horizontal dashed lines in the two-phase region of the Cu–Ni T versus composition phase diagram in Figure 6.7. On any tie line the compositions of the two phases in equilibrium (liquid and solid solution) are given by the intersections of the tie line with the solidus C_α (or C'_α) and liquidus C_L (or C'_L) lines. Figure 6.7 illustrates this for an overall composition C_o of 60% Cu by weight, which exists along the vertical dashed line that intersects both tie lines at 60% Cu. Consider that this 60% Cu alloy with Ni is at 1250°C as indicated by C_o located on the top tie line. This tie line intersects the liquidus at about 70% Cu. This means that the liquid phase in equilibrium with the α solid solution is about 70% Cu and 30% Ni (C_L). The intersection of this tie line with the soludus and C_S yields the composition for the solid phase alloy, α, of about 37% Cu and 63% Ni.

If this C_o is cooled from 1250°C to 1200°C, C_o remains the same (the overall composition of the sample is not modified simply by cooling!), but C_L evolves to C'_L and C_S to C'_S, as indicated by the arrows and a new tie line (the bottom dashed line) can be drawn. This new tie line at 1200°C connects C'_L and C'_S. While the overall composition has not changed, the composition of the liquid phase has changed from about 70% Cu to about 82% Cu, and the solid α phase has changed from about 37% to about 58% Cu. So both the liquid and α phase have increased in Cu. This may sound a bit strange at first, but we will see below that this is not a mistake.

In summary, in the two-phase region of a T-composition phase diagram, a tie line can be drawn at any temperature through the composition and temperature desired. This
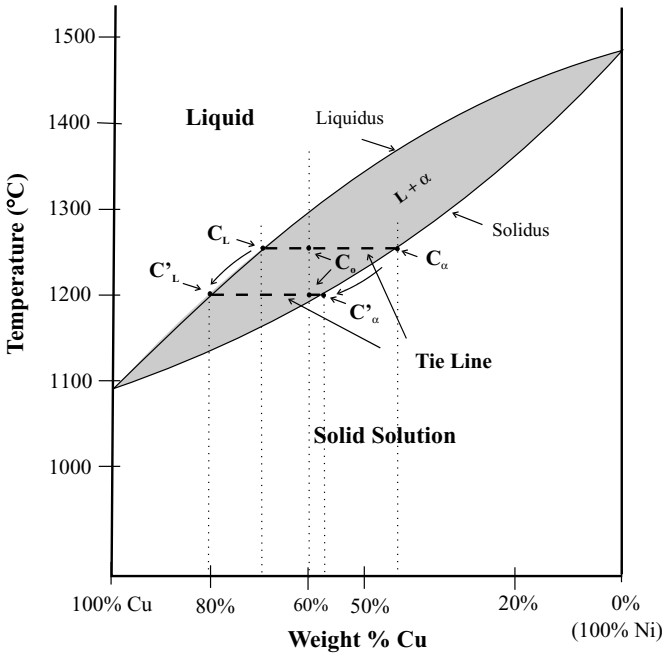
Figure 6.7   Temperature–composition phase diagram for the Cu–Ni system. Tie lines (*dashed horizontal lines*) are shown for cooling a 60% Cu alloy from 1250°C to 1200°C.

horizontal line extends to the ends of the two-phase region. The composition of the liquid phase is given as C_L and the solid phase as C_α, namely the intersections of the tie line with the liquidus and soludus, respectively. At the T of this tie line C_L is in equilibrium with C_α. As cooling takes place, another line parallel to the first will display different intersections with the liquidus and solidus, indicating different liquid and solid phase compositions but with the same overall composition C_o. On any tie line the weight fraction (or composition fraction) of each phase can be obtained in a simple way, and this will be considered next.

6.2.6   The Lever Rule

Before the lever rule is derived, it is useful to state what it is and then the goal of the derivation will be clear at the outset. Essentially the lever rule quantifies the information about the amount of each of the phases present, as was derived from the tie line principle. In Figure 6.8, which is a slightly modified phase diagram from Figure 6.7, it is seen that the segment of either tie line to the left of C_o is labeled S_α and the segment to the right is labeled S_L. We will show below that these segment lengths, S_α and S_L, are respectively proportional to the amount of the α and L phases present which is a statement of the lever rule. We commence the derivation with the definitions. Let

$$f_\alpha = \text{weight fraction of } \alpha \quad \text{and} \quad f_L = \text{weight fraction of L}$$
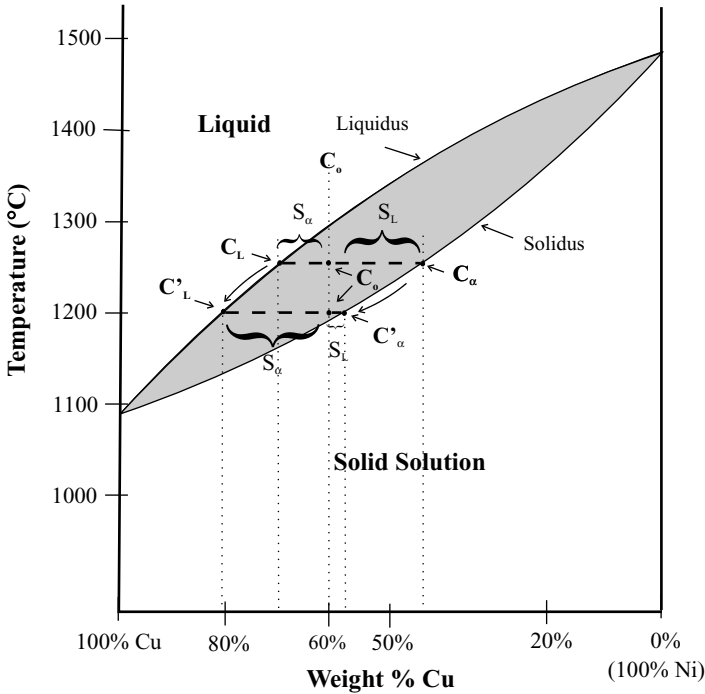
Figure 6.8 Temperature–composition phase diagram for the Cu–Ni system showing tie lines and illustrating the lever rule using tie line segments (*S*'s).

Since the weight fractions of the components add to 1, we can write the relationship

$$f_\alpha + f_L = 1 \tag{6.27}$$

The overall composition C_o is given by the sum of the fraction of amounts of the two phases each multiplied by the composition as

$$C_o = f_\alpha C_\alpha + f_L C_L = f_\alpha C_\alpha + C_L(1 - f_\alpha) \tag{6.28}$$

$$C_o = C_\alpha C_\alpha + C_L - C_L f_\alpha \tag{6.29}$$

Then solving for the fractions, we obtain

$$f_\alpha = \frac{C_o - C_L}{C_\alpha - C_L} \quad \text{and} \quad f_L = \frac{C_\alpha - C_o}{C_\alpha - C_L} \quad \text{or} \quad (1 - f_\alpha) \tag{6.30}$$

Notice that S_α (= $C_o - C_L$) is the part of the tie line farthest from α, that S_L (= $C_\alpha - C_o$) is the part farthest from L, and that the denominator $C_\alpha - C_L$ is the total length of the tie line. Then the f's obtained above are fractions of the tie line, the fraction of α is proportional to the segment of the tie line on the opposite side of C_o, and likewise for the

L phase. Phase diagrams are often displayed in weight fractions especially in the materials engineering community. The lever rule has thus far been defined in terms of weight percent. The lever rule can also be expressed similarly in terms of mole fractions. Consider that Figure 6.8 is plotted in terms of the mole fraction of copper (X_{Cu}) rather than weight fraction or percent. Of course, since weight fraction and mole fraction for a given alloy are not in general numerically the same, the figure would be scaled differently. Ignoring that fact for the present, we can start by defining the total number of moles of the liquid and alloy phases as n_t:

$$n_t = n_\alpha + n_L \tag{6.31}$$

where n_α is the number of moles of the alloy phase and n_L is the number of moles of the liquid phase. Each phase (α and L) has both Cu and Ni. The total amount of Cu, $n_t X_{Cu}$, can be obtained by summing the Cu in both phases as

$$n_t X_{Cu} = n_\alpha X_{Cu} + n_L X_{Cu} \tag{6.32}$$

Also the Cu in α is $n_\alpha X_{Cu}(\alpha)$, and the Cu in the L phase is $n_L X_{Cu}(L)$. Thus

$$n_t X_{Cu} = n_\alpha X_{Cu}(\alpha) + n_L X_{Cu}(L) \tag{6.33}$$

Equating the expressions for $n_t X_{Cu}$, we obtain

$$n_\alpha X_{Cu} + n_L X_{Cu} = n_\alpha X_{Cu}(\alpha) + n_L X_{Cu}(L) \tag{6.34}$$

Now rearranging, we obtain

$$n_\alpha (X_{Cu} - X_{Cu}(\alpha)) = n_L (X_{Cu}(L) - X_{Cu}) \tag{6.35}$$

With the aid of Figure 6.9, which is the same phase diagram as Figures 6.7 and 6.8 except it is in terms of mole fraction, we observe that the expression in parenthesis on the left is the length of the line segment to the right of $X_{Cu}(S_L)$, and that on the right is the length of the line segment to the left of $X_{Cu}(S_\alpha)$. Using these line segments, we can write

$$n_\alpha(S_L) = n_L(S_\alpha) \tag{6.36}$$

and

$$n_\alpha \propto S_\alpha \tag{6.37}$$

This means that the amount of the solid solution phase, α, is proportional to the line segment on the opposite side of the composition of the two-phase mixture (X_{Cu}). Likewise

$$n_L \propto S_L \tag{6.38}$$

So we see that the amount of the liquid phase, L, is proportional to the line segment on the opposite side of the composition of the two-phase mixture (X_{Cu}). It should be noted that Figure 6.9 is nearly the same as Figure 6.8 in terms of the scales for the weight and
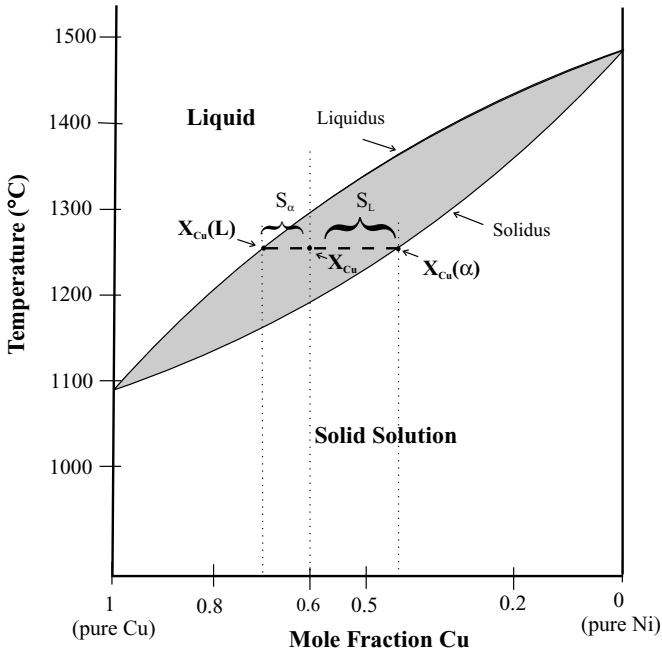
Figure 6.9   Temperature–composition phase diagram for the Cu–Ni system using mole fractions.

mole fractions. However, as was mentioned above, weight fraction and mole fraction are not in general the same. The formula for weight fraction (WF) is

$$WF_i = \frac{\text{Weight}_i}{\Sigma(\text{all weights})} \tag{6.39}$$

and for mole fraction (X) is

$$X_i = \frac{\text{Mass}_i/\text{MW}_i}{\Sigma\,\text{Mass}_i/\text{MW}_i} \tag{6.40}$$

Thus for weights and masses in grams the fractions will differ by the molecular weights for the constituents. For Cu and Ni the molecular (atomic) weights are about 58.7 g/mol and 63.5 g/mol, respectively. For our chosen value for the weight fraction for the Cu/Ni system of 0.60, the mole fraction of Cu is about 0.62. So coincidentally the mass and mole fractions are close in value to each other, and the difference ignored for the purpose of Figure 6.9.

In summary, the tie line principle enables the determination of the composition of the phases, and the lever rule enables the determination of amounts of each phase. There exist a wide variety of phase diagrams. Rather than attempt to survey a large number of cases here, the emphasis will be on general principles that apply to many kinds of phase diagrams. The topics addressed below relating to real phase diagrams include complete

solid solubility, partial solubility, compound formation, homogeneity range, and ternary and pseudobinary phase diagrams.

6.2.7   Examples of Phase Equilibria

The equilibrium phase diagram for Cu–Ni shown in Figure 6.7 illustrates that there is complete miscibility in the solid phase for the Cu–Ni alloy. This means that any composition of Cu in Ni, and vice versa, can be prepared under ambient conditions. Recall from Chapter 2 that an alloy assumes the crystal structure of the host material. The question that arises is, Which constituent is host for a 50–50 alloy? While this seems to be unanswerable, nature obviates the question. It is found that both Cu and Ni are FCC with lattice parameters of 0.3615 nm and 0.3524 nm, respectively. Furthermore Cu and Ni are side by side in the Periodic Table of the Elements, and thus both have similar electronic structure (differing by one electron) and masses. Consequently it is understandable that Cu and Ni can interchange for each other on the FCC lattice and result in complete miscibility. Similarly Figure 6.10 displays the complete solid miscibility phase diagram for Ge–Si. These elements also have the same crystal structure (diamond cubic) and similar lattice constants (0.5658 and 0.5431 nm for Ge and Si, respectively) and are adjacent in the same group in the Periodic Table. While Si's electronic structure had 2 electrons in its outer M shell as $2p^2$ and Ge has 2 electrons in its outer N shell as $3p^2$, the electronic energy band structures are similar. Thus complete miscibility is anticipated.

The complete solid solubility ideas presented above can be contrasted with complete solid immiscibility as shown for the Bi–Cd phase diagram shown in Figure 6.11. Bi
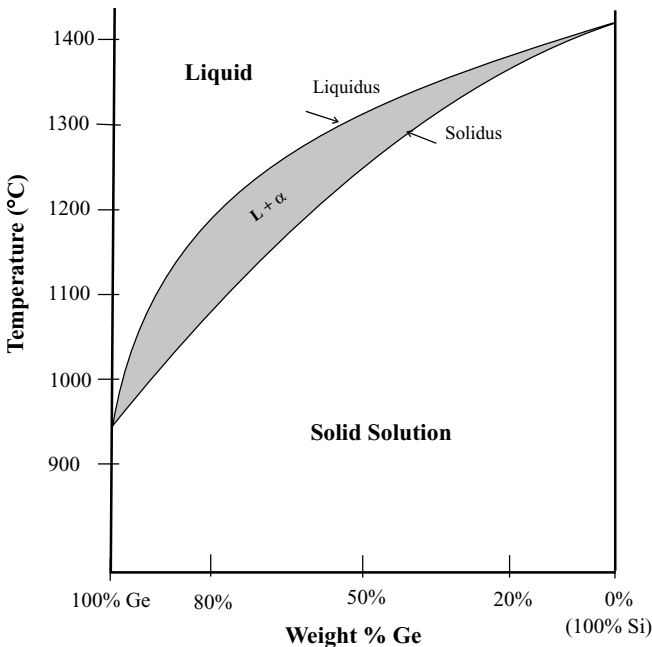
Figure 6.10   Temperature–composition phase diagram for the Ge–Si system.
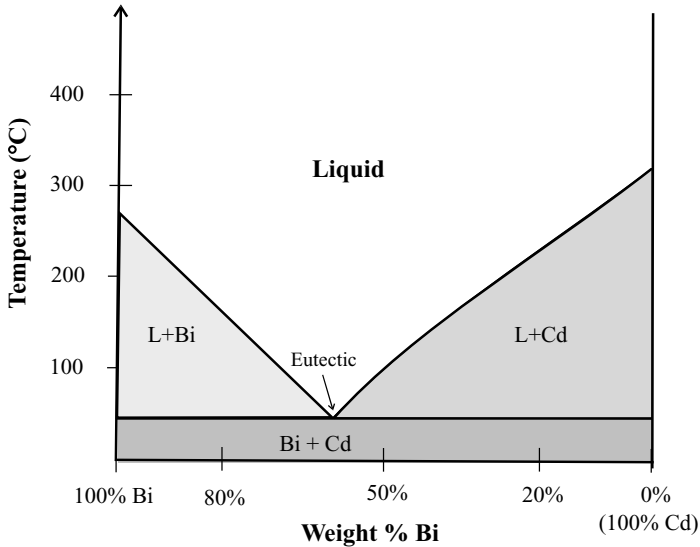
Figure 6.11   Temperature–composition phase diagram for the Bi–Cd system.

(rhombohedral) and Cd (hexagonal) have different crystal structures, widely different lattice parameters (Bi: 0.4736 nm and $\alpha = 57.14°$; Cd: 0.2979 and 0.5617 nm), and different electronic structures, and they are not adjacent (horizontally or vertically) in the Periodic Table. Thus a large degree of immiscibility is anticipated. Also in this phase diagram there is a point where three phases (L, Bi, and Cd) are in equilibrium, a triple or invariant point. In this kind of phase diagram this invariant point is called the eutectic, and it is the lowest temperature at which solidification occurs. If we prepare a liquid of Bi and Cd at the eutectic composition and then slowly cool the liquid, at exactly the eutectic temperature the liquid will convert to Bi and Cd, so the three phases coexist at the eutectic temperature. This reaction can be expressed as

$$L \rightarrow Solid_1 + Solid_2 \qquad (6.41)$$

where in this case the solids are Bi and Cd. For two components and three coexisting phases the phase rule predicts no degrees of freedom, an invariant point:

$$F = C - P + 1 \,(\text{at } P = \text{constant}) = 2 - 3 + 1 = 0 \qquad (6.42)$$

It is also possible to have other invariant points. If a solid phase, α, converts upon cooling to two other solid phases, another invariant point ($F = 0$) would exist. This reaction is given as

$$\alpha \rightarrow \beta + \gamma \qquad (6.43)$$

and is called a eutectoid reaction. The invariant point is called a eutectoid. Also the following reactions involving three phases yield invariant points:

$$\alpha + L \rightarrow \beta \qquad \text{peritectic} \qquad\qquad (6.44)$$

$$\alpha + \beta \rightarrow \gamma \qquad \text{peritectoid} \qquad\qquad (6.45)$$

$$L_1 \rightarrow \alpha + L_2 \qquad \text{monotectic} \qquad\qquad (6.46)$$

While immiscibility is commonplace, total or near total immiscibility is rare, and limited solubility is more often observed. Figure 6.12 shows a limited solubility phase diagram for the Pb–Sn system. On the 100% Pb and the 100% Sn sides of the phase diagram limited miscibility is seen. On the Pb side of the diagram, the solid solution (alloy) that is formed is called the α phase in which the crystal structure is that of Pb with Sn atoms taking substitutional sites on the Pb structure. On the Sn side of the phase diagram, the alloy formed is called the β phase, and is basically the crystal structure of Sn with some Pb atoms on substitutional sites. Pb has the FCC structure and Sn the tetragonal structure. Pb is larger than Sn, so one can expect less Pb in Sn than Sn in Pb, and this is observed with the α phase extending to about 19% Sn in Pb while the β phase extends to less than 6% Pb in Sn. In between these two alloy phases there is immiscibility and a mixture of the immiscible α and β phases is found. The ratio of the composition and amount of each of these phases can be determined at any T up to the eutectic (about 183°C) using the tie line principle and lever rule as was described above.

An example of how the quantitative information is extracted from the phase diagrams is illustrated with the use of the Sn–Pb Phase diagram shown in Figure 6.12. We commence at 250°C for 80/20 = Pb/Sn (see the dashed line in Figure 6.12 with the filled circle), and follow the changes when the system is cooled to 200°C. At 250°C the approximate compositions are $C_\alpha = 87\%$ Pb and $C_L = 67\%$ Pb, as can be obtained from a tie line at 250°C (solid line at 250°C). Now apply the lever rule to determine the amounts:
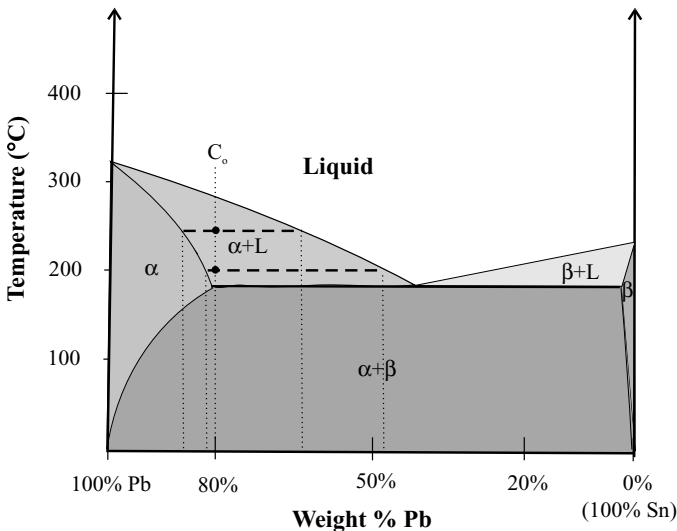
Figure 6.12   Temperature–composition phase diagram for the Pb–Sn system.

$$f_\alpha = \frac{C_L - C_0}{C_L - C_\alpha}$$

with $C_L - C_\alpha = 20$ and $C_L - C_0 = 67 - 80$, then

$$f_\alpha = 65\,\text{wt}\%\,\alpha$$

$$f_L = \frac{C_0 - C_\alpha}{20} = \frac{80 - 87}{20} = 35\,\text{wt}\%\,L$$

Now at 200°C the approximate compositions are

$$C_\alpha = 83\,\text{wt}\% \quad \text{and} \quad CL = 45\,\text{wt}\%\,\text{Pb}$$

$$f_\alpha = \frac{45 - 80}{45 - 83} = 92\,\text{wt}\%\,\alpha$$

$$f_L = \frac{80 - 83}{45 - 83} = 8\,\text{wt}\%\,L$$

Thus, as the T is lowered, the compositions of both α and L move toward enrichment of Sn, the component that melts at lower T.

Starting with 1 g of alloy at 250°C, we have

$$\text{wt Pb} = f_\alpha C_\alpha + f_L C_L = 0.65 \cdot 0.87 + 0.35 \cdot 0.67 = 0.8\,\text{g}$$

For 1 g of alloy at 200°C,

$$\text{wt Pb} = 0.92 \cdot 0.83 + 0.08 \cdot 0.45 = 0.8\,\text{g}$$

Thus the Pb, which is 80% by weight, remains at 80%.

Thus far we have considered cases where the constituents have either formed or not formed an alloy in the solid phase. Now we discuss the case where, instead of forming an alloy or solid solution, a new chemical compound is formed. It should be recalled that this means that a structure that is different from either of the starting components is produced. Figure 6.13 for the Mg–Ni system shows compound formation in addition to the features that we have already seen. A vertical line on a phase diagram is a compound where the width of the line is a measure of the compound's stoichiometry and is called the homogeneity range. For example, a geometric line with zero width is indicative of perfect stoichiometry. However, many compounds have finite homogeneity ranges, as is seen not only for intermetallic compounds as shown in Figure 6.13 but also for many other materials.

Solid solutions occur at both the Mg and Ni sides of the phase diagram yielding the α and β alloy phases, respectively. At about 0.46 and 0.17 weight fraction of Mg, the compounds Mg_2Ni and $MgNi_2$ form, respectively. These compounds are denoted by vertical lines on the phase diagram. A two-phase region exists in between these compounds indicative of their immiscibility. Two eutectics are seen near 0.77 and 0.10 weight fraction Mg, as are indicated by the $L \rightarrow$ solid phase$_1$ + solid phase$_2$ reaction occurring at the eutectic T's for these invariant points. The compound $MgNi_2$ melts above 1000°C to
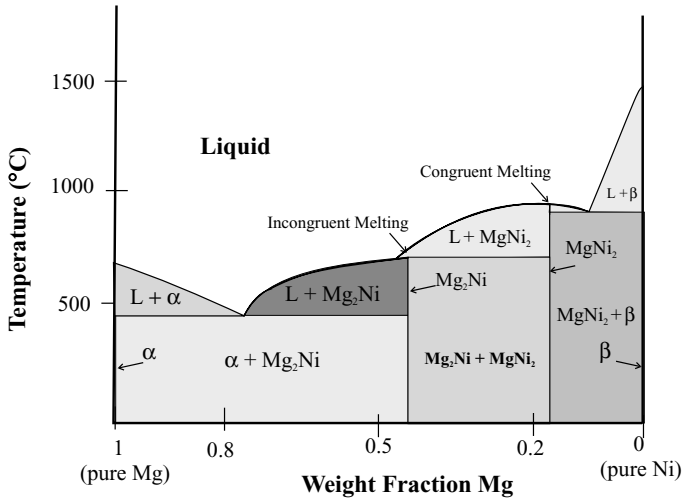
Figure 6.13   Temperature–composition phase diagram for the Mg–Ni system.

liquid of the same composition. This is called congruent melting. However, the compound Mg_2Ni melts above 700°C. However the composition of the liquid has more Mg than is indicated by the stoichiometry of Mg_2Ni, and a solid phase $MgNi_2$ is also produced. A tie line drawn above the melt in the two-phase region indicates the resulting L and solid phase compositions. This kind of melting where the melted compound does not exist in the melt is called incongruent melting.

While the vertical lines that are indicative of compounds appear narrow, a characteristic of stoichiometric compounds, in reality there is always some width to the lines. The linewidth represents a range of stoichiometry, albeit sometimes very small, in which the compound exists. It is possible that some important materials properties can change significantly across the width of the homogeneity range. This fact is illustrated in Figure 6.14 for GaAs, an important electronic and optical material. Note that the compound GaAs forms, and there is no alloy formation of GaAs with either Ga or As. Consequently both sides of GaAs display two phase regions; GaAs melts congruently. The interesting nuance is the width of the vertical line used to indicate the compound GaAs. The line shows measurable width at any temperature and below about 800°C the width is about 0.02 atomic fraction, and this width is mainly on the Ga rich side. This means that a Ga rich GaAs compound is formed. Above 1100°C the width is maximum and about 0.15 atomic fraction and nearly symmetrical about the 0.50 atomic fraction. The Ga rich side of the stoichiometry yields an N-type material, which means that the dominant electrical charge carriers are electrons. Also this N-type material is several orders of magnitude more conductive than a more stoichiometric material. More will be said later about semiconductors in Chapters 9 through 11. Suffice for the present discussion that the homogeneity range is a very important materials parameter to know something about.

The preceding discussion has dealt exclusively with single component and binary phase diagrams. However, phase equilibria in systems that have three components are
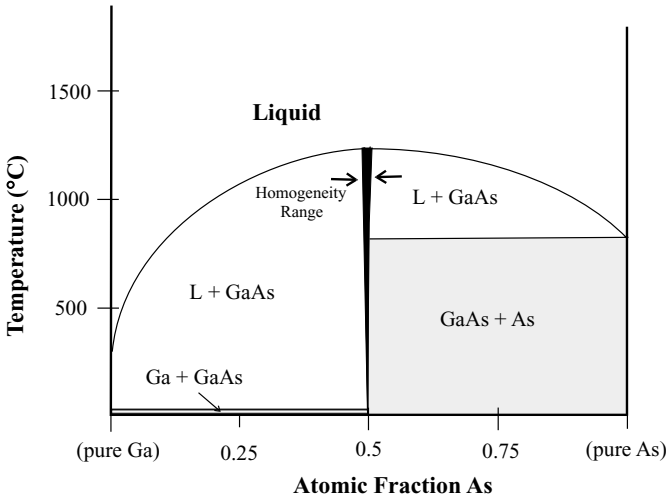
Figure 6.14   Temperature–composition phase diagram for the Ga–As system.

also important. Consider, for example, ceramic materials that are usually combinations of various oxides, and specifically the phase equilibria in the alumosilicate materials system that contain Al, Si, and O, and that can be represented on a ternary phase diagram. Figure 6.15a shows a ternary diagram for the Al, Si, and O components on which the composition map of existent phases can be displayed. In this kind of phase diagram involving three elements, many different structures (phases), compounds, and solutions can exist. For the uninitiated the triangular representation can be complex and difficult to comprehend. Furthermore the two-dimensional ternary diagram is for a single temperature, and even more complicated, the three-dimensional diagram is required to display the temperature information. While a three-dimensional diagram is complete, such a diagram typically displays more information than is required. An experimentalist usually starts to prepare the desired materials in, say, the Al, Si, O system by combining various proportions of powders of Al_2O_3 and SiO_2 that are readily available and easy to handle. The phase diagram in Figure 6.15a can be simplified by considering Al_2O_3 and SiO_2 as the starting components. On Figure 6.15a this situation is indicated by the dashed line that connects these two binary compounds. In between these end points are the phase equilibria than can exist for this limited part of the entire ternary diagram. This line is called a pseudobinary cut of the ternary diagram. Figure 6.15b is a pseudobinary cut of the Al, Si, O ternary diagram with Al_2O_3 and SiO_2 the starting components. The pseudobinary diagram has all the normal features that were previously discussed.

6.3   NUCLEATION AND GROWTH OF PHASES

6.3.1   Thermodynamics of Phase Transformations

An important application of the thermodynamics of phase equilibria is to determine the lowest energy pathway for new phases to appear. One aspect of this general issue of phase transformation is nucleation. Nucleation is the process by which the first vestige of a new
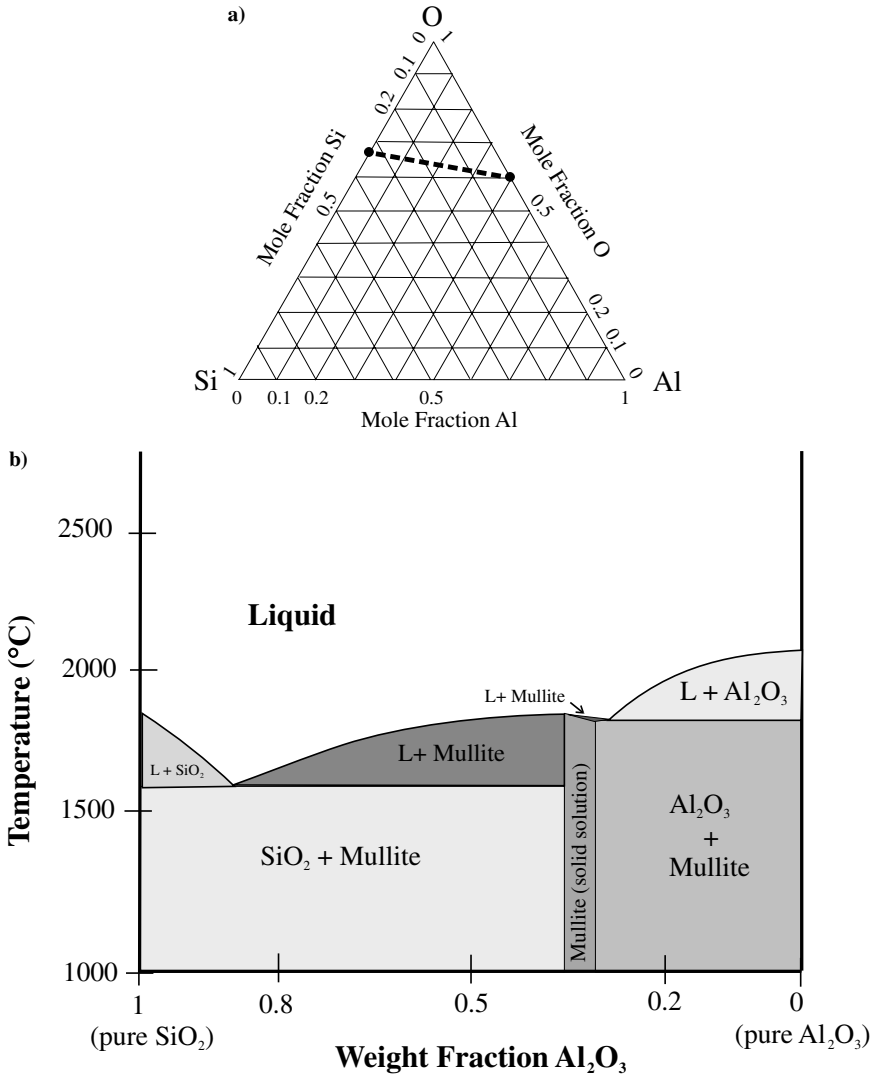
Figure 6.15   (*a*) Triangular diagram used to represent the ternary Al–Si–O system showing the line connecting SiO₂ with Al₂O₃; (*b*) details within the pseudobinary cut from (*a*).

phase appears. The smallest parts of the new phase are called nuclei and are the collected building blocks (atoms or molecules) of the new phase.

We first consider the thermodynamics for the transformation of the α phase to the β phase:

$$\alpha \rightarrow \beta \qquad\qquad (6.47)$$

The very initial appearance of β phase is called nucleation that results in nuclei with a distribution of sizes. Below we will discuss the critical nuclei size as the minimum size

for a stable nucleus. Once a stable nucleus forms, the newly emergent phase can grow and result in the coalescence of the nuclei. For growth of the new phase material, transport to the stable nuclei is crucial.

Consider the two phases, α and β, in equilibrium at temperature T_E. Consider that α is more stable at $T > T_E$ and β at $T < T_E$. At $T = T_E$,

$$G_\alpha = G_\beta \tag{6.48}$$

and

$$\Delta G_V = G_\beta - G_\alpha = 0 \tag{6.49}$$

where G_v is the usual free energy for a volume of material, namely the chemical free energy. This Gibbs free energy is purposely subscripted as the volume free energy in order to distinguish it from the surface free energy discussed below. Recall the following equilibrium relationship:

$$\Delta G_v = \Delta E_v - T_E \cdot \Delta S_v = 0 \qquad \text{where } E_v \approx H_v \tag{6.50}$$

The approximate equivalence of ΔE and ΔH is made for solids where under usual conditions of p, pdV is insignificant. From this we obtain for the entropy

$$\Delta S_v = \frac{\Delta E_v}{T_E} \tag{6.51}$$

Now near the equilibrium T, we have $T \approx T_E$. Using this relation, we obtain

$$\Delta G_v = \Delta E_v - T\left[\frac{\Delta E_v}{T_E}\right] = \frac{\Delta E_v(T_E - T)}{T_E} \tag{6.52}$$

With $\Delta T = T_E - T$, and assuming that both ΔS and ΔE (or ΔH) are not functions of temperature (for ΔS and $\Delta E \neq f(T)$), we obtain

$$\Delta G_v = \Delta E_v \cdot \frac{\Delta T}{T_E} \tag{6.53}$$

ΔT is the driving force for this phase transformation from α to β, and reverse, and it is typically referred to as the super cooling. When $\Delta T \neq 0$ a transformation occurs.

Now we can use these relationships to analyze specific situations. If heat is evolved for $\alpha \rightarrow \beta$, the reaction is exothermic and ΔE_v is negative. Thus at higher T and with $-\Delta E_v$, ΔT will be more negative ($T > T_E$). Then ΔG_v becomes more positive, and the transformation as written ($\alpha \rightarrow \beta$) is less favorable. If, on the other hand, the reaction is endothermic, ΔE_v is positive, the transformation is more favorable at higher T. The direction of the transformation is obtained from this thermodynamic treatment, but at this point nothing is learned about kinetics (the pathway) of the transformation. To this end the specific thermodynamics relating to the formation of the new phase needs to be understood, and this is called the thermodynamics of nucleation.

6.3.2 Nucleation

Figure 6.16 shows the nuclei of β that form from the above-mentioned phase transformation in the previously homogeneous α phase. To calculate the number of nuclei, n^*, we again consider the two-state problem, and the fact that it will cost energy to produce the β phase nuclei. The result is the Boltzmann factor in terms of the energetics of the specific case at hand:

$$n^* = N \exp\left(-\frac{\Delta G^*}{RT}\right) \tag{6.54}$$

where N is the number of sites available for nuclei formation and ΔG^* is the energy that must be supplied for nuclei formation (i.e., the barrier for formation).

We return to the problem of calculating ΔG (and then ΔG^*) for the α → β phase transformation by first assuming the formation of small spherical particles of β in the α phase. Later this assumption about the nuclei shape is justified. As the β phase forms in the α phase, an interface is created between the α phase and the β nuclei. This interface has an associated surface energy, γ, in units of energy per area produced. Figure 6.17 illustrates the origin of this surface energy. If we focus on one atom in the depicted two-dimensional solid, we notice that the atom is bonded to four adjacent atoms. For the purpose of tabulating the energy required to form a new surface, the bulk atoms first need to be freed by breaking the four bonds for each atom and then transporting the freed atom(s) to form the new surface where three bonds reform. From an energy accounting point of view, it costs four units of bond energy to free an atom and three units are returned upon surface bond formation. Thus there is a net expenditure of one unit of bond energy to form a one atom surface in this simple model. The formation of a real surface is, of course, more complicated with structural rearrangements. Nevertheless, this simple model demonstrates that the total amount of energy needed to form a surface is always positive and proportional to the number of atoms on the new surface, hence to the surface area produced. Remember that up to now we have
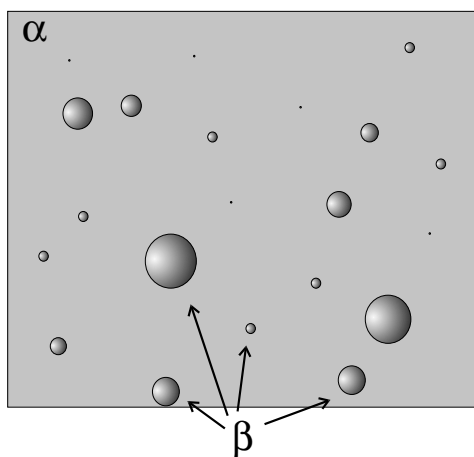
Figure 6.16 An α phase with spherical nuclei of β phase.
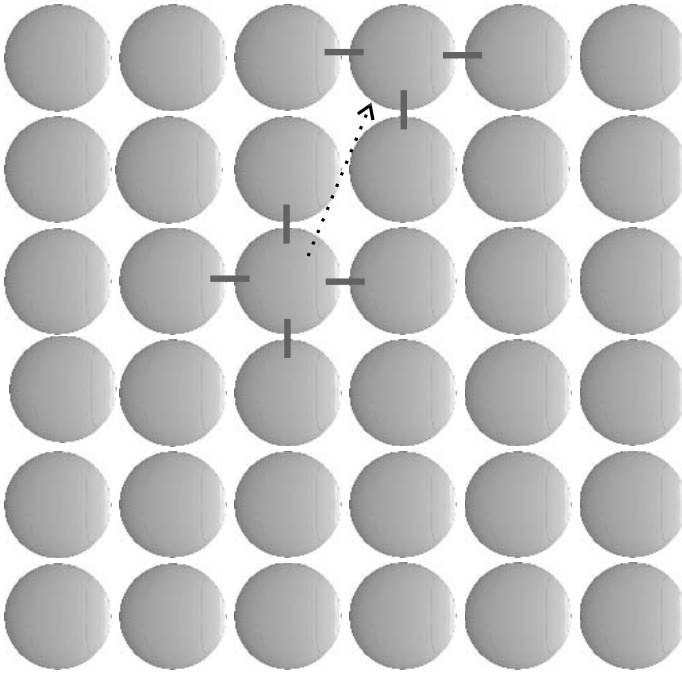
Figure 6.17 Bulk atoms with four bounds and surface atoms with three bonds.

been considering energy per volume and not per area when calculating thermodynamic properties.

We return to consider the formation of a spherical nucleus of β with radius r, and calculate the total energy for this process, ΔG_{tot}, using the following relationship:

$$\Delta G_{tot} = \text{Volume free energy change} + \text{Surface energy change} \qquad (6.55)$$

For a spherical nucleus the volume free energy term is the change in the chemical free energy associated with the transformation, ΔG_v. However, as was mentioned above, this energy is calculated in units of energy per volume. In order to add the volume free energy to the surface term, we need to remove the geometrical part by multiplying by the volume for the spherical nucleus, to yield

$$\left(\frac{4}{3}\pi r^3\right) \cdot \Delta G_v \qquad (6.56)$$

As we saw above, we can adjust the conditions and even the chemistry such that this volume term will be favorable meaning, ΔG_v is negative. Also, as was discussed above, the surface energy term, γ, is energy per area. So before adding to the volume term, it must be multiplied by the spherical nucleus area, to yield

$$\left(4\pi r^2\right) \cdot \gamma \qquad (6.57)$$

This term is always positive because it always takes energy to create a surface. It is now noticed that the sphere exhibits the smallest surface to volume ratio. Hence with γ positive, the spherical shaped nuclei will provide the smallest energy requirement, and hence be the favored shape. We can add the volume and surface energy terms to obtain the total energy:

$$\Delta G_{tot} = \left(\frac{4\pi r^3}{3}\right) \cdot \Delta G_v + \left(4\pi r^2\right) \cdot \gamma \qquad (6.58)$$

The changes in the two energy terms on the right-hand side of equation (6.58) depend on nuclei size, as shown in Figure 6.18. As was discussed above for the volume term, the conditions can be adjusted so that this term is negative and decreases with increasing r. Thus, with the surface term positive and growing more positive with increasing r, it is inferred that the sum of these terms has an inflection point which is labeled as $r*$ in Figure 6.18. For small nuclei sizes where $r < r*$, the surface term is dominant because the square of small quantities is larger that the cube. For the case where $r > r*$, the negative volume term dominates and pulls the transformation process to the right. This figure teaches that a positive ΔG_{tot} results for small nuclei ($r < r*$) and negative ΔG_{tot} for larger nuclei ($r > r*$). This means that the small nuclei are thermodynamically unstable relative to larger nuclei. At $r*$ the barrier, $\Delta G*$, is obtained, and $r*$, it is calculated below.

Before proceeding further with this development, it is useful to try to imagine what is occurring and why. It is straightforward to imagine atoms or molecules moving around and colliding with some of the collisions of α phase species converting to β phase, since we have already made this probable by adjusting the conditions to yield a negative ΔG_v. The initial result is that a small number of at first small nuclei of β form. If these nuclei are smaller than the critical size ($r < r*$), they are unstable, and there is a good chance that before they can grow larger via more of the fruitful collisions, they will disappear to reform α phase. Statistically even some of the small nuclei will exist long enough to grow. Those lucky few of the $r < r*$ nuclei that grow become more stable, and their chances of growing even larger steadily improves. We can imagine that these events are
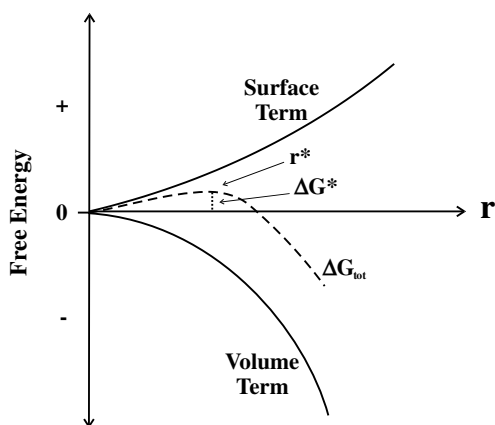
Figure 6.18 Free energy with surface and volume terms plotted versus the size of a spherical nucleus.

occurring with a huge number of atoms and/or molecules. So ultimately the number of statistical survivors is measurable, and we observe the formation of the β phase. This statistical notion is useful, since it teaches that many events are occurring on a microscopic scale. Some of the events are favored and some are not. Yet they occur because there are so many reactants. Even low probability events can occur, however infrequently. Ultimately the events that lead to observable stable phases are the favored events. Nature selects these via this statistical process, and thermodynamics (its laws) typically concerns the initial and final states of a large number of reactants and products, in order to tally the overall energetics.

Now returning to r^* at the inflection point, r^* is obtained at $d\Delta G_{tot}/dr = 0$ as follows:

$$\frac{d\Delta G_{tot}}{dr} = \frac{d\left(\left(4\pi r^3/3\right)\cdot\Delta G_v\right)}{dr} + d\left(4\pi r^2\cdot\gamma\right)dr = 0 \tag{6.59}$$

where $r = r^*$. Then solving for r^* and using the result above $\Delta G_v = \Delta E_v\cdot\Delta T/T_E$, we obtain

$$r^* = \frac{-2\gamma}{\Delta G_v} = -2\gamma\left(\frac{T_E}{\Delta E_v\cdot\Delta T}\right) \tag{6.60}$$

With this value for r^* in the formula for ΔG_{tot} the following is obtained:

$$\Delta G_{tot} = \left(\frac{4\pi r^{*3}}{3}\right)\cdot\Delta G_v + 4\pi r^{*2}\cdot\gamma \tag{6.61}$$

The final result for the activation energy for nucleation is

$$\Delta G^* = \left[\frac{16\pi r^2\cdot T_E^2}{3\Delta E_v^2\cdot\Delta T^2}\right] \tag{6.62}$$

The rate of nucleation dn^*/dt is proportional to the probabilities for nucleation and to mass transport, which is often by diffusion (where E_D is the activation energy for diffusion). It given as

$$\frac{dn^*}{dt} \propto \exp\left(\frac{-\Delta G^*}{kT}\right)\cdot\exp\left(\frac{-E_D}{kT}\right) \tag{6.63}$$

The rate of growth can then be given as

$$\text{Rate of growth} \propto \frac{dn^*}{dt}\cdot\exp\left(\frac{-E_D}{kT}\right) \propto \exp\left(\frac{-\Delta G^*}{kT}\right)\cdot\exp\left(\frac{-E_D}{kT}\right)\cdot\exp\left(\frac{-E_D}{kT}\right) \tag{6.64}$$

As the supercritical nuclei grow, they ultimately will encounter adjacent nuclei. A process of ripening then occurs where nuclei coalesce to form a smaller number of larger nuclei. This is shown in Figure 6.19. The driving force for ripening is simply that the larger nuclei have less surface area relative to their volume (smaller surface/volume ratio) and thus is a lower energy configuration. If this ripening and coalescence occurs on a surface, the regions in between nuclei can offer sites for sec-
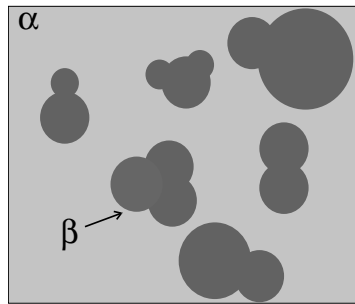
Figure 6.19   Coalescence of β nuclei in α phase.
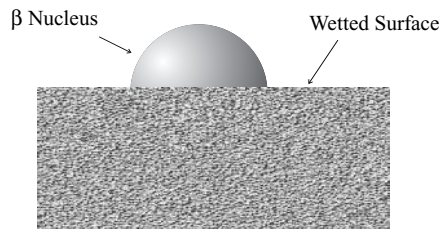
Figure 6.20   A β nucleus forming on a surface.

ondary nucleation, and ultimately the original surface will become completely covered with the new phase.

Up to now we have only considered homogeneous nucleation, which is where a nucleus forms in three dimensions and is specifically a spherical nucleus. However, it is well known that heterogeneous nucleation, for example, nucleation on a surface, occurs more readily. While the specific situation for heterogeneous nucleation needs to be considered, it is often due to the fact that less γ is required in the presence of another surface. For example, we considered above the formation of three-dimensional free standing spherical nuclei. The spherical shape minimized the surface–volume ratio and was favored. However, in the presence of a surface, a hemisphere can form and further lower the surface area, as shown in Figure 6.20. Thus it can be the case that heterogeneous nucleation is favored. Examples of this abound. Consider the cooling of water. Water cools to well below freezing, but ice does not form until a dust particle is present or the container surface is scratched exposing two-dimensional surface nucleation sites. In general, two-dimensional nucleation is complicated because there is the interaction of the surface with the nucleus, which may or may not be suitable for nucleus stability and which may dictate the nucleus shape and hence its energy. This is known as wetting behavior and will not be treated further here.

RELATED READING

P. A. Thornton and V. J. Colangelo. 1985. *Fundamental of Engineering Materials*. Prentice Hall, Englewood Cliffs, NJ. A readable elementary text for a first course in materials science.

C. R. Barrett, W. D. Nix, and A. S. Tetelman. 1973. *The Principles of Engineering Materials*. Prentice Hall, Englewood Cliffs, NJ. A readable elementary text for a first course in materials science.

A. N. Campbell and N. O. Smith. 1951. *Phase Rule*. Dover, New York. Many details about all kinds of phase equilibria.

EXERCISES

1. Ice, the solid form of water, floats on water. Explain why this is sensible from the phase diagram (Figure 6.1).

2. For Fe (Figure 6.3) at $P = 1$ atm and with T varying from 100°C to 2000°C, determine the phases present and the number of intrinsic variables needed to specify that phase as T rises using the Gibbs phase rule. Write the reactions occurring at the phase boundaries.

3. Explain how you could calculate the equilibrium phase diagram for S in terms of what information you need and the formulas to use.

4. List and discuss reasons why Ge and Si and Cu and Ni are completely soluble in the solid state while Pb and Sn are not (Figures 6.10, 6.9, and 6.12).

5. Starting with a molten composition of 80% Pb in Sn cool to room temperature (Figure 6.12). Write the reactions that occur at each phase boundary during the cooling process.

6. For the composition in problem 5, indicate the phases present and their compositions and the relative amounts of the phases present at 260°C.

7. Repeat problem 6 but for 200°C. Show that all the phases present increase in the amount of Pb. Compare you answers to exercises 6 and 7, and explain how this can occur when no material is added.

8. Relate the homogeneity range and its variation in temperature for GaAs (Figure 6.14) with point defects in the material.

9. Construct a binary liquid–solid phase diagram of T versus composition (mole fraction or atomic %) at 1 atm for pure element A (mp = 300°C) and pure element B (mp = 400°C). On the diagram include (1) a eutectic at 200°C and 50–50 composition, (2) limited solubility of B in A, (3) limited solubility of A in B, (4) a compound AB_2 that has a finite homogeneity range. Label each phase field. Use α, β, γ, etc., for alloys and define the alloy (e.g., A in B).

10. Explain why the assumption of spherical nuclei is a good assumption.

MECHANICAL PROPERTIES OF SOLIDS—ELASTICITY

7.1 INTRODUCTION

In this chapter on mechanical properties of solids we will deal mainly with the behavior of solid materials in response to applied forces. The questions that arise in this area are practical ones. How will the material change shape or break when a certain force is applied? How strong is the material? Will the material restore to its original shape after the distorting force is removed. The answers to these kinds of questions lie in the basic structure and morphology of the material. In this chapter and in Chapter 8 the ideas that are used to answer these questions, and to explain other mechanical behavior, are presented.

For the purpose of visualizing the evolution of mechanical properties of solids, consider a bar of solid material that can interact with a periodic disturbance that applies force to the solid, as indicated in Figure 7.1. In this case the disturbance is a periodic compression wave in air. The wave is a longitudinal wave (disturbance in the direction of propagation) composed of compressions and rarefactions of the air surrounding the solid, as opposed to a transverse wave (e.g., an electromagnetic wave) where there is no disturbance in the direction of propagation but rather normal to the propagation direction. As the wave interacts with the bar, compressions and rarefactions that occur in the wave are transmitted to the solid, yielding periodic density variations in the solid. The disturbances in the solid can be short or long wavelength disturbances. Short wavelengths are comparable with atomic, unit cell, and/or bonding dimensions, while long wavelengths are large compared with unit cell dimensions. Suppose that the solid in Figure 7.1 is made of spherical hard atoms connected to each other by bonds that act as springs as shown in Figure 7.2. In this billiard ball and spring model for a solid, the case of short wavelength can be envisioned as resonances in the natural vibrations or motions of the
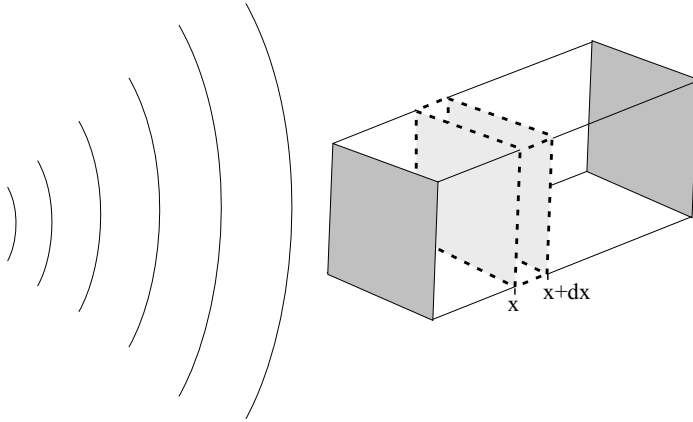
Figure 7.1   A transverse wave impinging upon a bar of solid.
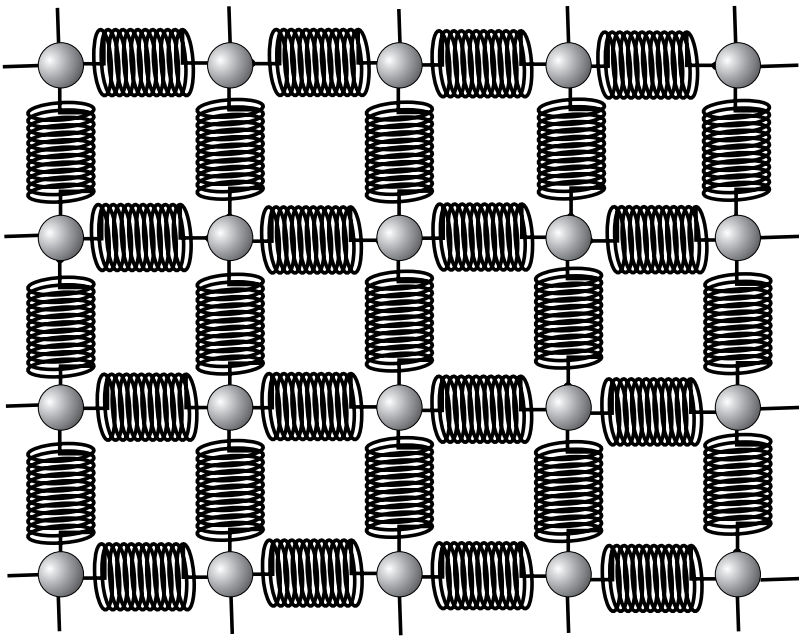
Figure 7.2   A solid approximated using a ball and spring model.

atoms. The rules for which vibrations are allowed and not allowed lie within the realm of quantum mechanics. The allowed resonances are called phonons, and they represent quantized lattice vibrations. These phonons are very important in determining some properties, such as heat capacity and the conductivity of heat and electricity. As the vibrations are excited by resonant energy, they can interact with and scatter other radiation.

In this text we do not consider short wavelength or phonon phenomena. Rather, we consider macroscopic disturbances at wavelength scales that are large compared with unit cell dimensions, as shown in Figure 7.1 with the periodic compression (or rarefaction) waves relative to a macroscopic bar of solid that has a large number of atoms. In this instance we can imagine large macroscopic volumes of the solid being compressed or stretched. The discussion of mechanical properties of solids in this chapter and the next will be mainly on how the solid responds to such mechanical disturbances.

In this chapter we deal with elasticity, which is the full recovery of a solid after the application of a deforming force. For most materials elastic behavior is only observed for very small (<0.1%) deformations, and long before the material may permanently deform or fracture. Thus elastic behavior is not as interesting as permanent deformation and fracture, but it is a behavior regime that is well understood. It is the basis for other more complex behavior that will follow in Chapter 8.

7.2   ELASTICITY RELATIONSHIPS

Elasticity refers to the ability of a solid to return to its original shape after the removal of an applied force that was sufficient to cause deformation. In the limit of small and totally recoverable deformations, elasticity is governed by Hooke's law:

$$\mathbf{F} = -k \cdot \Delta l \tag{7.1}$$

where the small deformation Δl in one dimension (1-D) is the result of an applied force, $\mathbf{F}$, with k as the constant of proportionality. As was noted above, in modeling mechanical properties, we can imagine the solid's atoms to be like billiard balls connected and held in place by chemical bonds in the form of springs (Figure 7.2). For small deformations the deformation is linearly proportional to the force with the (spring) constant of proportionality, k. This k is not related to the k space we computed in Chapter 3 and which we will use later in this chapter. The −sign indicates that $\mathbf{F}$ is a restoring force responding to the deformation like a spring (i.e., the applied and restoration forces are in opposite directions). The force $\mathbf{F}$ is applied per unit cross-sectional area A. $\mathbf{F}/A$ is the stress, and it has units of pressure. The response in the material due to the stress is called a strain. The stress (σ) and strain (ε) are related according to Hooke's law as

$$\sigma = \frac{\mathbf{F}}{A} = E \cdot \varepsilon \tag{7.2}$$

Note that the stress is directly proportional to strain with the constant of proportionality being E and called Young's modulus. Figure 7.3a shows a stress–strain curve for an elastic solid. At small strains (or stress), the stress and strain are linearly proportional. However, at large applied strains (or stress), the dependency is nonlinear, and the deformation is likely not fully recoverable. The onset of the nonlinear, nonrecoverable regime is the end of elasticity, or the elastic limit. This limit is indicated by the dashed line in Figure 7.3a. After the elastic regime is the nonlinear plastic regime that ends with fracture, as denoted by x in the figure. (Plasticity and plastic deformation will be discussed in Chapter 8.)

Figure 7.3b compares stress–strain behavior for three different solids. From the left, the first solid elastically deforms up to a certain stress (or strain) and then fractures
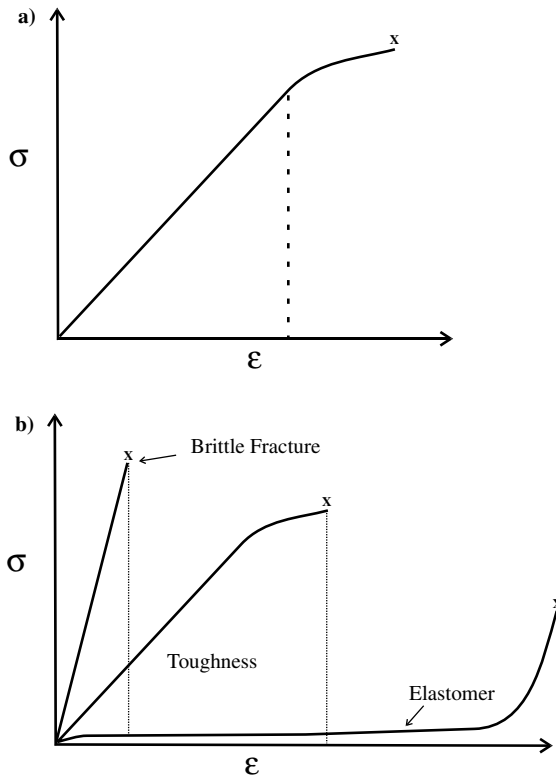
Figure 7.3   (*a*) Stress σ versus strain ε curve for a material in the elastic region and with the onset of the plastic region (*dashed line*); (*b*) three σ versus ε curves illustrating different kinds of mechanical behavior. × indicates fracture.

without plastic deformation. This is indicative of brittle solids such as concrete or ceramics. The next solid is the same as shown in Figure 7.3*a*. The solid first displays elastic behavior, then with continual stress displays plastic behavior, and finally fractures. This kind of behavior is called ductility, and it is typical of metals where the material can tolerate higher strain before failure but irrecoverably deforms at the higher strains. The last solid is the most unique in that it takes little stress to produce large deformations. Furthermore, even large deformations, greater than 100%, are often fully recoverable. Such materials are called elastomers, and a rubber band is a good example of elastomeric behavior. These three materials have decidedly different toughness. Toughness is defined as the area under the stress–strain curve, and it is compared for two of the solids in Figure 7.3*b*. Note that the area under the curve for the material with higher ductility is greater, so it has more toughness.

In the following sections elasticity and elastic behavior are explored in detail, starting with definitions of the terms to be used in this chapter.

7.2.1   True versus Engineering Strain

True strain, ε, is defined using the differential

$$d\varepsilon = \frac{dl}{l} \tag{7.3}$$

The formula basically states that the instantaneous change in length (dl) per length (l) yielding true strain is a unitless ratio. From this 1-D relationship the 1-D strain can be obtained as the limit of the sum or as the integral

$$\varepsilon = \int_o^{l_f} \frac{dl}{l} = \ln\left[\frac{l_f}{l_o}\right] \tag{7.4}$$

The "f" and "o" subscripts refer to final and initial lengths, respectively.

The engineering strain is the observed change in dimension relative to or as a ratio to the original dimension:

$$e = \frac{\Delta l}{l} = \frac{l_f - l_o}{l_o} \tag{7.5}$$

Clearly, these definitions are mathematically different, so it is useful to see under what circumstances these definitions lead to similar results.

Table 7.1 summarizes a thought experiment in which a bar of material was subjected to increasing 1-D stress that produced changes in length (l) in the direction of the applied stress. In the left side column is the time sequence during which the measurements were made of the length l recorded in the second column. In the third column is the calculated value for the true strain, ε, using equation (7.4), and in the last column is the engineering strain, e, from equation (7.5). Note that for deformations up to about 15% there is only a small difference in the two strains, but the engineering strain is always larger. With larger strains (50%) there is larger difference. However, such large strains are well outside the elasticity limits (i.e., such large strains on most materials are not fully recoverable). Thus, as a practical matter, it is better to use the engineering strain for simpler elastic calculations of strain. Note further that for small deformations both strains are additive. (Later in Chapter 8 we will consider permanent or plastic deformations where the strains can be large. In that case it is best to use true strain.)

Table 7.1   Comparison of true and engineering strains

Time	Measured l	Calculated ε	Calculated e
0	1	—	—
1	1.05	0.04879	0.05000
2	1.1	0.04652	0.04762
3	1.15	0.04445	0.04545
		Sum = 0.13976	Sum = 0.14307
4	1.5	0.26570	0.30435
		Sum = 0.40546	Sum = 0.44741
		ln 1.5 = 0.40546	

7.2.2 Nature of Elasticity and Young's Modulus

As was mentioned earlier, because a solid is composed of discrete atoms, in studying lattice vibrations one must consider discreteness. For example, if an array of atoms is perturbed by a periodic mechanical disturbance with a wavelength (λ) conformable to the vibrational wavelength of the atoms, then phonons are excited in the material. When the exciting λ is long relative to the discreteness, the solid is considered to be a continuum and the resulting deformations (if small) are called elastic waves. These waves are the concerted motion of many individual atoms.

Consider a solid bar with a longitudinal wave disturbance in the direction of propagation shown in Figure 7.1. Suppose that the disturbance is due to atmospheric compression, and that it interacts with the solid and causes a displacement of dx (of many atoms in concert), as shown by element dx in Figure 7.1. We consider $\Delta u(x)$ to be the elastic displacement of the atoms so that $\Delta u(x)$ is the displacement from the equilibrium separation of atoms (a_o) yielding $\Delta u(x) = (a - a_o)$ under a force, **F**. Then we can write the strain in differential form $d\varepsilon$, in terms of $du(x)$, as

$$d\varepsilon = \frac{d[u(x)]}{dx} \tag{7.6}$$

A better understanding of the nature of an elastic deformation is obtained by considering the forces that exist between the atoms in the solid as a result of chemical bonding. Figure 7.4a shows a typical potential energy versus atomic coordinate separation curve for chemical bonding (a so-called Morse potential). For simplicity, we use a monatomic solid with a bond energy, Φ. $\Phi(u)$ can be expanded in a Taylor series around a_o if the following mathematical conditions are met:

1. $\Phi(u)$ is continuous.
2. $d\Phi/du = 0$, at $\Delta u(x) = 0$, which is at $a = a_o$.
3. $\Delta u(x) \ll a_o$, which is a small deformation on atomic scale.

The resulting Taylor expansion is

$$\Phi(u) = \Phi_o + \left(\frac{d\Phi}{du}\right)_o \Delta u(x) + \frac{1}{2}\left(\frac{d^2\Phi}{du^2}\right)_o \Delta u(x)^2 + \ldots \tag{7.7}$$

From the second condition above, $(d\Phi/du)_o = 0$, we see that there is an extrema at a_o, so we can eliminate the second term. From the third condition, $\Delta u \ll a_o$, it is clear that the high power terms get rapidly smaller with increasing powers, so we can safely eliminate the terms higher than squared. Finally, we obtain

$$\Phi(u) = \Phi_o + \frac{1}{2}\left(\frac{d^2\Phi}{du^2}\right)_o \Delta u(x)^2 \tag{7.8}$$

This expression is made meaningful by first remembering that work is force exerted through a displacement (recall equation 4.3). If we consider the potential energy or the bond energy as related to the work of separation among atoms, then

a)

b)
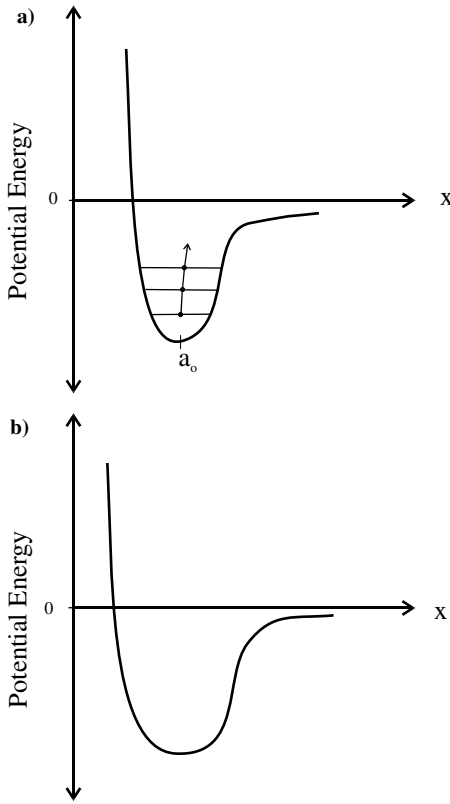
Figure 7.4 (*a*) Potential energy versus atomic separation curve for a solid showing several temperature levels; (*b*) a similar curve with weaker atomic interactions.

$$d(\Phi) = -\mathbf{F} \cdot d(\Delta u(x)) \tag{7.9}$$

and from the result above for $\Phi(u)$ we obtain

$$\frac{d\Phi}{d(\Delta u(x))} = -\mathbf{F} = \left(\frac{d^2\Phi}{du^2}\right)_o \Delta u(x) \tag{7.10}$$

We can conclude that the force is related to the displacement ($\Delta u(x)$) through the curvature at a_o, which is the second derivative at a_o, $(d^2\Phi/du^2)_o$. Furthermore, when this result is compared to Hooke's law, we obtain an expression for E:

$$\frac{d^2\Phi}{du^2} = E \tag{7.11}$$

The modulus of elasticity, Young's modulus, is then a measure of the sharpness of the potential energy (PE) curves near a_o, as given by equation (7.11). Figure 7.4*a* and *b* com-

Table 7.2   Young's modulus and thermal expansion coefficients for selected materials

Material	Young's Modulus, E (10^{11} dynes/cm^2)	Thermal Expansion Coefficient, α (ppm/°C)
Natural rubber	10^{-4}	650
Nylon	0.3	100
SiO$_2$	10	0.8
Si	20	8
Al$_2$O$_3$	34	9
W$_2$C	61	7
Diamond	78	1.2

pares two different PE curves. The sharper featured one, Figure 7.4a, depicts a solid with tighter bonding and a larger E; Figure 7.4b shows smaller curvature, hence smaller E. If one considers the same stress (**F**/A) applied to both solids, the one with the smallest E deforms more, which is due to the weaker interatomic forces. It is interesting to note that within elasticity, the superposition principle obtains, and forces are linearly additive with the predictable deformations as $\mathbf{F}_1 + \mathbf{F}_2$ yields $u_1 + u_2$.

Table 7.2 lists selected E values for some common materials. Note that rubber has a very small E, so it can deform a large amount for a given applied stress compared to tungsten carbide or diamond. Also among the common materials note that as E increases, the material becomes harder and more brittle. In Chapter 8 we will consider the unique features of rubber and other so-called elastomers that can deform or strain a huge amount (>100%) and yet return to their original shape.

Thermal expansion is also readily understood by reference to the PE curves in Figure 7.4. Figure 7.4a shows horizontal lines in the PE curve that represent different temperatures. The center of each line is different due to the asymmetry in the potential curve. The center point represents the mean nearest neighbor distance. For higher temperatures the mean separation increases to larger interatomic separations. The magnitude of this change is called the thermal expansion. The coefficient, α, governs thermal expansion and is defined as

$$\alpha = \frac{\Delta l}{l}(°C) \tag{7.12}$$

Of course, we expect a different α whenever a different potential obtains, as is likely for different crystallographic directions. Typically an average value is reported. Table 7.2 also displays some values of α for several materials. Note that α for rubber and nylon are large compared to α for inorganic materials; SiO$_2$ has one of the smallest α values. Interestingly, when films of rubber and SiO$_2$ are deposited on a substrate of Si at some elevated temperature, say 100°C, no stresses occur due to thermal expansion. However, when the sample is cooled from the deposition temperature to room temperature, both the Si substrate and the film (rubber or SiO$_2$) contract. The rubber will contract almost 100 times more than Si, and SiO$_2$ will contract 10 times less than Si. For the case of a rubber film on Si, the large contraction of the rubber and the relatively small contraction of the Si cause a compressive stress to develop near the Si surface and a tensile stress to develop in the rubber film. For the case of an SiO$_2$ film on Si, the opposite occurs. Namely the
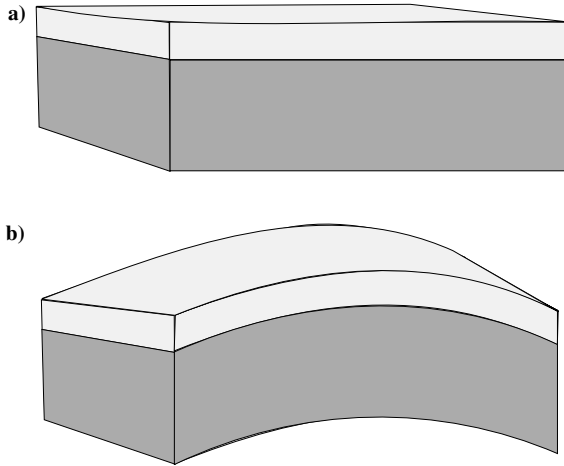
Figure 7.5   (*a*) A solid of different material layers; (*b*) the same bar as (*a*) but after a change in temperature showing strain resulting from different thermal expansion coefficients.

small α for SiO_2 and the relatively large α for Si cause a compressive stress in the oxide film and tensile stress in the Si surface. These stresses can deform the substrate. Figure 7.5 illustrates the case for a film with lower α than that of the substrate. In Figure 7.5*a* the film-substrate sample is undeformed at the deposition temperature, and in Figure 7.5*b* the same sample is observed after cooling to room temperature. This figure greatly exaggerates the deformation imparted by a thin film to a thick substrate. However, for materials of equal thickness with dissimilar α's, significant bending can occur and this effect was used in older thermal switches.

7.3   AN ANALYSIS OF STRESS BY THE EQUATION OF MOTION

In this section we develop a method for the analysis of the equation of motion using the segment indicated by x, $x + dx$ in Figure 7.1. We suppose that the longitudinal wave compresses the solid, and that this compressed segment travels down the solid bar as the wave advances.

Our analysis commences with the basic Newtonian formula for force

$$\mathbf{F} = m\mathbf{a} = \rho A dx \cdot \frac{d^2 u(x)}{dt^2} \tag{7.13}$$

where the units for the right hand expression are (mass/volume $\cdot$ area $\cdot$ dx $\cdot$ distance/time^2 and yields mass $\cdot$ acceleration. Using this relationship and remembering that $\sigma = E\varepsilon$, $\sigma = \mathbf{F}/A$ and $d\varepsilon = du(x)/dx$ (equations 7.2 and 7.6), we can rewrite the stress formulas as follows:

$$\mathbf{F} = [\sigma(x + dx) - \sigma(x)] \cdot A \tag{7.14}$$

where **F** is the net force acting on the segment. Now we write the change in stress in differential form:

$$\frac{\partial \sigma}{\partial x} \cdot dx = \sigma(x + dx) - \sigma(x) \qquad (7.15)$$

Substituting for σ, $\sigma = E\varepsilon$, we obtain

$$\frac{\partial \sigma}{\partial x} = \frac{\partial (E\varepsilon)}{\partial x} = \frac{\partial \{E \cdot [du(x)/dx]\}}{dx} = E \cdot \frac{\partial^2 u(x)}{\partial x^2} \qquad (7.16)$$

Then

$$\mathbf{F} = E \cdot \frac{\partial^2 u(x)}{\partial x^2} \cdot A \cdot dx \qquad (7.17)$$

Equating the two results above (equations 7.13 and 7.17) for **F** yields

$$\rho A dx \cdot \frac{d^2 u(x)}{dt^2} = E \cdot \frac{\partial^2 u(x)}{\partial x^2} \cdot A \cdot dx \qquad (7.18)$$

Equation (7.18) can be rearranged to yield

$$\frac{\partial^2 u(x)}{\partial x^2} - \frac{\rho}{E} \cdot \frac{d^2 u(x)}{dt^2} = 0 \qquad (7.19)$$

which is a 1-D wave equation. A solution to this equation is an exponential of the form

$$u = C \cdot e^{i(kx - \omega t)} \qquad (7.20)$$

where $k = 2\pi/\lambda$ (recall k space from Chapter 3), ω is the frequency, and C is the amplitude. Also, since the momentum p is given as

$$p = \frac{h}{\lambda} \qquad (7.21)$$

which is the de Broglie relationship (see Chapter 9), we can write k as

$$k = \frac{2\pi p}{h} \qquad (7.22)$$

To show that $u(x)$ above is a solution, it is necessary to differentiate twice with respect to both time and position:

$$\frac{du}{dx} = ikCe^{i(kx - \omega t)}$$
$$\frac{d^2 u}{dx^2} = i^2 k^2 Ce^{i(kx - \omega t)} \qquad (7.23)$$

and

$$\frac{du}{dt} = -i\omega C e^{i(kx-\omega t)}$$

$$\frac{d^2u}{dt^2} = i^2\omega^2 C e^{i(kx-\omega t)} \tag{7.24}$$

Combined, they yield

$$-k^2 C e^{i(kx-\omega t)} + \frac{\rho}{E}\omega^2 C e^{i(kx-\omega t)} = 0 \tag{7.25}$$

For this equation to hold, the following relationships must also hold:

$$\omega^2 = \frac{E}{\rho}k^2, \quad \omega = \left[\frac{E}{\rho}\right]^{1/2} \tag{7.26}$$

so we obtain the interesting result

$$\omega = v_s k \tag{7.27}$$

The relationship above, shown in Figure 7.6, is called a dispersion relationship because one variable is the energy and v_s is the speed of sound in the medium. The v_s term has the units of velocity. For example, E can be in units of dynes/cm^2 with a dyne a unit of force as g-cm/s^2 and density, ρ, in units of g/cm^3, and so for E/ρ the units are cm^2/s^2 or v^2. To get a sense of the magnitude of E, recall that a typical velocity of sound is about 5×10^5 cm/s and ρ varies from about 2 to 8 g/cm^3, so we use 5 g/cm^3. These reasonable values give an order of magnitude approximation for E of about 10^{12} dynes/cm^2.
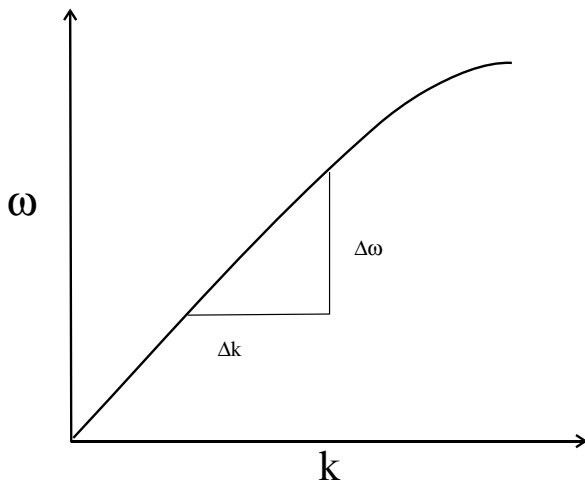
Figure 7.6 Frequency ω versus lattice vector **k** curve for a solid.
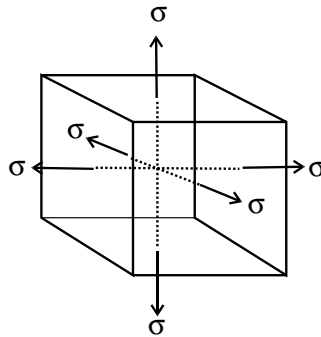
Figure 7.7   Deformation of a cube resulting from normal stresses (σ) applied to the cube's faces.
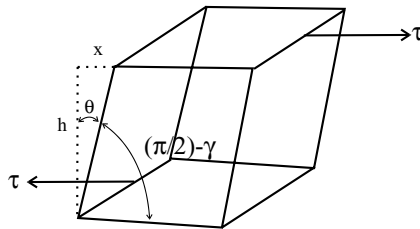
Figure 7.8   Deformation of a cube resulting from shear stresses (τ) applied to the cube's faces.

There are experimental techniques available in which the speed of propagation of acoustic waves can be measured for materials. With a value of v_s measured experimentally for a particular acoustic wavelength, and a value for the density of the material, ρ, a value for the Young's modulus, E, can be obtained.

7.4   HOOKE'S LAW FOR PURE DILATATION AND PURE SHEAR

We first treat pure dilatation, a term that means the change in shape due to some force. While dilatation is usually thought of as an expansion, it can be both + and −. A pure dilatation is shown in three dimensions (3-D) in Figure 7.7. Note that the cube is uniformly deformed by the application of equal forces to its faces so that σ/A is the same on each face. In 3-D the fractional change in volume, $\Delta V/V$, is the strain. Thus

$$\sigma = \frac{B \Delta V}{V} \tag{7.28}$$

where B is the bulk modulus.

Pure shear produces no change in volume, only a change in shape, as shown in Figure 7.8. A measure of the shear strain, γ, is given by the angle θ. From Figure 7.8, $\tan \theta = x/h$, and for small θ, $\tan \theta = \theta$. Thus $\gamma = x/h$, and the shear stress, $\tau \propto \gamma$, is given as
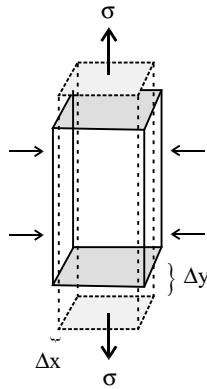
Figure 7.9   Lateral deformation (Δx and Δy) of a solid (*horizontal arrows*) resulting from applied normal stresses (σ).

$$\tau = G\gamma \tag{7.29}$$

where G is the shear modulus.

7.5   POISSON'S RATIO

An axial force per area, σ, applied to the top and bottom of a solid (in the y direction), as shown in Figure 7.9, causes strains (deformations) laterally (in the x and z directions), as indicated by the arrows for the x direction in the figure. The ratio of the induced transverse strain (in x) relative to the axial strain (y) is the Poisson ratio

$$\nu = -\frac{\varepsilon_x}{\varepsilon_y} \tag{7.30}$$

For many metals the range for the Poisson ratio is $\nu \approx 0.25$–0.35 and 0.18 for SiO_2, and since it is a ratio of like values, it is unitless. Intuition tells us, and it is commonly observed, that for a tensile axial strain in y, a positive ν results that is indicative of a negative strain in x. However, in rare materials and for certain directions in these materials, a negative ν is observed. These unusual materials are called auxetic materials.

7.6   RELATIONSHIPS AMONG E, ε, AND ν

The relationship(s) among the various mechanical properties help us calculate unknown properties or at least estimate unknown properties. The relationships among E, ε, and ν are readily obtained by Hooke's law and superposition. Superposition allows small (elastic) deformations from different applied stresses (forces per area) to be added, as was discussed above. We commence by considering the orthogonal x, y, z coordinates and the application of a stress in the x direction, σ_x, that causes strains ε_x and also lateral

strains in y, ε_y and z, ε_z. Using equation (7.30), we obtain the three orthogonal strains:

$$\varepsilon_x = \frac{\sigma_x}{E}, \quad \varepsilon_y = \frac{-\nu_y \sigma_x}{E}, \quad \varepsilon_z = \frac{-\nu_z \sigma_x}{E} \tag{7.31}$$

Similarly, for stresses applied in the y, σ_y and z, σ_z directions, the following relationships are obtained:

$$\varepsilon_y = \frac{\sigma_y}{E}, \quad \varepsilon_x = \frac{-\nu_x \sigma_y}{E}, \quad \varepsilon_z = \frac{-\nu_z \sigma_y}{E} \tag{7.32}$$

and

$$\varepsilon_z = \frac{\sigma_z}{E}, \quad \varepsilon_x = \frac{-\nu_x \sigma_z}{E}, \quad \varepsilon_y = \frac{-\nu_y \sigma_z}{E} \tag{7.33}$$

Using superposition, we proceed to combine the deformations in each of the x, y, z directions; that is, we combine ε_x, ε_y, and ε_z. Consider that a stress applied in the x direction not only causes a deformation in x but also in y and z, according to the Poisson ratio, and the sign of the lateral deformations is opposite to that in the applied direction. The result of superposition, assuming that ν is isotropic, is

$$\varepsilon_x = \frac{\sigma_x}{E} - \frac{\nu \sigma_y}{E} - \frac{\nu \sigma_z}{E} \tag{7.34}$$

$$\varepsilon_y = \frac{\sigma_y}{E} - \frac{\nu \sigma_x}{E} - \frac{\nu \sigma_z}{E} \tag{7.35}$$

$$\varepsilon_z = \frac{\sigma_z}{E} - \frac{\nu \sigma_x}{E} - \frac{\nu \sigma_y}{E} \tag{7.36}$$

The assumption that ν is isotropic is not entirely true in general, but it greatly simplifies the algebra. Because for most materials there is a small range of ν, the approximation will not cause serious error. To further simplify the algebra and the final result, we consider a hydrostatic applied pressure, p. Recall that stress is F/A, or pressure. A hydrostatic pressure is an equal pressure in all directions so that $\sigma_x = \sigma_y = \sigma_z = \sigma = p$. From this we obtain the following summary formulas:

$$\varepsilon = \frac{\sigma}{E} - 2\nu \frac{\sigma}{E}$$
$$\sigma = \left[\frac{E}{1 - 2\nu} \right] \varepsilon \tag{7.37}$$

This result is essentially a correction to Hooke's law as it was expressed above in equation (7.2): $\sigma = E\varepsilon$. The fact that lateral deformations occur, and are summarized by the Poisson ratio enables the correction to be made by tallying the lateral deformations using the superposition principle. As was mentioned above, ν values around 0.25 are expected

for most materials (metals being higher and ionic solids lower). This leads to a correction by a factor of about 2 as

$$\frac{1}{1-2\nu} = \frac{1}{1-0.5} \approx 2 \tag{7.38}$$

7.7   RELATIONSHIPS AMONG *E*, *G*, AND ν

Refer to Figure 7.8, which displays a cube deformed to an oblique parallelepiped, and Hooke's law (equation 7.2), which we used to derive a formula for the shear stress τ in terms of the shear strain γ (equation 7.29) as τ = Gγ, where *G* is the shear modulus. This formula is generalized for the *x*, *y*, *z* coordinates as

$$\gamma_{xy} = \frac{\tau_{xy}}{G}, \quad \gamma_{yz} = \frac{\tau_{yz}}{G}, \quad \gamma_{zx} = \frac{\tau_{zx}}{G} \tag{7.39}$$

At this juncture we have three constants for a material that describe elasticity, *E*, *G*, and ν. Below we will show that these elastic constants are related as:

$$G = \frac{E}{2(1+\nu)} \tag{7.40}$$

Assuming that the approximate veracity of equation (7.40) as we will justify below, we can deduce the approximation

$$G \approx \frac{E}{2.5} \tag{7.41}$$

This relationship enables an estimation of *G* from a measured value of *E*. Also for *E* > *G* it follows that shear deformations are larger than normal deformations due to an applied stress. If then one considers how much deformation a material can withstand before the material fails (fractures) or permanently deforms (plastic deformation), it can be reasoned that the material will more likely fail or permanently deform as a result of shear strains. We return to this important insight in Chapter 8.

To derive the relationships among *E*, *G*, and ν, we consider the prismatic element of a cube formed by a diagonal plane that bisects the cube in Figure 7.8 to yield the prismatic element seen on the right in Figure 7.10. The π/4 deforms to β in the left-hand prism of Figure 7.10, and β is given as

$$\beta = \frac{(\pi/2 - \gamma)}{2} = \frac{\pi}{4} - \frac{\gamma}{2} \tag{7.42}$$

For this value for β to be obtained, a cube with angles originally at π/2 must deform under stress to π/2 − γ as shown in Figure 7.8. So the prismatic element with angles of π/4 deforms as shown at the left side of Figure 7.10 to π/4 − γ/2. Now using the identity

$$\tan(A - B) = \frac{\tan A - \tan B}{1 + \tan A \tan B} \tag{7.43}$$
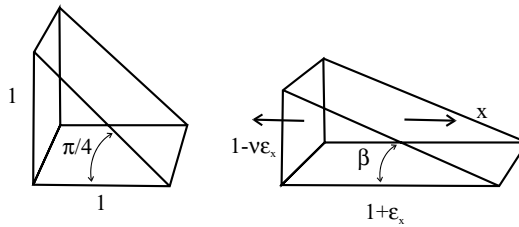
Figure 7.10   Prismatic element of a cube (*left panel*) compared to a deformed element (*right panel*).

where A and B are any angles, we can obtain an expression for $\tan \beta$ with substitutions $\pi/4 = A$ and $\gamma/2 = B$:

$$\tan\beta = \frac{\tan \pi/4 - \tan \gamma/2}{1 + \tan \pi/4 \tan \gamma/2} = \frac{1 - \tan \gamma/2}{1 + \tan \gamma/2} \tag{7.44}$$

This expression can be simplified by considering that $\tan 45° = 1$ and that, for elastic deformations, $\gamma/2$ is small. Then the tan of a small argument is approximately

$$\tan\beta = \frac{1 - (\gamma/2)}{1 + (\gamma/2)} = \frac{2 - \gamma}{2 + \gamma} \tag{7.45}$$

For the deformed element on the right side of Figure 7.10, we obtain the relation

$$\tan\beta = \frac{1 - \nu\varepsilon_x}{1 + \varepsilon_x} \tag{7.46}$$

We can equate the two relationships for $\tan \beta$ (equations 7.45 and 7.46) and obtain

$$(2 - \gamma)(1 + \varepsilon_x) = (2 + \gamma)(1 - \nu\varepsilon_x) \tag{7.47}$$

This equation yields the following value for γ:

$$\gamma = \frac{\varepsilon_x(1 + \nu)}{1 + \varepsilon_x[(1 - \nu)/2]} \tag{7.48}$$

Next we make some approximations to obtain a physically meaningful and simple result. Since ν was approximated to be about 0.25, and for small elastic deformations ε_x is 0.1 or less, $\varepsilon_x \cdot (1 - \nu)/2 < 0.04$. Because the product of ε_x and $(1 - \nu)/2$ is small relative to 1, we obtain γ as

$$\gamma = \varepsilon_x(1 + \nu) \tag{7.49}$$

Recall that $\gamma = \tau/G$ and $\varepsilon_x = \sigma_x/E$ substitution in equation (7.49) yields

$$\frac{\tau}{G} = \varepsilon_x(1 + \nu) = \frac{(1 + \nu)\sigma_x}{E} \tag{7.50}$$

Then, solving for *G*, we have

$$G = \left(\frac{E}{1+\nu}\right)\frac{\tau_{max}}{\sigma_x} \tag{7.51}$$

where τ_{max} is the maximum shear stress. Below it will be shown that the maximum shear stress is obtained at 45° from the application of a normal stress to a solid. With the result above for *G* and the following relationships to be developed below (see equation 7.60), we obtain

$$\sigma_x = \frac{F}{A}, \quad \tau_{max} = \frac{F}{2A} \text{ at } 45°$$
$$\text{then } \frac{\tau_{max}}{\sigma_x} = \frac{1}{2} \tag{7.52}$$

From this we proceed to the desired relationship among *E*, *G*, and ν:

$$G = \frac{E}{2(1+\nu)} \tag{7.53}$$

To fully understand the derivation above, it is necessary to resolve the relevant forces and stresses, that will yield the relationships used above. Figure 7.11 shows applied forces to a bar of solid. The forces are on the same axis, $\mathbf{F}_{AP}$ and $\mathbf{F}'_{AP}$, and are applied to planes A_θ and A, respectively. These planes represent arbitrary cuts to the bar of material. Both $\mathbf{F}_{AP}$ and $\mathbf{F}'_{AP}$ are perpendicular to plane A, but because plane A_θ is at angle θ to plane A, $\mathbf{F}_{AP}$ is not perpendicular to plane A_θ. The applied forces can be resolved in any desired direction. The two components of any stress are the normal component, σ, which is perpendicular to any plane in question, say, the A_θ plane, and the shear component, τ, that is in the A_θ plane. These components of stress are now calculated for the A_θ plane from the given applied stresses. The normal component of force on the A_θ plane is $\mathbf{F}_N$ and is given as

$$\mathbf{F}_N = \mathbf{F}_{AP}\cos\theta \tag{7.54}$$
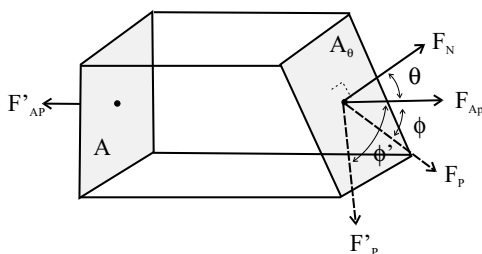
Figure 7.11 Applied normal forces ($\mathbf{F}_{AP}$) to plane *A*. The forces are resolved on plane A_θ in terms of the normal ($\mathbf{F}_N$) and in plane or shear force components ($\mathbf{F}_P$).

This force is converted to the normal stress by dividing by the area A_θ:

$$A_\theta = \frac{A}{\cos\theta} \tag{7.55}$$

With the normal force and the plane area, the normal stress σ is given as

$$\sigma = \frac{\mathbf{F}_{AP}\cos^2\theta}{A} \tag{7.56}$$

The force in the A_θ plane, the shear force, is $\mathbf{F}_P$:

$$\mathbf{F}_P = \mathbf{F}_{AP}\cos\phi \tag{7.57}$$

This force is converted to the shear stress, τ, by dividing by A_θ as above:

$$\tau = \frac{\mathbf{F}_{AP}\cos\phi\cos\theta}{A} = \sigma_{AP}\cos\phi\cos\theta \tag{7.58}$$

The formulas for σ and τ in response to the applied force illustrate Schmid's law. The shear force can have a range of directions in the A_θ plane given another force vector, as drawn in Figure 7.11, where $\mathbf{F}_P'$ and the resolution of $\mathbf{F}_{AP}$ to different directions require different ϕ's and thus have different values. Notice that the normal force is a maximum when $\theta = 0$:

$$\sigma = \sigma_{max} = \frac{\mathbf{F}_{AP}}{A} \tag{7.59}$$

When $\theta = 0$, then $\phi = 90°$ and $\tau = 0$. The shear force is a maximum when $\phi = \theta = 45°$ (as was mentioned above), which is given as

$$\tau = \tau_{max} = \frac{\mathbf{F}_{AP}\cos 45\cos 45}{A} = \frac{\mathbf{F}_{AP}}{2A} \tag{7.60}$$

7.8   RESOLVING THE NORMAL FORCES

In continuing the method above for resolving forces, the shear stress on any plane in a solid can be obtained from knowledge of the normal stresses, which are shown in Figure 7.12. In this figure the shaded plane is the one in question, and its area A_θ can be found from the relationship

$$A_\theta = \frac{A_y}{\cos\theta} = \frac{A_x}{\sin\theta} \tag{7.61}$$

Using the fact that $\sigma = \mathbf{F}/A$, we can write the expression for the shear stress τ resulting from σ_y and then from σ_x to obtain the following force relationships:

$$\mathbf{F}_{\tau_y} = \mathbf{F}_y\cos(90° - \theta) = \mathbf{F}_y\sin\theta \tag{7.62}$$
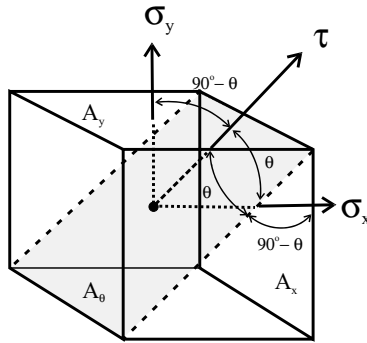
Figure 7.12 Applied normal stresses (σ) are resolved on the A_θ plane in terms of the shear stress (τ).

and

$$\mathbf{F}_{\tau_x} = \mathbf{F}_x \cos\theta \tag{7.63}$$

The forces in Figure 7.12 can be summed at equilibrium as

$$\mathbf{F}_x - \mathbf{F}_y - \mathbf{F}_\tau = 0 \tag{7.64}$$

This expression can be rewritten making substitutions of σA for each $\mathbf{F}$, as

$$\sigma_x A_\theta \sin\theta \cos\theta - \sigma_y A_\theta \cos\theta \sin\theta - \tau A_\theta = 0 \tag{7.65}$$

Solving equation (7.65) for τ yields the following:

$$\tau = (\sigma_x - \sigma_y)\sin\theta \cos\theta \tag{7.66}$$

By the identity,

$$\sin 2\theta = 2\sin\theta \cdot \cos\theta \tag{7.67}$$

and we obtain the final result:

$$\tau = \frac{(\sigma_x - \sigma_y)}{2}\sin 2\theta \tag{7.68}$$

RELATED READING

C. R. Barrett, W. D. Nix, and A. S. Tetelman 1973. *The Principles of Engineering Materials*. Prentice Hall, Eaglewood Cliffs, NJ. A readable elementary text for a first course in materials science.

J. M. Gere and S. P. Timoshenko 1984. *Mechanics of Materials*. Brooks/Cole Engineering, Monterey, CA. An authoritative treatment of mechanics of materials, starting from the basics of elasticity and plasticity and including many important practical examples.

P. A. Thornton and V. J. Colangelo. 1985. *Fundamental of Engineering Materials*. Prentice Hall, Eaglewood Cliffs, NJ. A readable elementary text for a first course in materials science.

EXERCISES

1. A spherical Al pressure vessel has 18 in i.d. and 0.25 in wall thickness. The ultimate stress is 24×10^3 psi and yield stress in tension is 16×10^3 psi. The tank must have a safety factor of 2.1 with respect to ultimate stress and 1.5 with respect to yield stress. What is the maximum allowable pressure in the tank?

2. A load of 1000 lb is suspended from each of two identically sized wires of 0.25″ diameter. One wire is steel ($E = 30 \times 10^6$ psi) and the other is Al ($E = 10.5 \times 10^6$ psi). Determine the axial engineering strain in each, and discuss the atomistic differences in the materials that lead to this result.

3. From the table below calculate the Poisson ratio for each material and discuss why the values are similar or not (depending on your results).

Material	E (psi · 10^6)	G (psi · 10^6)
Carbon steel	30	12
Alloy steel	30	12
Cast iron	15	6
Al	10	4
Brass	14	6
Cu	17	6

4. An Al bar and an alloy steel bar (see table above) are each subjected to 24,000 psi tensile stress. Calculate the lateral strains for both bars. Discuss your answer.

5. For a cubic crystal calculate the force in the [110] for a force of 184 N in the [100].

6. Applied to the cubic material shown below is a force **F** on the (100) plane. Calculate the shear stresses in the (110) plane in the two directions shown. Are the shear stresses equal?
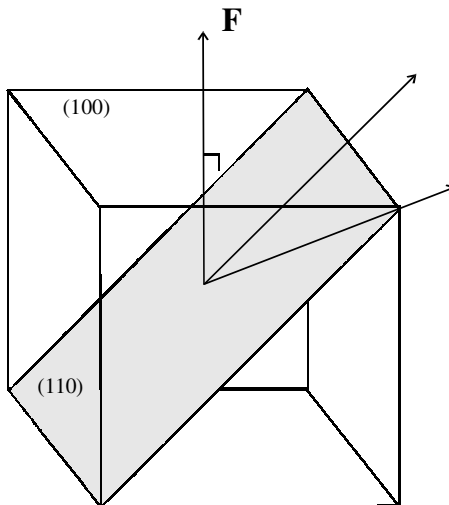
7. Explain why Hooke's law requires a correction by Poisson's ratio. Include in your discussion how the Poisson ratio would effect the elastic deformation due to an applied force.

8. Explain why E for different materials is in general different.

<div style="text-align: right; font-size: 4em;">8</div>

MECHANICAL PROPERTIES OF SOLIDS—PLASTICITY

8.1 INTRODUCTION

Plastic deformation refers to nonrecoverable strains caused, for example, by stresses that exceed the elastic limit for a particular material. Elastic behavior was illustrated in Figure 7.3*a*. The nonrecoverable stress strain behavior is represented to the right of the dashed line in Figure 7.3*a*. It is common practice to deform a paper clip by a small amount in order to clip two sheets of paper. The paper clip withstands this small deformation, and its original shape returns once it is removed from the papers. However, when we bend the paper clip significantly for a large group of papers, it remains in the deformed position. This deformation is an example of a plastic deformation. In the first part of this chapter we review observations about plastic deformation. Then we compare these observations with simple theory, and observations are made about the origins of plastic deformation. In crystalline materials dislocations are the main factor in plastic deformation. Creep is another important mechanism for plastic deformation and will be briefly discussed. From crystalline materials we proceed to a treatment of polymeric materials, and models used to describe the deformation of these kinds of materials.

8.2 PLASTICITY OBSERVATIONS

When a bar of single crystal material is stressed and deformed, characteristic observations can be made relative to plastic deformation. The stresses and resulting deformations are shown schematically in Figure 8.1. In Figure 8.1*a*, a bar of single crystal material is put in tension via stress (σ) applied perpendicular to the shaded crystal faces, as shown by the arrows. An additional shaded plane is shown inside the bar at some arbi-
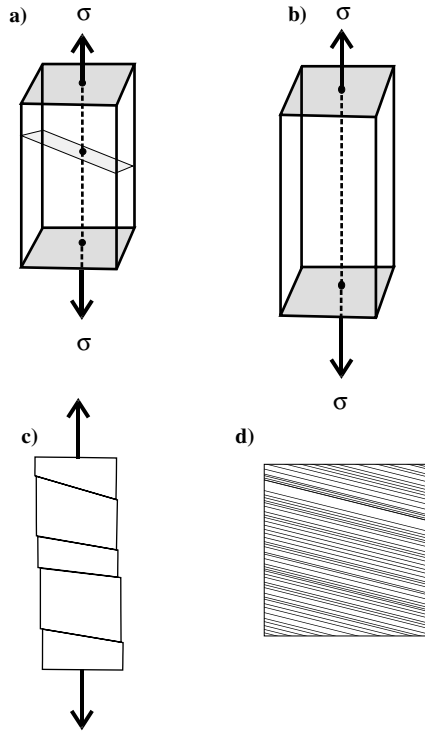
Figure 8.1   (*a*) Tensile stress σ applied to a bar of solid; (*b*) elongation of the bar due to stress; (*c*) macroscopic deformation due to elongation; (*d*) slip bands.

trary angle. As we saw in Chapter 7, the applied forces can be resolved as normal and shear forces on this or any other plane in the solid. Figure 8.1*b* shows the deformation (elongation) of the crystal under tensile stress. If the deformation is large, there will be a small recoverable component (elastic) and a nonrecoverable component to the deformation, the so-called plastic deformation. Close examination of the crystal either during or after a nonrecoverable deformation reveals that the outside surface becomes stepped, as shown in Figure 8.1*c*. The steps lead to striations at specific angles relative to the application of the applied stress. The angles correspond to specific planes that have slipped laterally (relative to σ), causing the steps. The lateral motion of the planes must be due primarily to shear forces resolved on the planes that move, and those planes are called slip planes. Recall also from Chapter 7 that the shear stresses are more efficacious at producing strain owing to the smaller G relative to E. Specifically, we found in Chapter 7 that

$$G \approx \frac{E}{2.5} \qquad (7.41)$$

From this result we can express the strains in terms of E as $\gamma = 2.5\tau/E$ (from $\tau = G\gamma$) while $\varepsilon = \sigma/E$ (from $\sigma = E\varepsilon$). Thus the shear strains can easily be larger than normal strains.

Of course, each plane that slips moves only by atomic dimensions. Thus, to observe the slip, one is in reality observing the lateral motion of many planes. After deformation, further observations of the outside surface reveals the step accumulations as the slip bands depicted in Figure 8.1*d*. Recall that this scenario was briefly discussed in Chapter 4, using Figure 4.4 where dislocations were implicated in the formation of the striations or slip bands. Naturally, edge and screw dislocations arise as a result of the motion of one plane relative to another. Therefore it is only a small stretch of the imagination to relate dislocation motion to plastic deformation in crystals. However, while it is unavoidable to implicate dislocations in crystal plastic deformation, the details are far less obvious. For the dislocations to be causative of macroscopic deformation, there must be many dislocations each with Burger's vector **b** of atomic dimension that, when added, yields a macroscopic deformation. The dislocation must be able to move to the crystal surface, so as to be additive, and the number of dislocations must be variable. For example, one can deform a solid short of fracture. If dislocations are causative, then the number must be proportional to the magnitude of the deformation. On this latter point it has been observed in many studies that as deformation increases, so does the dislocation density. In this chapter the details about how dislocations can move and multiply in crystals are discussed.

Both amorphous and crystalline solids have been observed to plastically deform by means other than dislocations, and this is discussed later in this chapter. The mechanical properties of polymeric solids are of great industrial interest, and they present interesting scientific cases. For example, we know that a rubber band can be stretched (deformed) to hundreds of percent of its original length, yet it can elastically relax to its initial shape. This is called elastomeric behavior, and will be discussed. Also many solids can exhibit a time-dependent elastic and plastic response to an applied stress or strain, and this kind of behavior is also discussed. Different simple models for the mechanical behavior of solids are developed.

8.3   ROLE OF DISLOCATIONS

First we address the issue of how dislocations move through a crystalline solid, and this can be understood with the help of Figure 8.2. Figure 8.2*a* shows an edge dislocation at some distance from the surface. The row of atoms forming the extra half plane at the top of the crystal (indicated by the ⊥ symbol) is shown not to be bonded to the atoms below it in the undisturbed part of the crystal. However, with the shear stress (τ) applied as indicated by the arrows, the top part of the crystal distorts left, and the bottom part to the right. The net result of this distortion is that atoms that were distant from one another are now pushed closer, and some bonding occurs that effectively moves the extra half plane to the left, as is indicated by the ⊥ symbol in Figure 8.2*b*, now one spacing left of its original position. The end result of continual stress is that the extra half plane intersects the surface where it obviously cannot move further. The accumulation of these planes leads to the stepped structure discussed above, and a macroscopic distortion via the sum of many such movements of dislocations. This mechanism for the motion of a dislocation is an energy efficient method. Rather than the motion of an entire half plane the distance to the surface, the bond-breaking and bond-making mechanism is akin to the motion of a periodic disturbance that is a relatively efficient pathway for transport. While all pathways are possible, the pathways that are actually observed are those that are most efficient, meaning those with the highest probability.
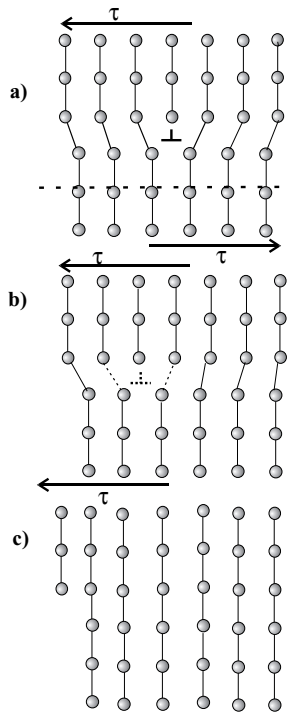
Figure 8.2   (*a*) Shear stress τ applied to a solid that has an edge dislocation; (*b*) the distortion causes motion of the dislocation via bond breaking and making; (*c*) the dislocation eventually moves to the surface.

Table 8.1   Slip systems for face (FCC) and body (BCC) centered and diamond cubic (DC) lattices

Cubic Lattice	Slip Plane	Slip Direction	Burgers Vector
FCC	{111}	⟨110⟩	$a/2\langle110\rangle$
BCC	{110} and {112}	⟨111⟩	$a/2\langle111\rangle$
DC	{111}	⟨110⟩	$a/2\langle110\rangle$

In the discussion above it was mentioned that slip lines form as a result of the resolved forces onto certain planes that are prone to slip. A plane and the preferred directions for slip are called a slip system. Table 8.1 includes slip system information for common cubic crystal systems as well as Burger's vector for the stable dislocations in the crystal. Notice that the slip planes are all low index planes. Recall from Chapter 2 that low index planes are planes with the highest density of atoms. Furthermore the preferred slip directions are the directions of closest approach (see also Chapter 2 on close packing). If we again consider the ball and spring model that was used in Chapter 7 (Figure 7.2), the slip systems are rationalized. Using Figure 7.2 that is sketched for a single plane, imagine

other planes and directions also connected by springs (the chemical bonds), but that the springs for other than nearest neighbors are looser springs representative of weaker bonding than for nearest neighbors. Then imagine gripping the solid and pulling it so as to put the solid in tension by trying to pull it apart. The springs resist and distend so as to reduce the applied force. The tightest springs are the ones that bear the applied force. Now returning to the dislocation issue, it can be realized that the shortest bonding distances on the densely packed planes are those that "feel" the applied stress and distort as a result of the stress. Consequently the most dense planes and directions comprise the predominant slip system. This is not to say that other planes and directions will not permit slip. However, once again, the most probable path is that dictated by the slip system, and on average, it is the pathway most observed.

Previously it was mentioned that dislocations are implicated in the plastic deformation of crystals. If this is true then a crystal devoid of dislocations will resist plastic deformation. The argument made here is a useful one for both understanding and quantifying the degree to which dislocations affect the strength of materials. We commence this argument by first considering a simple model for the strength of materials.

We consider in Figure 8.3a two rows of an undistorted crystalline solid and the same rows in Figure 8.3b with an applied shear stress to the top and bottom rows, as indicated by the arrows. In Figure 8.3b the distortion caused by the stress is measured as x (with $\gamma = x/a$) and can advance to $b/2$, but then from $b/2$ forward the distortion is actually decreasing as the top and bottom rows come back to registry. Given that as x increases,
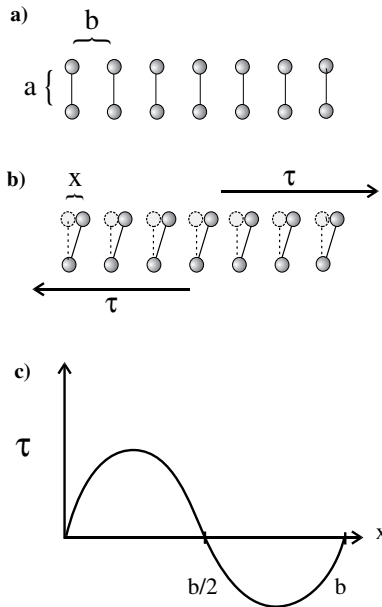
Figure 8.3   (a) Two planes in registry; (b) disregistry as a stress τ is applied; (c) periodic stress as the planes move past one another.

γ increases and thus τ likewise increases, we can plot the variation of τ with x. This periodic variation is shown in Figure 8.3c. This periodic variation can be written as

$$\tau = \tau_{max}\sin\left(2\pi \cdot \frac{x}{b}\right) \tag{8.1}$$

This formula shows that τ varies sinusoidally with the fraction x/b and the fraction $2\pi \cdot (x/b)$ is the x/b fraction of 360°. If the strain is small, we can use the fact that the sine of a small argument can be approximated by the argument, so that the following approximation can be used:

$$\tau = \tau_{max}2\pi \cdot \frac{x}{b} \tag{8.2}$$

With $\tau = G\gamma$ and $\gamma = x/a$ and with $a \approx b$, we can further approximate τ as

$$\tau_{max} \approx \frac{G}{2\pi} \tag{8.3}$$

or τ_{max} is approximately $0.1G$. G values for materials (metals) typically lie in the 10^7 psi range, so the maximum shear stress achievable by a material (maximum strength) is about 10^6 psi. However, typically observed strengths for crystalline materials range from 3 to 5 orders of magnitude less than the maximum, or τ_{max} is about $(10^{-3}$ to $10^{-5}) \cdot G$. Close scrutiny of these results is revealing. It is observed that for dislocation free metals the value of τ_{max} is about 0.1 to $0.01G$, which is close to the approximate calculation made above, considering all the assumptions made and the simplicity of the model for strength. Furthermore, in materials where dislocations cannot readily move such as ceramics and for noncrystalline materials where dislocations do not exist, the actual and theoretical strengths are also reasonably close in magnitude. Therefore it is quite clear that for metals dislocations play a very important role in deformation and strength.

To complete our understanding about the implication of dislocations in plastic deformation, it is useful to make further calculations about dislocations, particularly the energy for a dislocation and the energy to move a dislocation. It is intuitive that the energy of a dislocation must be proportional to the deformation associated with the dislocation. Recall that the Burger vector, **b**, is a measure of the deformation of a dislocation, and thus the energy of a dislocation is proportional to its **b**. With the use of Figure 8.4, which displays a screw dislocation, we can obtain an expression for the strain, γ. The dislocation shown in Figure 8.4a in the annulus with width dr is unrolled in Figure 8.4b, where $2\pi r$ is the circumference of the annulus and **b** is the Burger's vector. The shear strain expression that we will need below is given as

$$\gamma = \frac{\mathbf{b}}{2\pi r} \tag{8.4}$$

To assess whether dislocation will form, it is necessary to calculate the elastic energy for a dislocation. In particular, since a dislocation is a line, we need to calculate the energy per length designated by Γ. We commence with a calculation of the energy per volume element in which the dislocation resides. We use the fact that energy is given by the integral of force times displacement (recall equation 4.3 written for work) as

$$\text{Energy} = \int \mathbf{F} \cdot dx \tag{8.5}$$
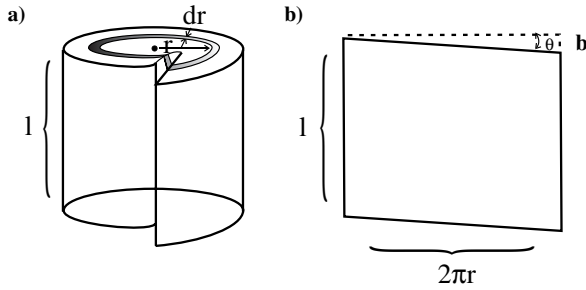
Figure 8.4 (*a*) Screw dislocation intersecting a surface; (*b*) the "unrolled" annulus containing the dislocation.

Then energy per volume can be obtained using $\tau = G\gamma$, and the fact that volume (V) is area (A) times height x:

$$\frac{\text{Energy}}{\text{Volume}} = \frac{\int \mathbf{F} \cdot dx}{V} = \frac{\int \tau \cdot A \cdot dx}{A \cdot x} = \frac{\int G \cdot \gamma \cdot dx}{x} \tag{8.6}$$

From equation (8.4) for γ and Figure 8.4*b* and from the fact that $d\gamma = dx/x$, we obtain for the energy per volume the following formula:

$$\frac{\text{Energy}}{\text{Volume}} = \frac{\int G \cdot \gamma \cdot dx}{x} = \int G \cdot \gamma \cdot d\gamma = \frac{1}{2}G\gamma^2 = \frac{G}{2}\left(\frac{\mathbf{b}}{2\pi r}\right)^2 \tag{8.7}$$

Now we convert this expression to energy by multiplying by volume and integrating and then dividing by the length l of a dislocation to obtain the energy/length, Γ, as

$$\Gamma = \frac{\text{Energy}}{\text{Length}} = \int \frac{\text{Energy}}{\text{Volume}} \cdot \frac{\text{Volume}}{l} = \int \frac{G}{2}\left(\frac{b}{2\pi r}\right)^2 \cdot \frac{2\pi r l \, dr}{l} \tag{8.8}$$

This integration is carried out over the length of the dislocation from 0 to r. However, we are using elastic theory for the expressions, and elastic theory does not hold for large deformations such as those at the core of the dislocation (near $r = 0$) where the deformation is largest. Therefore the core dislocation energy (E_{core}) must be separately calculated and then added to the result from equation (8.8),

$$\Gamma = \int_{r_c}^{r} \frac{G}{2}\left(\frac{\mathbf{b}^2}{2\pi r}\right) dr + E_{core} = \frac{G\mathbf{b}^2}{4\pi}\ln\frac{r}{r_c} + E_{core} \tag{8.9}$$

with the limits of integration commencing where the dislocation core deformation can be approximated by elastic theory formulas (r_c) and extends to the end of the dislocation r. There are several approximations that can be made to simplify this expression. First we do not here attempt to reproduce calculations of the energy for a typical dislocation core. However, estimations of this energy reveal that the core energy is typically

less than 10% of Γ. Thus, for the purposes here, we neglect this small contribution to Γ. For macroscopic dislocations of the order of 1 cm ($r = 1\,cm$) and for dislocations cores with approximate size of 10^{-6} cm, and $E_{core} = 0$, $\Gamma \approx Gb^2$. This final approximation is important because it shows the final desired result that Γ is proportional to the square of Burger's vector:

$$\Gamma \propto \mathbf{b}^2 \tag{8.10}$$

Before proceeding with the use of this expression to approximate dislocation energies, we illustrate the use of this expression to test the stability of dislocations. Specifically, this relationship for Γ can be evaluated for various values of the square of Burger's vector of dislocations. We commence with a brief review of vector algebra. For the vectors,

$$\mathbf{V}_1 = a_1\mathbf{i} + b_1\mathbf{j} + c_1\mathbf{k} \quad \text{and} \quad \mathbf{V}_2 = a_2\mathbf{i} + b_2\mathbf{j} + c_2\mathbf{k} \tag{8.11}$$

The vector sum is given as

$$\mathbf{V}_1 + \mathbf{V}_2 = (a_1 + b_2)\mathbf{i} + (b_1 + b_2)\mathbf{j} + (c_1 + c_2)\mathbf{k} \tag{8.12}$$

The vector dot product can be written as

$$\mathbf{V}_1 \cdot \mathbf{V}_2 = \{a_1a_2 + b_1b_2 + c_1c_2\}\cos\theta \tag{8.13}$$

If $\mathbf{V}_1 = \mathbf{V}_2$, the dot product is given as

$$\mathbf{V}_1 \cdot \mathbf{V}_1 = a_1^2 + b_1^2 + c_1^2 \tag{8.14}$$

The form of a Burger vector is, of course, that of any vector. Therefore we can use vector algebra to predict what happens when dislocations meet and can combine. For example, for the following dislocations, adding we can write, using vector algebra,

$$\frac{1}{2}\langle 111 \rangle + \frac{1}{2}\langle 1\,\bar{1}\,\bar{1} \rangle \rightarrow \langle 100 \rangle \tag{8.15}$$

Furthermore dislocation reactions can be tested. For the example above we can calculate Γ for each of the three dislocations and compare the results, and then make a judgment whether the reaction will tend to proceed to the right as it is written or the opposite. For the $\frac{1}{2}\langle 111 \rangle$ dislocation, $\Gamma = a^2(\frac{1}{4}\,\frac{1}{4}\,\frac{1}{4}) = 3/4a^2$, and for the $\frac{1}{2}\langle 1\bar{1}\bar{1} \rangle \Gamma = 3/4a^2$, and the right-hand side of reaction (8.15) for the $a\langle 100 \rangle$ yields $\Gamma = a^2$. Thus the left-hand side energy is $1.5a^2$, and the right-hand side is $1a^2$. Reaction (8.15) can clearly proceed to the right toward lower energy.

We return to a calculation of the energy of a dislocation according to equation (8.9) for the energy per unit length, Γ. For many materials, especially metals, a value of $G \approx 10^7$ psi or 7×10^4 dynes/cm^2 (1 psi $\approx 7 \times 10^4$ dynes/cm^2). We can assume values for the length of dislocations from, say, 0.2 nm to be 1 to 10 nm and beyond. This yields a total energy of 1.6×10^{-12}, 1.6×10^{-11}, and 1.6×10^{-10} dynes-cm for the 0.2, 1, and 10 nm dislocation lengths, respectively. Converting to eV (1 eV $\approx 1.6 \times 10^{-12}$ dyne-cm), we obtain 1 eV for the smallest to 100 eV for the largest of the three dislocation lengths. Recall that

in Chapter 4 it was indicated that point defects such as vacancies and interstitials require about 1 eV for formation, and so merely at room temperature (≈ 0.025 eV) a finite number of such defects will exist. However, except for the smallest dislocations, the energy required to form dislocations is at least an order of magnitude larger than the energy required for point defects. Thus with the number of dislocations being exponentially related to the negative of the energy required, we can safely predict that the formation of dislocations is not spontaneous at room temperature. This means that applied forces are necessary to produce dislocations. As was discussed above, it has been observed that under applied forces dislocations are produced that move and can multiply.

Recall from Chapter 4 that an important factor for the production of point defects was the configurational entropy derived from placing disorder into a perfect lattice. Line defects, however, contribute far less configurational entropy (i.e., disorder) because the dislocation is restricted in the number of ways a particular dislocation can be arranged in the solid. Thus the entropy of disorder gained does not help to offset the larger energy needed to produce a dislocation. Dislocations are not formed spontaneously because they require an external stress.

The motion of dislocations is also necessary for plastic deformation. The dislocations once produced are able to move to the surface of the solid and thereby accumulate, add, and ultimately yield a macroscopic plastic deformation. Experiments have shown that the energy required to move a dislocation varies with the kind of solid. As we saw, the most efficient kind of dislocation motion requires bond breaking and bond making. Thus it is consistent to consider the bonding in order to obtain insights into dislocation motion. For example, we know that metals typically have weaker bonding, compared to ionic and/or covalent solids. This is consistent with observations that indicate that the stress necessary to move dislocations (σ_{move}) is greater for ionic and covalent solids than for metals, and this can be quantitatively summarized as follows:

$$10^2 \text{ psi (for metals)} \leq \sigma_{move} \geq 10^4 \text{ psi (for covalent or ionic solids)}$$

The energy required can be calculated from these observed stress values and then compared to the energy available at room temperature (≈ 0.025 eV). For this calculation we need to estimate the area over which a dislocation is operative. For the sizes of dislocation lines of from 1 to 10 nm we can estimate area of 10^{-12} to about 10^{-14} cm^2. The force can be calculated using these area estimates from $\mathbf{F} = \sigma \cdot A$ and then the energy from $\int \mathbf{F} \cdot dx$, where x is the magnitude of the distortion or the magnitude of the Burger vector. We estimate this to be about 0.1 nm. With the conversions of psi to dynes (1 psi $= 7 \times 10^4$ dynes/cm^2) where a dyne is (g·cm)/s^2·cm^2), and then converting to eV (1 eV $= 1.6 \times 10^{-12}$ dynes-cm) for comparison, we obtain the results for the two areas in Table 8.2. Note in the table from the energies in the top row for metals at room temperature (0.025 eV)

Table 8.2 Energy to move dislocations

Applied Stress (psi)	Energy (eV) area = 10^{-12} cm^2	Energy (eV) area = 10^{-14} cm^2
10^2	3×10^{-2}	3×10^{-4}
10^4	3	3×10^{-2}

Table 8.3   Approximations

Theoretical strength	$G/2\pi \approx 10^{-1}\,G$
Actual strength (with dislocations)	$(10^{-3}–10^{-5})G$
Energy of a dislocation	$\geq 10\,\text{eV}$
Energy to move a dislocation	$<1\,\text{eV}$

that the dislocations in metals can move at room temperature. For other materials, dislocation motion will depend on specific stresses and temperature, and can be expected in the weaker bonded materials.

Before proceeding farther it is useful to summarize the main ideas developed thus far using the approximations made above and displayed in Table 8.3. From the first two rows of Table 8.3 are the dislocations implicated in plastic deformation and strength of materials. They apply to crystalline materials in which dislocations can be defined. The third row indicates that the dislocation energy is relatively high and that dislocations do not form spontaneously in most materials and require external stress to be produced. Once dislocations are produced, the fourth row indicates that the dislocations can move easily via a periodic motion of bond breaking and bond making. All the conclusions are appropriate for metals where the bonding is weaker and dislocations can move easily. Other crystalline materials can also exhibit plastic deformation via dislocations but the specifics of the materials need careful scrutiny.

The last major point is that the dislocation density must increase during the application of progressively increasing stress in order to account for increasing plastic deformation with stress. There are many ways that this can occur, and here we introduce just a few. First and easiest to imagine is the expansion of a dislocation line or loop, as shown in Figure 8.5a and b. Dislocation density ρ_D is given by a product of the number of dislocations, n, with each multiplied by its length, l, and divided by the volume, V, considered as

$$\rho_D = \frac{\sum_i n_i l_i}{V} \tag{8.16}$$

While it is intuitive that an increased number of dislocations in a volume will increase the dislocation density, it is be less obvious that the increase in length of any one dislocation in the volume without increasing the number will increase the dislocation density. The line dislocation depicted in Figure 8.5a can be stretched by the application of stresses to the solid from the solid line to the dotted line. The dislocation length l then increases, and according to equation (8.16), so does the dislocation density. Likewise the stretching of the dislocation loop shown in Figure 8.5b increases the perimeter of the loop, and thus causes an increase in dislocation density.

For the dislocation loop we can calculate the stress necessary for expansion and to increase ρ_D by recalling that the energy per length is given by $\Gamma = G\mathbf{b}^2$. For the loop this energy per length is multiplied by the length $2\pi r$ to give the total energy. For the expansion from r to $r + dr$ the extra energy needed is then given as $2\pi\, G\mathbf{b}^2 \cdot dr$. Also for expansion the energy or work of expansion is given by $\int \mathbf{F} \cdot dr$ where the force per length can be given by $\tau \cdot \mathbf{b}$ or the force per area multiplied by the length of the dislocation. This force per length is in turn multiplied by the length of $2\pi r$ to yield for the work a value $\tau \cdot \mathbf{b} 2\pi r \cdot dr$. Basically there is work necessary to expand the loop ($2\pi\, G\mathbf{b}^2 \cdot dr$) and there is
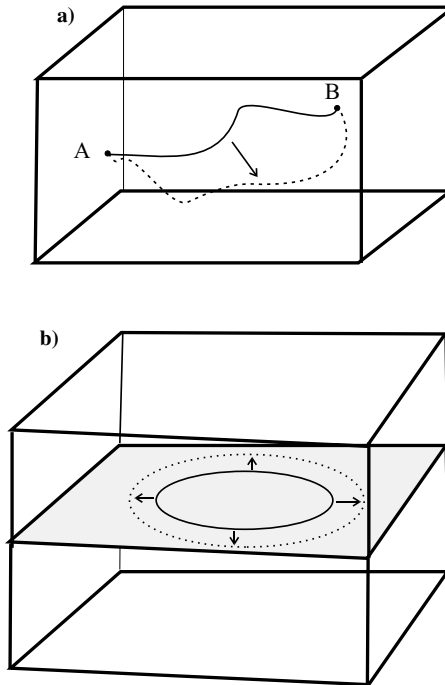
Figure 8.5   (*a*) A dislocation line pinned at *A* and *B* in a solid and stretched via stress; (*b*) a dislocation loop that is stretched via stress.

work of resistance or line tension ($\tau \cdot \mathbf{b}2\pi r \cdot dr$). This situation can be imagined as the stretching of a rubber band. We need to overcome the line tension of the rubber band to expand it. At equilibrium these two work terms need to be equal. So the result for τ to achieve equilibrium is given as

$$\tau = \frac{G\mathbf{b}}{r} \tag{8.17}$$

and the loop dislocation density will increase when $\tau > G\mathbf{b}/r$ (i.e., the loop will expand).

For the line dislocation shown in Figure 8.5*a* that is pinned at positions *A* and *B*, the line tension *T* is the resistance. The energy for a dislocation ($\Gamma = G\mathbf{b}^2$) from the pinning at two positions is

$$2\Gamma = \tau \mathbf{b} l \tag{8.18}$$

at equilibrium, and hence the condition for the line expansion is

$$\tau > \frac{2\Gamma}{\mathbf{b}l} \quad \text{or} \quad \tau > \frac{2G\mathbf{b}^2}{\mathbf{b}l} \quad \text{or} \quad \frac{2G\mathbf{b}}{l} \tag{8.19}$$

In addition to increasing ρ_D by dislocation expansion, there are mechanisms that can increase the number of dislocations. The best-known mechanism is the Frank-Read
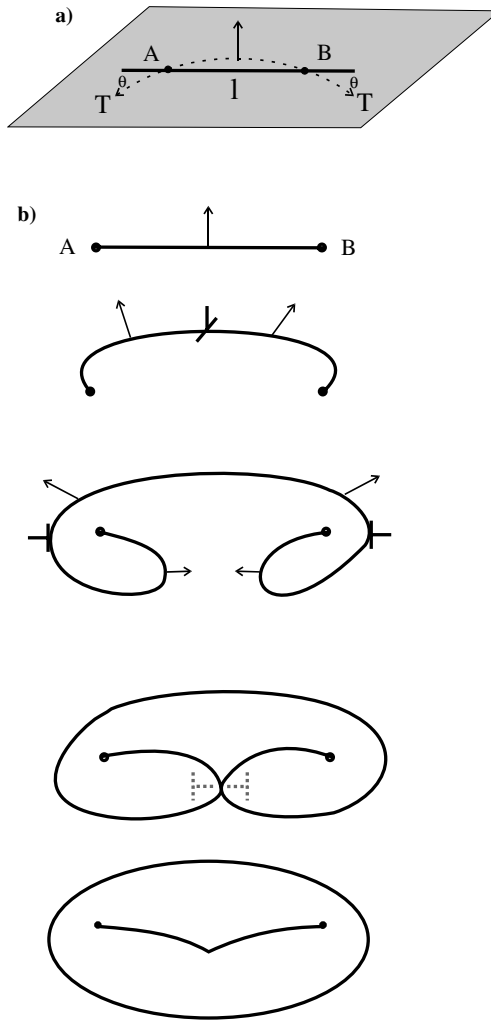
Figure 8.6   (*a*) A dislocation on a plane, pinned at *A* and *B* and stressed; (*b*) the evolution of the stressed dislocation yielding a new dislocation.

source. This mechanism can be understood with the use of Figure 8.6. Figure 8.6*a* shows a dislocation on the shaded plane that is pinned at points *A* and *B* with length *l*. A shear stress τ is applied in such a way as to stretch the dislocation in the direction shown by the arrow as was discussed above. Since the dislocation is pinned at *A* and *B*, it cannot move, but it can expand. The dislocation has a line tension Γ that is the resistance to the expansion caused by τ. The stretching and resistance forces can be equated as was done above to yield

$$\tau \mathbf{b} l = 2\Gamma \sin\theta \qquad (8.20)$$

where **b** is Burger's vector. At $\theta = 90°$ $\Gamma = \Gamma_{max}$, and the following formula is obtained:

$$\tau = \frac{2\Gamma}{\mathbf{b}l} = \frac{2G\mathbf{b}^2}{\mathbf{b}l} = \frac{2G\mathbf{b}}{l} \tag{8.21}$$

As the applied stress increases and exceeds τ, the dislocation expands, as shown by the successive images in Figure 8.6b. Notice that in the top image of the dislocation, the direction of the dislocation is indicated by the half plane symbol, ⊥. As the dislocation expands, it bows out, as shown in the second panel, and the direction of the half planes becomes again shown. Ultimately, as the dislocation expands further, the right- and left-hand bowed sections touch, as shown in the fourth panel of Figure 8.6b. At this point in the evolution of the expanding dislocation where the dislocation parts touch, the dislocation half planes annihilate, as shown by the dashed dislocation symbol, ⊥. A new dislocation line forms and the loop that has formed stays in place as shown in the bottom panel of Figure 8.6b. In summary, the original pinned dislocation, when stretched, literally punches out dislocation loops and reconstitutes itself. This source of dislocations under shear stresses is called the Frank-Read source and is observed in many crystal systems, such as Si where series of concentric loops surrounding a dislocation line have been reported.

It is also possible for plastic deformation to occur without dislocations. One such mechanism is called creep, and we discuss now a specific kind of creep called Nabarro-Herring creep using Figure 8.7. Figure 8.7 pictorially summarizes actual Si oxidation
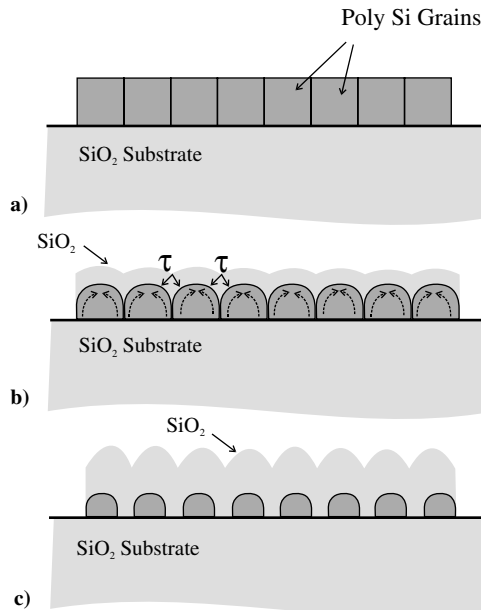
Figure 8.7 (a) Polycrystalline Si grains on a substrate; (b) oxidation causes SiO_2 formation on the free surface and in grain boundaries causing stress that leads to Si migration from the grain boundaries; (c) the intergranular oxidation causes thickness fluctuations in the oxide.

experiments performed some years ago. Specifically polycrystalline Si (poly Si) grains are subjected to O_2 gas at elevated temperatures. Before proceeding, it is well to know that the oxidation of single crystal Si in O_2 results in the oxidation of Si to form SiO_2. The oxidation of the free surface of a single crystal of Si produces a uniform film of SiO_2. Further a comparison of the molar volumes for Si ($V_{M,Si}$) with that of SiO_2 ($V_{M,SiO2}$) yields a value for the ratio of $(V_{M,Si})/(V_{M,SiO2}) = 2.2$. This means that for every atom of Si converted to a molecule of SiO_2, the new solid SiO_2 produced has 2.2 time the volume of Si that it replaces. In metallurgy this molar volume ratio of product to reactant is often referred to as the Pilling-Bedworth ratio. Oxides with large ratios such as that for SiO_2 and Si usually result in the oxide flaking off (also called scaling) as it is formed. This failure in the coating is due to the fact that the stresses developed with the molar volume expansion exceed the mechanical strength of the coating. However, it is observed that the SiO_2 film does not fail when Si is oxidized. The fact that scaling does not occur is attributed to the viscous flow of the amorphous oxide produced, and the flow relieves the stresses that can accumulate from to the large molar volume expansion that occurs during Si oxidation (viscous flow is discussed below and refers to the motion of groups of atoms or molecules). Returning to the creep problem, we see in Figure 8.7*a* a regular arrangement of grains of Si on a SiO_2 substrate, whose regularity is shown only for illustration purposes. The substrate takes no part in the scenario and is merely a relatively inert support. When this polycrystalline Si is exposed to O_2 at elevated temperatures, the Si is thermodynamically driven to react with O_2 to form SiO_2, and at temperatures below about 1200°C the oxide will be amorphous. Now the difference between the oxidation of single crystal Si, and the polycrystalline Si shown in the figure is due to the grain boundaries. For single crystal Si only the free surface of Si oxidizes, but for polycrystalline Si, both the free surface and grain boundaries (which are boundaries with disorder) can oxidize. In the geometry shown in Figure 8.7 the grain boundaries will oxidize more slowly than the free surface, because O_2 needs to penetrate the boundaries, but will nevertheless oxidize. The relatively large volume of oxide (2.2 times the volume of Si) that forms on the free surface can readily expand in the free direction, provided it can flow (later we deal with viscosities and the ability to flow). At high temperature the oxide viscosity is sufficiently small that the oxide can flow in the free direction. However, the oxide that is confined in the grain boundaries cannot flow as readily as that on the free surface. Consequently a lateral stress develops due to the lack of free volume that is necessary to accommodate the relatively large volume of SiO_2 forming with the Si. When the stress in the grain boundaries due to oxidation reaches sufficiently high values, the flow of Si and/or SiO_2 occurs to relieve the stress. In this materials system, a flow of Si is observed from the grain boundary region toward the top of the free side of the grains, as shown by the dashed arrows in the grains in Figure 8.7*b*. Such a flow of Si atoms reduces the intergranular stress by creating volume for the forming SiO_2. At the same time more Si at the free surface can oxidize with minimal accumulation of stress. This flow of Si (or whatever else) in response to a lateral stress is called Nabarro-Herring creep. Often rather than talk about the flow of atoms, materials scientists refer to the flow of vacancies (volume) in the opposite direction.

In summary, in the plastic deformation of crystalline materials, dislocations are clearly implicated for metals and may also be crucial for other materials. Dislocations can accumulate via motion, and they multiply under stress to produce plastic deformation. Other mechanisms for plastic deformation such as creep are also possible.

8.4   DEFORMATION OF NONCRYSTALLINE MATERIALS

Noncrystalline or amorphous materials can plastically deform but without dislocations. In amorphous materials dislocations cannot be defined because this class of materials does not have long range order. Many amorphous materials of importance are either so-called network solids or polymers, and this dichotomy provides a reference frame from which to discuss the deformation of amorphous materials. The network solids include inorganic ceramics such as SiO_2 as was discussed in Chapter 2 and illustrated in Figure 2.2 as a continuous network of SiO_4 tetrahedra. Network solids typically have short range order dictated by chemical bonding that defines the basic building blocks, and these blocks are connected to form networks that extend throughout the material. Polymers are similar up to this point, but in addition to having a network, the networks have anisotropic bonds. In one direction the chemical bonding is strong in defining the building blocks and combination of building blocks, the polymer backbone, in a one-dimensional strand. These strands or chains are then bonded together, often in a twisted and tangled array. It is this interchain bonding that discriminates a polymer from a network solid. Typically it is much weaker than the backbone bonding yielding an anisotropy, and in some cases the chains can be extensively tangled. If a polymer with extensively tangled chains is stretched so as to straighten the tangles, then it is the same as a network solid. Thus, in order to understand the mechanical properties of amorphous materials, we commence with a discussion of network amorphous solids, and then add the feature of tangled chains and note the difference.

8.4.1   Thermal Behavior of Amorphous Solids

In order to obtain a coherent picture of amorphous solids, it is useful to commence with the thermal behavior of the solids. Figure 8.8 shows the typical change in the molar volume of a solid with temperature. The behavior is best understood starting from the molten state at high temperature, and at first slowly cooling. A single trajectory is seen in the liquid state and is indicated by the solid line above the temperature T_m. However, at the melting temperature, T_m, the liquid converts to its crystalline state, C. Notice that the slopes of the liquid and solid lines are different and are indicative of different thermal expansion coefficients for liquid and solid forms. Usually this liquid to crystalline solid transformation yields the most compact form for the material, and hence the smallest molar volume. For small molecules and atomic solids this is the pathway that is most often observed. For example, to obtain the pathway leading to an amorphous solid for metals, extremely rapid cooling must be performed. One method of producing amorphous metals involves actually setting off an explosion in the liquid that splatters the molten metal onto liquid nitrogen cooled (77 K) plates. This so-called splat cooling involves the very rapid cooling of the metal before the atoms can attain crystalline order, and the maintenance of the low temperature to prevent or at least reduce atomic migration. However, for larger networks of atoms and molecules, cooling even at nominal rates can yield amorphous solids, because the required rearrangement has few pathways (entropy, S) and requires considerable energy (enthalpy, H). It is seen in Figure 8.8 that two amorphous solids, A and A', are produced by cooling the liquid at rates too fast to achieve the crystalline state C for the solid. In fact a series of amorphous solids can be produced, each somewhat different in molar volume with only two of this set shown in the figure. While the solids A, A', and C are composed of the same building blocks, they have different structures due to the final arrangement of the building blocks. The bottom-
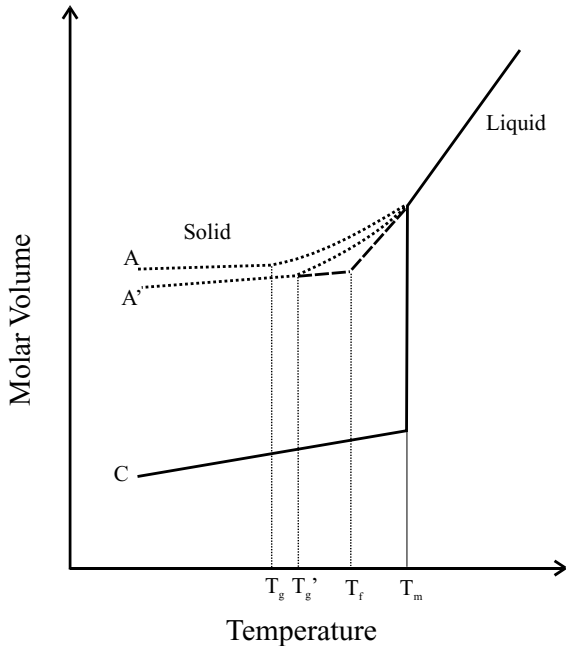
Figure 8.8 Molar volume versus temperature for a solid formed at different cooling rates from the liquid, and resulting in different structures. Melt temperature, T_m, fictive temperature, T_f, and glass transition temperatures T_g indicated.

most solid, the crystalline one, C, has more order while the topmost one A has the least order. The different structures are usually identified by the temperature at which the structure in going from liquid to solid no longer changes. This temperature is called the glass transition temperature and labeled as T_g. Notice that A and A' have different T_g's. The different T_g's represent different arrangements of the building blocks for the network solids and/or polymers, and different structures are obtained using different preparation procedures. In the case above different cooling rates were used to produce the different solid state structures. Also in Figure 8.8 is an extrapolated temperature, T_f, that is called the fictive temperature. T_f is the temperature at which the extrapolation of the solid and liquid volume versus T lines meet. T_f represents a structure that would exist at the juncture. As such it is not an actual structure but does serve to differentiate different amorphous structures of the same material. T_f was referred to in the older literature on glasses but is hardly used anymore. Sometimes the amorphous structural form for a network solid is referred to as an undercooled liquid.

The measurement of the characteristic temperature T_g is now possible by a variety of precise, commercially available techniques; most measure thermal properties of the material as a function of T. As a practical matter, several of the important temperatures associated with amorphous network solids, mostly oxide glasses, are related to the viscosity of the glass or the viscosity range. We will define viscosity more thoroughly below, but the viscosity, η, is essentially a measure of the resistance of a material to shear forces. The units for η are poise, which is g·cm/s. A material with a lower η will flow more easily under an applied force, while a material with a high viscosity resists flow or flows more slowly under an applied force. Because amorphous solids are sometimes referred to as

Table 8.4   Important viscosity ranges for glasses

Range Identity	Viscosity Range (poise)
Glass transition temperature, T_g	10^{13}
Annealing range	$\sim 10^{13}$
Working range	10^4–10^6
Softening or melting range	$< 10^4$
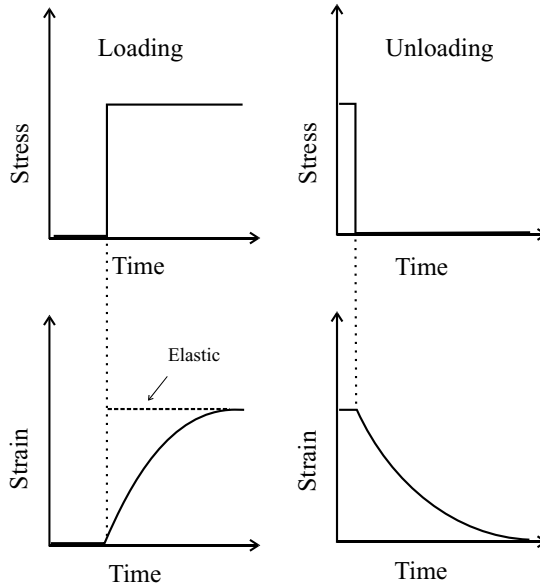
Figure 8.9   Left side shows applied stress-time (loading) to a solid and the anelastic strain response; right side shows the stress unloading and the anelastic strain response.

undercooled liquids, it is natural to also refer to the viscosity of a solid. Table 8.4 summarizes the important viscosity ranges for glasses.

The transition temperature is at 10^{13} poise, which is near the so-called annealing range. This is the range used by glassblowers to anneal away stresses in the finished work. Of course, when actually shaping the glass, a much lower viscosity is required, and this is called the working range. Even in the working range the glassblower doesn't want the glass to be so runny as not to be workable, as it would be at lower viscosity in the softening or melting range where the glass appears to be liquid.

8.4.2   Time-Dependent Deformation of Amorphous Materials

Up to now we have ignored the time dependence of a deformation, or the time elapsed between the application of a stress and the resultant strain. For purely elastic (totally recoverable) deformation, the time difference between stress and strain is zero. The phenomenon where there is a fully recoverable, but time-dependent recovery, is called anelasticity. Anelasticity can be understood by referring to the loading and unloading curves shown in Figure 8.9. These curves show the application of stress (loading) in the top left

panel, and the removal of the applied stress (unloading) in the top right panel; in the bottom two panels the corresponding strains are displayed with particular attention to the time phase. Note in Figure 8.9 that for anelasticity, the application of a step function shaped applied stress yields a time-dependent strain that ultimately reaches a maximum, but compared with a purely elastic response (dashed), the maximum is time delayed. When the stress is removed (unloading), once again the anelastic response is a time delayed, but with full recovery. The difference between purely elastic behavior and anelastic behavior lies in the time lag between stress and strain.

For the plastic deformation of crystalline materials, dislocations and/or creep mechanisms were found to be operative. For plastic deformation of noncrystalline materials, dislocations cannot be invoked. Rather for noncrystalline materials plastic deformation is characterized by viscoelastic behavior. Viscoelastic behavior, or viscoelasticity, is somewhat analogous to anelasticity in that there is a time lag between stress and strain. However, viscoelastic deformation is not fully recoverable as is anelasticity, and the result is plastic deformation. Specifically, viscoelastic behavior is characterized by the shear stress τ being proportional to the strain rate $\dot{\gamma}$ as

$$\tau \propto \gamma \quad \text{and} \quad \tau = \eta \dot{\gamma} \tag{8.22}$$

where η, the constant of proportionality, is called the viscosity. As was mentioned above, the units for η are poise, which is $g \cdot cm/s$, and as is seen from the defining formula above, this unit comes from the stress (dynes/cm^2 where a dyne is $g \cdot cm/s^2$) divided by the strain rate $\dot{\gamma}(s^{-1})$ yields $g \cdot cm/s$ or the poise. A plot of τ versus $\dot{\gamma}$ should yield a straight line with a slope η, and this is commonly referred to as ideal Newtonian behavior for the material. Newtonian fluid behavior is shown in Figure 8.10 as the solid line. The dashed line shows non ideal viscoelastic behavior where there is a deviation at larger strain rates. Specifically, the deviation at higher strain rates indicates that lower shear stresses are required for non-Newtonian behavior to cause given shear rates as compared to Newtonian liquids. Another way to look at it is that lower values of the viscosity (the slope) occur at higher stresses or higher shear rates. This is referred to as shear-thinning
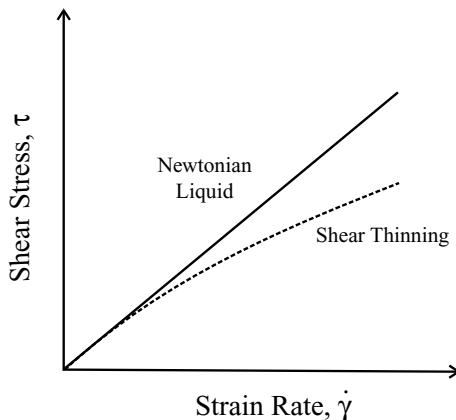
Figure 8.10 Shear stress versus strain for a Newtonian liquid (*solid line*) and a non-ideal Newtonian liquid (*dashed line*).
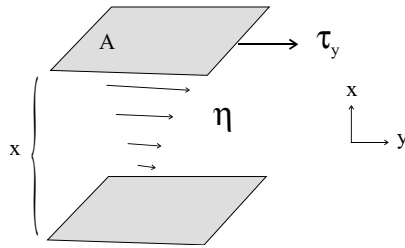
Figure 8.11 Two planes with relative motion caused by stress τ immersed in a liquid of viscosity η.

behavior, and it provides a useful property for fluids. For example, house ceiling paint is made somewhat thicker than wall paint to prevent dripping. However, in addition chemicals are added so that as the paint is spread (sheared) with a brush or roller (force is applied), and the paint thins so as to cover and be spreadable. The substances added are called thixotropic agents, and the shear-thinning fluid is called a thixotrope. Also some liquid rocket fuels are thixotropes. The fuels are essentially gelled solids at normal pressures so as to reduce hazards due to leaks, but when pumped at elevated pressure (stressed) the fuel flows readily.

It is useful to consider the origin of viscosity from a phenomenological viewpoint. Figure 8.11 shows two equal area (A) plates immersed in a fluid. A useful way to imagine this is to consider the two metal plates in a container of honey. Now apply a force **F** to the top plate, as is shown by the arrow, while keeping the bottom plate is stationary. The force applied per area yields τ_y with the coordinates shown in Figure 8.11. As the top plate is moved to the right by the applied force, the honey nearest the top plate will also move to the right. The farther from the top plate one moves, the less velocity will be imparted to the fluid, and this is indicated by the progressively smaller arrows in between the plates. Basically there is a velocity gradient for v_y that depends on the x position or $v_y(x)$. This gradient is approximated as $v_y(x)/x$. When $v_y = dy/dt$ and $\gamma = dy/x$, so the gradient is expressed as

$$\frac{v_y}{x} = \dot{\gamma} = \frac{1}{x}\dot{y} \tag{8.23}$$

From equation (8.22), $\tau_y \propto \dot{\gamma}$ with the proportionality constant η, so the velocity gradient is also proportional to the shear stress, with the viscosity determining the steepness or magnitude of the gradient. Next we see how the mechanical makeup of network solids models is constructed using elasticity and viscoelasticity to simulate particular solids.

8.4.3 Models for Network Solids

Models can be formulated using the laws governing elasticity and viscoelasticity, since most real materials have characteristics of both behaviors. Model parameters are essentially elastic and viscoelastic constants. These constants enable a mechanical description of the material, and thus comprise the mechanical properties of the material.

The two main model elements represent elasticity and viscoelasticity. For a purely elastic response a spring is used. Recall that we previously considered a solid to be com-
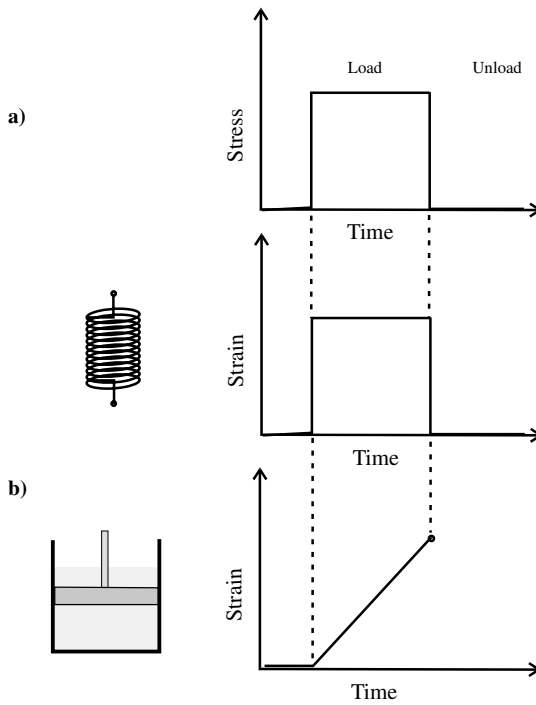
Figure 8.12   (*a*) Spring and (*b*) dashpot with strain response after loading.

posed of springs as chemical bonds. Figure 8.12*a* shows a spring and the loading and unloading response of a spring. Notice that the deformation of the spring, the strain, is in phase with the applied stress. For a purely viscous response a dashpot is used. A dashpot can be envisioned as a cylindrical container with one end open, and inserted in the open end of the container is a fitted piston. The container has a viscous fluid such as honey on both sides of the piston. A dashpot sketch and the loading and unloading responses for the dashpot are shown in Figure 8.12*b*. Notice that the dashpot deformation lags the applied stress, and that when the dashpot is unloaded, it remains at the last strain state. There is no force to return the dashpot to its original position. A combination of spring(s) and dashpot(s) can be used in various arrangements to simulate the mechanical behavior of a wide variety of amorphous solids.

The first arrangement to consider is a solid modeled as a spring and dashpot in series as shown in Figure 8.13. A solid that behaves approximately like a spring and dashpot in series is called a Maxwell solid, and many inorganic network solids display this behavior to some degree; one example is SiO_2. To the connected elements a stress σ is applied so as to cause a total deformation of ε. Both the spring and dashpot deform to ε_1 and ε_2, respectively. With the application of the deforming stress, the spring instantaneously distends to ε_1. Then the dashpot begins to distend ultimately reaching ε_2. When the stress is removed (unload) the spring instantaneously contracts to its original length (elastic response). However, the dashpot remains distended because there is no driving force to
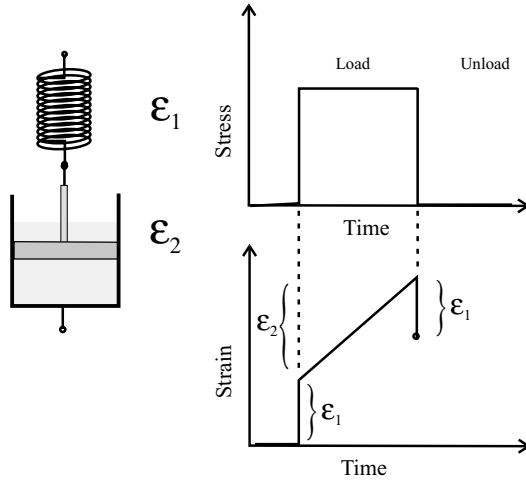
Figure 8.13 Maxwell model for a solid with a spring and dashpot in series and with loading and unloading.

recompress it. With the basic formulas for elasticity and viscoelasticity discussed above $\sigma = E\varepsilon_1$ and $\dot{\varepsilon}_2 = \sigma/\eta$, the following is obtained for the total deformation ε:

$$\varepsilon = \varepsilon_1 + \varepsilon_2 \tag{8.24}$$

Taking the time derivative, one obtains

$$\dot{\varepsilon} = \frac{\dot{\sigma}}{E} + \frac{\sigma}{\eta} \tag{8.25}$$

For a constant deformation, $\varepsilon = $ constant, $\dot{\varepsilon} = 0$, and we obtain the following:

$$-\frac{d\sigma}{dt} = \frac{E}{\eta}\sigma \tag{8.26}$$

The solution for this differential equation is

$$\sigma = \sigma_o e^{-t/\tau} \tag{8.27}$$

where $1/\tau = E/\eta$.

The next arrangement in complexity is to arrange the spring and dashpot in parallel as shown in Figure 8.14. This kind of solid is called a Voigt solid, and it is observed for some polymers where chain straightening occurs first as a stress is applied. The Voigt model introduces a time-dependent elasticity. In the Maxwell solid the applied stress was instantaneously taken up by the spring, and the dashpot did not respond at the instant
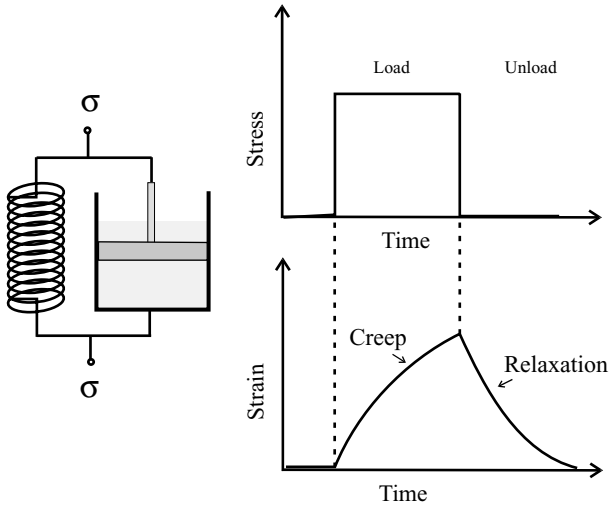
Figure 8.14   Voigt model for a solid with a spring and dashpot in parallel and with loading and unloading.

that the force was applied. However, the Voigt solid behaves differently. In this model the spring cannot instantaneously distend because it is in parallel with the relatively non-responsive dashpot. In fact the slowly responding dashpot controls or limits the rate of deformation and dictates that deformation occurs as a function of time. When the solid is unloaded, the spring is distended, and it stores energy that is then available to compress the dashpot. The applied stress can be written as the sum of the elastic and viscous components:

$$\sigma = \sigma_{elastic} + \sigma_{viscous} \tag{8.28}$$

Then substituting $E\varepsilon$ and $\eta\dot{\varepsilon}$ for the elastic and viscous stresses, respectively, we obtain

$$\sigma = E\varepsilon + \eta\dot{\varepsilon} \tag{8.29}$$

Notice that for the purpose of simplification we have equated the strains ε and γ. Now divide by η and rearrange to obtain a differential equation as follows:

$$\dot{\varepsilon} + \frac{E}{\eta}\varepsilon - \frac{\sigma}{\eta} = 0 \tag{8.30}$$

Rearranging obtains

$$d\varepsilon + dt\left(\frac{E}{\eta}\varepsilon - \frac{\sigma}{\eta}\right) = 0 \tag{8.31}$$

This expression results in a well-known integral form:

$$\int \frac{d\varepsilon}{(\alpha\varepsilon + \beta)} = -\int dt \tag{8.32}$$

Then equation (8.31) is readily integrated to yield

$$\frac{1}{\alpha}\ln(\alpha\varepsilon+\beta) = -t+C \tag{8.33}$$

The constant of integration C can be evaluated at $t = 0$, $\varepsilon = 0$ which yields $C = (1/\alpha)\ln\beta$. When inserted into equation (8.33), it yields the following result:

$$\frac{1}{\alpha}\ln(\alpha\varepsilon+\beta) = -t+\frac{1}{\alpha}\ln\beta \tag{8.34}$$

Equation (8.34) is simplified to the following:

$$\frac{1}{\alpha}\ln\left(\frac{\alpha}{\beta}\varepsilon+1\right) = -t \tag{8.35}$$

This expression has $\alpha = E/\eta$ and $\beta = -\sigma/\eta$, and thus $\alpha\beta = -E/\sigma$. Its exponential form is obtained as

$$\varepsilon = \frac{\sigma}{E}\left(1-e^{-(E/\eta)t}\right) \tag{8.36}$$

It is useful to explore the strain response using this final formula. As time increases, the exponential part decreases, since it is essentially $1/e^t$. Thus, as t increases, ε approaches σ/E, the purely elastic response. As the polymer chains become stretched (i.e., the tangles are straightened), the materials becomes more elastic in mechanical behavior. For larger values of the viscosity (e.g., for more viscous polymers), the exponential is essentially $1/e^{1/\eta}$, and as this term increases, ε decreases. Consequently the creep is slower than in a more viscous system.

To understand how the Maxwell and Voigt models are used to simulate a real solid, imagine a polymeric solid with tangled chains. As stress is applied, the solid deforms slowly in response to the stress straightening out the chains. Once the chains are taut, the solid begins to behave like a Maxwell solid with initially elastic behavior but followed by a creep response. The simulation for a Voigt solid is depicted in Figure 8.14 with a spring and dashpot parallel network.

The temperature dependence can also be modeled with a combination of models. A Burger solid is a series combination of Maxwell and Voigt models as is shown in Figure 8.15. The spring at the top simulates low-temperature bond stretching, and the bottom dashpot simulates high-temperature plastic deformation above T_m. The Voigt model simulates chain straightening that occurs above T_g.

The best way to use these models is to first obtain time-dependent stress and strain data for the solid sample in question. Temperature-dependent data are ideal. Then the elements are added in parallel or series to simulate the behavior. Last, data and model are compared, possibly using regression analysis, and the values for the model parameters are obtained and verified independently.

8.4.4 Elastomers

Some polymers exhibit elastic behavior for very large deformations. These solids are called elastomers, and this behavior was discussed in Chapter 7 and illustrated in Figure
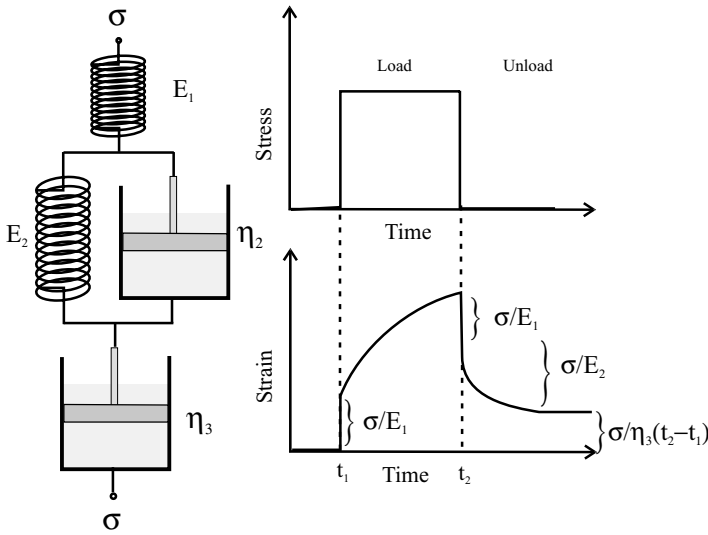
Figure 8.15   Burger model for a solid with a spring and dashpot in both series and parallel and with loading and unloading.

7.3b. Figure 7.3b shows that the elastomer displays more stiffness after a large deformation, which eventually leads to fracture. The unusual behavior displayed by elastomers is attributed to long tangled chains in the material. The extended elastic region is due to the long chains untangling and straightening as the material is deformed. After the chains are straightened, more conventional network solid elastic behavior is observed with the rapidly rising stress–strain behavior and finally fracture. The unique feature of elastomers is that after the straightening, the chains re-tangle as the stress is removed. This behavior is understood by recalling the configurational entropy notion introduced in Chapter 4 that derives from the Boltzmann relationship:

$$\Delta S_{tot} = k \ln \Omega \tag{4.14}$$

Ω is the ratio of the probability of the final to initial states as:

$$\Omega = \frac{w_f}{w_i} \tag{4.15}$$

where w_f is related to the number of ways of forming the final state and w_i is the number of ways of forming the initial state. S_{tot} refers to the total entropy, meaning the sum of $S_{sys} + S_{sur}$. If the T were the same for the system and surroundings, then no q would flow to the surroundings and $dS_{sur} = 0$. So we write the configurational entropy of the system (the solid) as

$$\Delta S_{sys} = \Delta S_{config} = k \ln \Omega = k \ln\left(\frac{w_f}{w_i}\right) \tag{8.37}$$
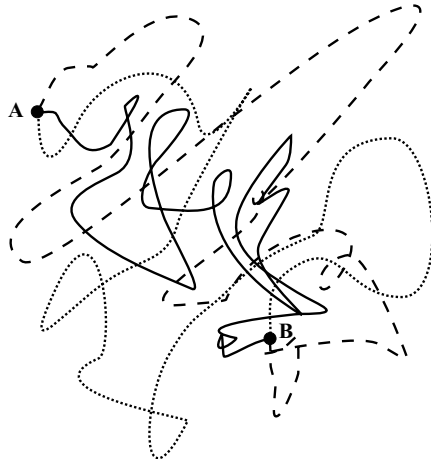
Figure 8.16   Several possible configurations for a polymer chain between two points *A* and *B*.

Now we consider the number of ways of arranging long chains in a solid. Figure 8.16 shows three chains of approximately equal length connected to the same *A* and *B* end points. This is, of course, three ways of arranging chains from literally an infinite set of ways, and in reality there are bonding and steric constraints that limit the number of ways. Because, there are a large number of ways to arrange the chains (large w_f), compared to one way to arrange a straightened chain between two points (w_i), it is easy to see that the configurational entropy will favor the tangled arrangement by a large measure.

The work change *dw* from extending the chains *dx* using a force **F** is given as

$$dw = \mathbf{F} \cdot dx \tag{8.38}$$

Using the Gibbs free energy relationship, $G = H$ (or *E* for solids) $- TS$, and remembering that *G* is the non-pV work in the system, we can also write the following:

$$\mathbf{F} = \frac{dw}{dx} = \left(\frac{\partial G}{\partial x}\right)_{T,P} = \left(\frac{\partial E}{\partial x}\right)_{T,P} - T\left(\frac{\partial S}{\partial x}\right)_{T,P} \tag{8.39}$$

If the elongation does not produce a significant change in bonding, then $\partial E/\partial x \approx 0$, and the change in entropy with elongation is given by the Boltzmann relationship:

$$\mathbf{F} = -T\left(\frac{\partial S}{\partial x}\right)_{T,P} = -Tk \ln\frac{w_f}{w_i} \tag{8.40}$$

Note that for straighter chains w_f is smaller while for tangled chains w_f is larger. As the chains are stretched, the ratio $w_f/w_i < 1$. For this reason the ln term in equation (8.40) is negative and **F** positive. Hence it is predicted that more **F** is needed at higher temperatures, meaning the Young's modulus increases with *T*, which is exactly what we observed

here. Therefore this model of chain straightening yields a consistent picture of elastomeric behavior.

RELATING READING

C. R. Barrett, W. D. Nix, and A. S. Tetelman. 1973. *The Principles of Engineering Materials.* Prentice Hall, Eaglewood Cliffs, NJ. A readable elementary text for a first course in materials science.

J. M. Gere and S. P. Timoshenko. 1984. *Mechanics of Materials.* Brooks/Cole Engineering, Monterey, CA. An authoritative treatment of mechanics of materials, starting from the basics of elasticity and plasticity and including many important practical examples.

P. A. Thornton and V. J. Colangelo. 1985. *Fundamental of Engineering Materials.* Prentice Hall, Eaglewood Cliffs, NJ. A readable elementary text for a first course in materials science.

EXERCISES

1. Discuss the fact that for the BC and FCC the minimum deformation occurs for slip systems.

2. Show why the theoretical strength of a material is seldom realized.

3. Explain why rubber tends to crystallize when it is stretched.

4. How can you distinguish experimentally between mechanisms of Nabarro-Herring creep, and dislocation generation and motion to explain plastic deformation in a material.

5. From the Maxwell model predict the effect of a higher viscosity on the stress response during loading with a constant strain. Likewise predict the effect of a higher Young's modulus.

6. Do the same analysis as in problem 5, but for the Voigt model.

ELECTRONIC STRUCTURE OF SOLIDS

9.1 INTRODUCTION

The electronic structure of solid materials is fundamental to understanding virtually all the properties of materials, including the arrangement of atoms and molecules, thermodynamic properties, mechanical properties, and electronic properties. It is the electronic properties that are the focus of this chapter and the next two chapters, but before these properties can be discussed in Chapter 10 and devices in Chapter 11, a basic understanding of the electronic structure is required.

This chapter commences with wave mechanics, including electron particle-wave duality, and then provides a quantum mechanical treatment of electrons in periodic structures. This discussion leads to the electronic energy band structure model that is used in much of this and the following chapters.

9.2 WAVES, ELECTRONS, AND THE WAVE FUNCTION

9.2.1 Representation of Waves

Figure 9.1 shows the sine function for simple harmonic motion (SHM). At $t = 0$ the wave traveling in this figure is represented as

$$\psi = f(x) = A\sin x \tag{9.1}$$

This representation can be modified to yield SHM at any time t for a wave traveling with velocity v:
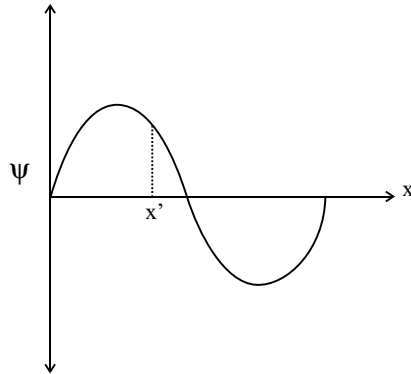
Figure 9.1   A periodic wave traveling in the *x* direction and observed at *x'*.

$$\psi = f(x - vt) = A\sin(x - vt) \tag{9.2}$$

In equation (9.2) the product vt subtracted from x expresses the fact that an observer at x' in Figure 9.1 "sees" the wave moving past from the left. Hence at t the wave amplitude that was previously at the $t = 0$ value is now at the vt value, and the amplitude now at x is given by the value $\sin(x - vt)$, indicating that the wave is traveling from left to right. The displacement in terms of angular measure in fractions of a wavelength, 2π, is given as $(x/\lambda)2\pi$. Since $k = 2\pi/\lambda$, we obtain

$$\Psi = A\sin(x - vt) = A\sin(kx - kvt) \tag{9.3}$$

By the relationships $v = \lambda\nu$ and $\omega = 2\pi\nu$, and then $v = \lambda\omega/2\pi = \omega/k$, the following representation is obtained:

$$\Psi = A\sin(kx - \omega t) \tag{9.4}$$

Of course, for a wave traveling both left and right, the wave is as follows:

$$\Psi = A\sin(kx - \omega t) + A\sin(kx + \omega t) \quad \text{or} \quad \Psi = A\sin(kx \pm \omega t) \tag{9.5}$$

Complex numbers (see the discussion about complex numbers in Chapter 3) are also used to represent harmonic waves. For one thing they are easily manipulated mathematically (differentiated, integrated, etc.). A complex number z is defined as

$$z = a + ib \tag{9.6}$$

where a is the real part, Re(z), and b the imaginary part, Im(z). On the imaginary number axes shown in Figure 9.2, we see that z projected on the real axis is $z\cos\theta$, and on the imaginary axis, it is $z\sin\theta$. Then these terms are added to find the resultant z as
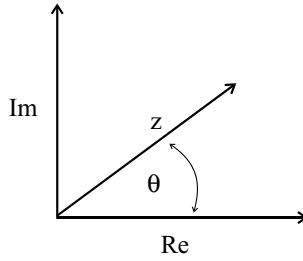
Figure 9.2   Imaginary number *z* is represented by projections onto real (Re) and imaginary (Im) coordinates.

$$z = |z|(\cos\theta + i\sin\theta) \tag{9.7}$$

Then, by the Euler formula, we have

$$e^{i\theta} = \cos\theta + i\sin\theta \tag{9.8}$$

It follows that an exponential representation is appropriate for ψ from equation (9.4):

$$\Psi = Ae^{(i(kx-\omega t))} \tag{9.9}$$

Including both left and right traveling waves and now using Ψ to denote the complete wave (rather than ψ), we have

$$\Psi = \left(Ae^{ikx} + Be^{-ikx}\right)e^{-i\omega t} \quad \text{and} \quad \Psi = \psi e^{-i\omega t} \tag{9.10}$$

Later we will be interested in the derivatives

$$\frac{\partial\Psi}{\partial t} = -i\omega\Psi \tag{9.11}$$

and

$$\frac{d\Psi}{dx} = \left(Aike^{ikx} - Bike^{-ikx}\right)e^{-i\omega t} \quad \text{and} \quad \frac{\partial^2\Psi}{\partial x^2} = -k^2\Psi \tag{9.12}$$

9.2.2   Matter Waves

In this section we relate pure waves with matter. We commence with a review of some physics formulas, and then piece the picture together. For pure waves the total energy carried by a wave is given as

$$E = h\upsilon \tag{9.13}$$

Here υ is the frequency and *h* is Plank's constant. For a particle the kinetic energy is given as

$$E_{\text{kin}} = \frac{1}{2}mv^2 = \frac{1}{2}pv = \frac{p^2}{2m} \tag{9.14}$$

Since p here is the momentum, $p = mv$, and for an electron, $p = m_e v$. Also, for matter, the total energy is given as

$$E = mc^2 = pc \tag{9.15}$$

The speed of light c is given by $c = \upsilon\lambda$, where λ is the wavelength. Since, according to de Broglie, matter and energy are unified, the total energies must be equal, so equations (9.13) and (9.15) yield the relationship

$$h\upsilon = pc \tag{9.16}$$

Substituting $c = \upsilon\lambda$ into equation (9.16) yields

$$h\upsilon = p\upsilon\lambda \tag{9.17}$$

Now the following relationship is obtained:

$$\lambda = \frac{h}{p} \quad \text{and} \quad \lambda \propto \frac{1}{m} \tag{9.18}$$

This remarkable result, known as the de Broglie relationship teaches that both particles and waves have wavelengths and that the wavelength of a particle is inversely proportional to the particle's mass. Interestingly, once this result became accepted, that there is a duality of waves and matter, the mechanics formally used for waves could be applied to particles of matter as well.

Before pursuing the mechanics associated with duality, we will perform some calculations to understand how the de Broglie relationship operates for electrons. Electronic rest mass is $m_e = 9.1 \times 10^{-31}$ kg. If we assume 1 eV electron energy or 1.6×10^{-19} J, with $v = (2E/m)^{1/2}$ from equation (9.14), we obtain an electron velocity $v = 5.9 \times 10^5$ m/s. This yields a momentum, $mv = 5.4 \times 10^{-25}$ kg-m/s. Then, using equation (9.18), we have $\lambda = h/p = 6.63 \times 10^{-34}$ J-s/5.4×10^{-25} kg-m/s $= 1.2 \times 10^{-9}$ m. The value for the wavelength is obtained as $\lambda = 1.2$ nm, or 1 eV electron energy yields a de Broglie wavelength of about 1 nm. Recall that in Chapter 3 similar calculations were done for neutrons, to achieve $\lambda = 0.1$ nm for an energy corresponding to about 400 K (called thermal neutrons).

At this point the foundation has been put in place to treat electrons as waves and employ wave mechanics. At the heart of that possibility of using wave mechanics to explore the electronic structure of matter is the issue of what mathematical functions are to be used for electrons within the wave mechanics machinery that embody the properties of matter in a wave-like function. These functions are called wave functions. Before we proceed to learn what they are, we will review the concept of superposition (also discussed in Chapter 3). Superposition is a helpful concept for wave functions.

9.2.3 Superposition

As we will see below, and have already seen in Chapter 3 on the subject of diffraction, it is important to consider several waves traveling in the same direction and/or imping-
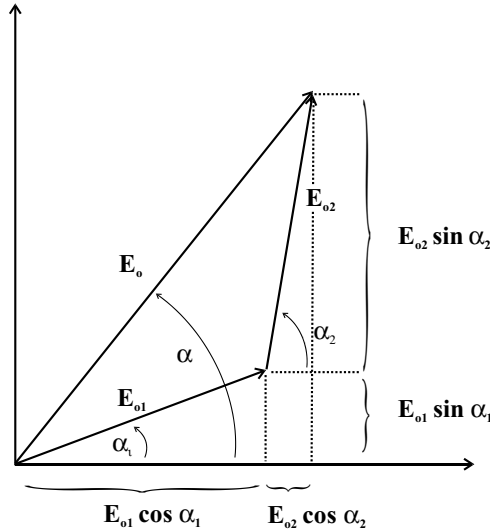
Figure 9.3 Superposition of vectors $\mathbf{E}_{01}$ and $\mathbf{E}_{02}$ to yield resultant $\mathbf{E}_0$.

ing at the same point. In many cases the combined effect of multiple waves can be obtained using the superposition principle. For two waves, 1 and 2, the net displacement at the point of intersection of the two waves, ψ is obtained as

$$\psi = \psi_1 + \psi_2 \tag{9.19}$$

For waves of the same frequency, superposition refers to the addition of waves while tracking both wave amplitude and phase. This is the idea behind diffraction where waves of the same frequency are scattered to yield information about the symmetry and position of structural elements (see Chapter 3). Vectors are used to track two quantities. In the addition of two waves, 1 and 2, we commence by assigning amplitudes $\mathbf{E}_{01}$ and $\mathbf{E}_{02}$ to the waves and phases α_1 and α_2. Figure 9.3 shows the addition of these two vectors to yield the resultant vector $\mathbf{E}_0$. This figure shows that the y components of the resultant are the two sine components, $\mathbf{E}_{01} \sin \alpha_1$ and $\mathbf{E}_{02} \sin \alpha_2$, and that the x components of the resultant are the two cosine components, $\mathbf{E}_{01} \cos \alpha_1$ and $\mathbf{E}_{02} \cos \alpha_2$. These components are separately added and then squared to yield the square of the resultant as given by the following:

$$\mathbf{E}_0^2 = (\mathbf{E}_{01} \sin \alpha_1 + \mathbf{E}_{02} \sin \alpha_2)^2 + (\mathbf{E}_{01} \cos \alpha_1 + \mathbf{E}_{02} \cos \alpha_2)^2 \tag{9.20}$$

Then each of the squared terms is expanded to yield

$$\mathbf{E}_0^2 = \mathbf{E}_{01}^2 \sin^2 \alpha_1 + \mathbf{E}_{02}^2 \sin^2 \alpha_2 + 2\mathbf{E}_{01}\mathbf{E}_{02} \sin \alpha_1 \sin \alpha_2 + \mathbf{E}_{01}^2 \cos^2 \alpha_1$$
$$+ \mathbf{E}_{02}^2 \cos^2 \alpha_1 + 2\mathbf{E}_{01}\mathbf{E}_{02} \cos \alpha_1 \cos \alpha_2 \tag{9.21}$$

By the identity $\sin^2 z + \cos^2 z = 1$, the following expression is obtained:

$$\mathbf{E}_0^2 = \mathbf{E}_{01}^2 + \mathbf{E}_{02}^2 + 2\mathbf{E}_{01}\mathbf{E}_{02}(\sin\alpha_1\sin\alpha_2 + \cos\alpha_1\cos\alpha_2) \tag{9.22}$$

By the identity $\cos(\alpha_1 - \alpha_2) = \sin\alpha_1\sin\alpha_2 + \cos\alpha_1\cos\alpha_2$, the final result is obtained for the addition of two waves:

$$\mathbf{E}_0^2 = \mathbf{E}_{01}^2 + \mathbf{E}_{02}^2 + 2\mathbf{E}_{01}\mathbf{E}_{02}\cos(\alpha_1 - \alpha_2) \tag{9.23}$$

By analogy with the two-wave result, we can generalize for N waves as follows:

$$\mathbf{E}_0^2 = \left[\sum_{i=1}^{N}\mathbf{E}_{0i}\sin\alpha_i\right]^2 + \left[\sum_{i=1}^{N}\mathbf{E}_{0i}\cos\alpha_i\right]^2 = \sum_{i=1}^{N}\mathbf{E}_{0i}^2 + 2\sum_{j>i}^{N}\sum_{i=1}^{N}\mathbf{E}_{0i}\mathbf{E}_{0j}\cos(\alpha_j - \alpha_i) \tag{9.24}$$

The sum of N waves of identical frequency is a wave of the same frequency with a new amplitude and phase. This result can be used to compare diffraction images where the phases are coherent with those where the phases are not.

If the α's are random, then $(\alpha_1 - \alpha_2)$ is also random. As in the random walk problem of Chapter 5, the sum over the random cosine terms will be 0. If we then consider that the amplitudes are nearly the same, we can write

$$\mathbf{E}_0^2 = N\mathbf{E}_{01}^2 \tag{9.25}$$

However, if the phases are equal, then $(\alpha_1 - \alpha_2)$ is 0 and the cosine is 1. Hence we obtain

$$\mathbf{E}_0^2 = \sum_{i=1}^{N}\mathbf{E}_{0i}^2 + 2\sum_{j>i}^{N}\sum_{i=1}^{N}\mathbf{E}_{0i}\mathbf{E}_{0j} = \left[\sum_{1}^{N}\mathbf{E}_{0i}\right]^2 = N^2\mathbf{E}_{01}^2 \tag{9.26}$$

This can be verified for the case of 2 waves ($N = 2$) with equal amplitudes where the following is obtained:

$$\mathbf{E}_0^2 = \sum_{i=1}^{N}\mathbf{E}_{0i}^2 + 2\sum_{j>i}^{N}\sum_{i=1}^{N}\mathbf{E}_{0i}\mathbf{E}_{0j} = 2\mathbf{E}_{0i}^2 + 2\mathbf{E}_{0i}^2 = N^2\mathbf{E}_{0i}^2 \tag{9.27}$$

The result in equation (9.26) was discussed in Chapter 3 for diffraction where the phases are coherent, and equation (9.25) applies to incoherent scattering. The difference in the coherent and incoherent phases is between N and the square of N, where N is a large number. Thus coherent phase diffraction yields a large signal relative to the background.

Another useful result is for waves traveling in different directions. Again, for similar amplitudes we can write for superposition

$$\mathbf{E} = \mathbf{E}_1 + \mathbf{E}_2 = \mathbf{E}_0\sin(kx + \omega t) + \mathbf{E}_0\sin(kx - \omega t) \tag{9.28}$$

Note that selection of sine functions is arbitrary. By the identity

$$\sin\alpha + \sin\beta = 2\sin\frac{1}{2}(\alpha + \beta)\cos\frac{1}{2}(\alpha - \beta) \tag{9.29}$$

with $\alpha = kx + \omega t$, $\beta = kx - \omega t$, we obtain the final result as

$$\mathbf{E} = 2\mathbf{E}_0\sin(kx)\cos(\omega t) \tag{9.30}$$
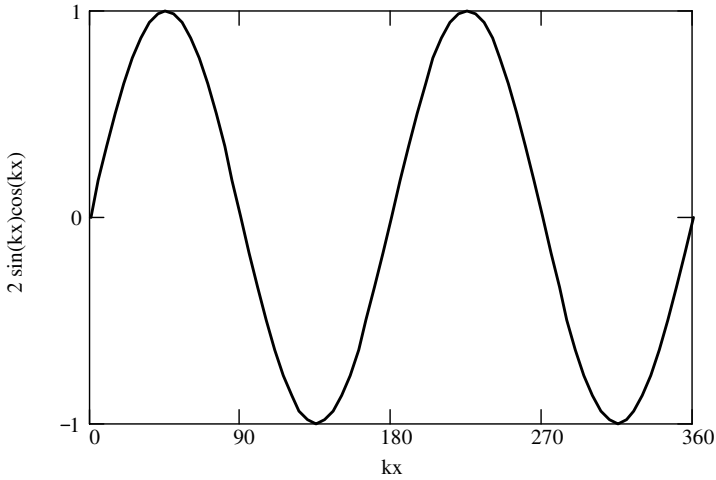
Figure 9.4   Plot of standing wave from superposition with nodes at $kx = m\lambda/4$.

This is a standing wave as shown in Figure 9.4. For $kx = 2\pi x/\lambda$ and $\omega t = 2\pi \nu t = 2\nu \pi v t/\lambda = kx$, the nodes are at $x = m\lambda/4$, where m is 0, 1, 2, . . . , and where either the sin or cos is zero.

The superposition case for our needs ahead is that of waves of nearly the same amplitude but with differing frequency ω and wave number k. Superposition of these waves (in this case cosine waves are arbitrarily chosen) yields

$$\mathbf{E} = \mathbf{E}_1 + \mathbf{E}_2 = \mathbf{E}_0 \cos(k_1 x - \omega_1 t) + \mathbf{E}_0 \cos(k_2 x - \omega_2 t) \tag{9.31}$$

By the identity

$$\cos\alpha + \cos\beta = 2\cos\frac{1}{2}(\alpha + \beta)\cos\frac{1}{2}(\alpha - \beta) \tag{9.32}$$

we obtain

$$\mathbf{E} = 2\mathbf{E}_0 \cos\left[\frac{k_1 + k_2}{2}x - \frac{\omega_1 + \omega_2}{2}t\right]\cos\left[\frac{k_1 - k_2}{2}x - \frac{\omega_1 - \omega_2}{2}t\right] \tag{9.33}$$

After making the substitutions

$$k_h = \frac{k_1 + k_2}{2}, \quad \omega_h = \frac{\omega_1 + \omega_2}{2}, \quad \text{and} \quad k_l = \frac{k_1 - k_2}{2}, \quad \omega_l = \frac{\omega_1 - \omega_2}{2} \tag{9.34}$$

we obtain the simplified result:

$$\mathbf{E} = 2\mathbf{E}_0 \cos[k_h x - \omega_h t]\cos[k_l x - \omega_l t] \tag{9.35}$$

We see that $\omega_h > \omega_l$, so the first term contains a high-frequency (sum of frequencies; see equation 9.34) wave and the second cosine term being the difference in frequencies is a low-frequency term. Figure 9.5a displays this kind of composite wave where, as above,
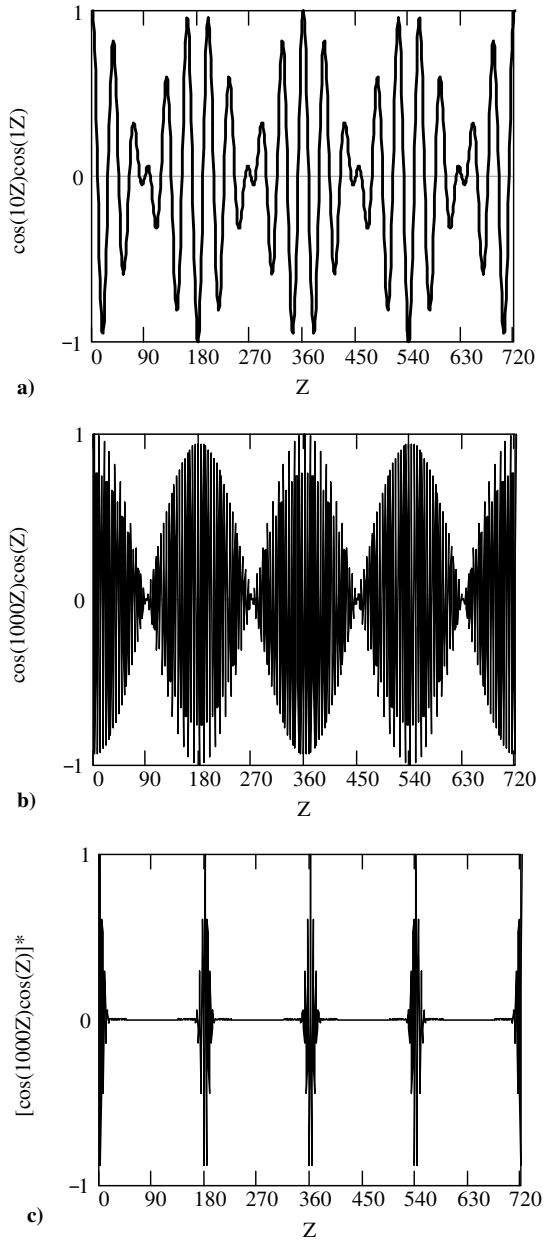
Figure 9.5 Three different modulated waves appearing more "particle-like" from (*a*) through (*c*). *indicates the function raised to the 100 power.

one periodic function alters or modulates the other. In Figure 9.5a we see two frequencies. The amplitude of the high-frequency component is changing at a lower-frequency. In effect there is a low-frequency envelope that contains the high-frequency component. The velocities for the two components are, in general, different. The high-frequency component of velocity is called the phase velocity, v_p, and this velocity can be calculated starting from the following relationship:

$$v = v\lambda = \frac{\omega}{k} \tag{9.36}$$

where ω is the angular frequency, $\omega = 2\pi v$. Then we apply equation (9.36) to high and low frequencies resulting from the addition and subtraction, respectively, of similar magnitude components. We return to equation (9.34) and obtain for the high-frequency component of the wave envelope

$$v_p = \frac{\omega_1 + \omega_2}{k_1 + k_2} \cong \frac{\omega}{k} \tag{9.37}$$

By similar reasoning, we can obtain the lower velocity for the envelope called the group velocity, v_g:

$$v_g = \frac{\omega_1 - \omega_2}{k_1 - k_2} \cong \frac{d\omega}{dk} \tag{9.38}$$

The result obtained here of adding together or superimposing multiple waves, as given by equation (9.35), can also be used to describe the phenomenon of adding intelligence waves (speech or music or images) to a carrier wave that can be transmitted as electromagnetic waves and received by a radio or TV, for example. It is useful to consider how this modulation is accomplished, in that it is related to defining wave functions. For speech, acoustic waves must first be transformed to varying electrical signals using a transducer. For speech or music, a commonly used transducer is a microphone that transforms the longitudinal acoustic waves to varying electric signals. Then these relatively slowly varying electric waves can be superimposed with relatively rapidly varying radio frequency (RF) waves. The RF waves are better suited to long-range transmission. The resulting modulated composite wave has a slowly varying amplitude that is the frequency of the speech and forms an envelope that contains the rapidly varying carrier wave. A radio is tuned to the carrier frequency, and a detector circuit is used that is not sensitive to the high-frequency carrier but rather detects the changing amplitude of the carrier, which is the intelligence. This lower frequency wave is then fed to another transducer (speaker) where it is changed back into audible speech. Figure 9.5b shows two-wave modulation where the high frequency carrier is amplitude modulated (AM) by a lower frequency wave. Of course the waves resulting from speech would not be periodic and would be much more complex than this example shown in Figure 9.5b.

The modulated wave propagates at the group velocity, and thus this velocity is more important because energy is also transmitted by the wave at v_g.

9.2.4   Electron Waves

The ideas and examples developed above about the superposition of waves are now extended beyond the addition of a few waves to the superposition of many waves with disparate frequencies and amplitudes. Figure 9.5a and b can be used to compare what happens as a result of modulation where two waves that are superimposed differ widely

in frequency. In Figure 9.5a the frequency of both the high- and low-frequency component can be discerned. However, in Figure 9.5b it is difficult to discern the frequency of the high-frequency component. If we consider the superposition of many waves with disparate frequencies, the possibilities are endless, but one relevant example is shown in Figure 9.5c. Notice in this case that wide nodes are produced that separate pulses that have a very complex waveform. It is virtually impossible to discern frequency for the high-frequency component. Even more complex modulations can produce pulses that have still larger separations, and also isolated pulses. It is now argued that these widely separated pulses have some properties similar to particles. In these complex waveform pulses, it is not possible to determine the frequency, hence the energy, in the pulse. As the pulse narrows, however, it becomes easier to determine its position. This is the basic idea contained in the Heisenberg uncertainty principle that applies to matter. It is expressed as

$$\Delta p \cdot \Delta x \ge \frac{\hbar}{2} \tag{9.39}$$

where Δp and Δx are uncertainties in the simultaneous determination of position and momentum (recall that momentum and kinetic energy are related as $E = p^2/2m$) of a particle. This expression indicates that it is not possible to precisely determine the energy and position of a small particle. To measure the position using, for example, photons, the position of the electron must be accurately detected by short wavelength photons such as γ or X rays. Any interaction of the probe photon with the electron to be measured can alter the electron momentum (or energy) and hence result in uncertainty. Thus complex modulation yields a wave packet that has properties similar to those of a small particle in that each particle has a readily defined position. For the standing wave shown in Figure 9.4, the frequency is readily discernable but not its position. Therefore one can take an intellectual leap and express a particle, say, an electron, as a complex wave resulting from the modulation or superposition of many simpler waves. The resultant description of the particle is called a "wave function," and the symbol Ψ is usually used. Once having made this scientific hypothesis, one can use the available machinery/mechanics and explore the various conclusions. This is, of course, the field of quantum mechanics, which has allowed a huge and ever-growing number of correct conclusions. All of the "miracles" of quantum mechanics commence with the wave function as an appropriate descriptor of the electron and/or other interesting particles, with the appropriate equations to solve for energy or other interesting properties.

We can summarize this section by realizing that the wave function Ψ, which is used to represent a particle in quantum mechanics, derives from the fact that complex waves resulting from superposition share some important behavior. This notion is at the heart of "duality" and provides the basis for quantum mechanics Below we will adopt the use of Ψ and use appropriate formulas to perform a variety of calculations aimed at elucidating the electronic structure of solids.

9.3   QUANTUM MECHANICS

In this section we adopt the wave function description of the electron, and develop some basic ideas about the Schrödinger equation (SE). Then we present several solutions for the SE that are useful for the understanding electronic structure. The objective is to arrive at a simple model for a solid, the so-called Kronig-Penney (KP) model. Despite its simplicity and assumptions, the KP model contains most of the essential features relating

to the origin of electronic structure of solids without the potentially complicated mathematics necessary for models that provide precise calculations.

9.3.1   Normalization

Recall the exponential form for Ψ expressed in equation (9.10):

$$\Psi = \psi e^{-i\omega t} \tag{9.10}$$

where

$$\psi = Ae^{ikx} + Be^{-ikx} \tag{9.40}$$

Clearly, Ψ is complex. Since the introduction of that representation, we have learned that the wave function Ψ can be used to represent electrons that are real. This dilemma can be overcome using a postulate for the probability $P(x, t)$ of finding an electron in some range of x (in 1-D), say, from x to $x + \delta x$ as

$$P(x,t) = \Psi\Psi^* dx \quad \text{in 1-D} \tag{9.41}$$

and

$$P = \Psi\Psi^* dV \quad \text{in 3-D} \tag{9.42}$$

This postulate renders the probability real, and $\Psi\Psi^*$ (the superscript * refers to the complex conjugate) can be interpreted as a measure of the electron density for Ψ as the wave function representing an electron. Furthermore it follows that the sum of all the probabilities P of finding an electron at all points must equal unity. This can be expressed by

$$\int_{-\infty}^{\infty} \Psi\Psi^* dV = 1 \tag{9.43}$$

The integral equation above is called the normalization condition for Ψ. For the form for Ψ, the complex conjugate for $e^{-i\omega t}$ is $e^{i\omega t}$ and hence $\Psi\Psi^* = \psi\psi^*$. For the case where ψ represents a standing wave, ψ is real and $\psi\psi^* = \psi^2$.

9.3.2   Dispersion of Electron Waves and the SE

Starting from the relationship in Section 9.2.2, equations (9.13) through (9.18), we can write, using $k = 2\pi/\lambda$ and $\Sigma = h/2\pi$,

$$p = \frac{h}{\lambda} = \hbar k = m_e v \tag{9.44}$$

The kinetic energy is given as

$$KE = \frac{1}{2}mv^2 = \frac{\hbar^2 k^2}{2m_e} = \frac{p^2}{2m_e} \tag{9.45}$$

The total energy is the sum of kinetic and potential energies (PE). The resulting dispersion relationship can be expressed as follows:

$$E = KE + PE = h\omega = \hbar\omega = \frac{\hbar^2 k^2}{2m_e} + V = \frac{p^2}{2m_e} + V \tag{9.46}$$

where V is the potential. This yields for ω,

$$\omega = \frac{\hbar k^2}{2m_e} + \frac{V}{\hbar} \tag{9.47}$$

These relationships will be used below.

Assume that a solution for a wave equation is a wave function of the form above:

$$\Psi = \left(Ae^{ikx} + Be^{-ikx}\right)e^{-i\omega t} \tag{9.48}$$

Then the derivatives with respect to time and x introduced above are as follows:

$$\frac{\partial \Psi}{\partial t} = -i\omega\Psi = -i\omega\psi e^{-i\omega t} \tag{9.11}$$

and

$$\frac{\partial^2 \Psi}{\partial x^2} = -k^2\Psi \tag{9.12}$$

Solving these derivatives for Ψ and equating the results yields

$$\frac{-k^2}{i}\frac{\partial \Psi}{\partial t} = -\omega\frac{\partial^2 \Psi}{\partial x^2} \tag{9.49}$$

Using the dispersion relationship for ω from above obtains

$$\frac{-k^2}{i}\frac{\partial \Psi}{\partial t} = -\frac{\partial^2 \Psi}{\partial x^2}\left\{\frac{\hbar k^2}{2m_e} + \frac{V}{\hbar}\right\} \tag{9.50}$$

Now divide equation (9.50) by k^2 and multiply by $\hbar$ to obtain

$$\frac{-\hbar}{i}\frac{\partial \Psi}{\partial t} = -\frac{\partial^2 \Psi}{\partial x^2}\left\{\frac{\hbar^2}{2m_e} + \frac{V}{k^2}\right\} \tag{9.51}$$

After multiplying the bracketed terms by the second derivative of Ψ and then substituting equation (9.12) above in the second term in brackets, we have

$$\frac{-\hbar}{i}\frac{\partial \Psi}{\partial t} = -\frac{\hbar^2}{2m_e}\frac{\partial^2 \Psi}{\partial x^2} + V\Psi \tag{9.52}$$

This equation can be further simplified by separating the spatial and temporal parts of Ψ. To accomplish the separation, we use the time derivative of Ψ from equation (9.11), keeping in mind that ψ is only a function of position. The result is

$$\frac{d\Psi^2}{dx^2} = \frac{d^2\Psi}{dx^2}e^{-i\omega t} \tag{9.53}$$

Now we can substitute these expressions into equation (9.52) above and obtain

$$\frac{d^2\psi}{dx^2}e^{-i\omega t} + \frac{2m_e i}{\hbar}\left(-i\omega\psi e^{-i\omega t}\right) - \frac{2m_e}{\hbar^2}V\left(\psi e^{-i\omega t}\right) = 0 \tag{9.54}$$

After multiplying through by $e^{i\omega t}$ and recognizing that $E = \Sigma\omega$, we obtain in 1-D,

$$\frac{d^2\psi}{dx^2} + \frac{2m_e}{\hbar^2}(E - V)\psi = 0 \tag{9.55}$$

This equation is the typical form for the time-independent SE, and the one that will be used in the following sections.

9.3.3 Classical and QM Wave Equations

Recall from Chapter 7, equation (7.19), that the classical wave equation contains a second spatial and second temporal derivative. This wave equation is easily obtained from equation (9.3) using the earlier representation of a wave:

$$\psi = f(x \pm vt) = A\sin(x \pm vt) \tag{9.56}$$

Consider a wave $y = f(x')$, where $x' = x \pm vt$ with x the distance traveled and v the velocity. This wave may be a harmonic wave where the function is trigonometric, as shown in equation (9.56). The chain rule is used next to develop formulas with position x and time t as variables. The first derivatives are

$$\frac{\partial x'}{\partial x} = 1 \quad \text{and} \quad \frac{\partial x}{\partial t} = \pm v \tag{9.57}$$

With these derivatives, and by the chain rule, we can obtain the following formulas:

$$\frac{\partial y}{\partial x} = \frac{\partial f}{\partial x'}\frac{\partial x'}{\partial x} = \frac{\partial f}{\partial x'}\cdot 1 \quad \text{and} \quad \frac{\partial^2 y}{\partial x^2} = \frac{\partial}{\partial x'}\frac{\partial f}{\partial x'} = \frac{\partial^2 f}{x'^2} \tag{9.58}$$

$$\frac{\partial y}{\partial t} = \frac{\partial f}{\partial x'}\frac{\partial x'}{\partial t} = \frac{\partial f}{\partial x'}\cdot \pm v \quad \text{and} \quad \frac{\partial^2 y}{\partial t^2} = \frac{\partial}{\partial x'}\left(\frac{\partial f}{\partial x'}\cdot \pm v\right)\cdot\frac{\partial x'}{\partial t} = \frac{\partial^2 f}{\partial x'^2}v^2 \tag{9.59}$$

Notice the similarity of the second spatial and temporal derivatives (right sides of equations 9.58 and 9.59). Then equating these derivatives yields

$$\frac{\partial^2 y}{\partial x^2} = \frac{1}{v^2}\frac{\partial^2 y}{\partial t^2} \quad \text{and in 3-D} \quad \nabla^2 y = \frac{1}{v^2}\frac{\partial^2 y}{\partial t^2} \tag{9.60}$$

This is the same classical wave equation we used in Chapter 7 as equation (7.19). However, as was shown above, the SE has a different form:

$$\frac{-\hbar}{i}\frac{\partial \Psi}{\partial t} = -\frac{\hbar^2}{2m_e}\frac{\partial^2 \Psi}{\partial x^2} + V\Psi \tag{9.52}$$

The SE has a first derivative with respect to time. The reason for the different form for the SE lies with the form for the dispersion equation (9.46), which derives from wave-particle duality. Consequently, when the proper dispersion formula that includes duality

is used in the formulation, the SE looks more like a diffusion equation (recall Fick's second law from Chapter 5, equations 5.19 and 5.21) than a classical wave equation. In the following section the SE is applied to arrive at the electronic structure for solids.

9.3.4 Solutions to the SE

Ultimately in this section we will solve the time-independent SE (equation 9.55) for a solid. Our objective will be to find the unperturbed allowed electronic energy levels in the solid. These allowed energy levels comprise the electronic structure of the material. The time-dependent solution to the SE is not useful here because we are interested in the equilibrium or time invariant electronic structure. However, to determine the time evolution of events that can disturb the equilibrium structure such as various spectroscopies that use a perturbing radiation, for example, a complete solution is required. Before proceeding directly to the task of solving the time-independent SE for a solid, we take a slight detour and apply the SE to several simpler problems, namely free and bound electrons. These solutions can yield insight into both the solution method of the SE and the anticipated results. In addition it should be recognized that one can obtain precise solutions by using realistic inputs to the SE. This is possible in some cases, and solutions have been found that are in close agreement with reality. However, the mathematical complexity, even for very simple systems, can be distracting. Thus the strategy we use here is to model systems that do not give precise results but rather physically correct and meaningful results by way of simple mathematics. This will lead us to the solution of the SE for the complicated case of a solid composed of many atoms, even though we do not obtain accurate numerical results.

9.3.4.1 Free Electron Solution to the SE   As was alluded to above, it is sometimes useful, and almost always simpler, to perform calculations on model systems as opposed to real systems. In this spirit we consider totally free electrons, even though we realize that electrons in solids are never completely free of the potential that exists in the solid. However, some electrons in some solids (good conductors, e.g., Cu and Au) behave as if they are free, if not entirely free. Thus, using the model of totally free electrons can yield some insight into how these kind of electrons in good conductors behave.

The time independent SE written in 1-D is

$$\frac{d^2\psi}{dx^2} + \frac{2m_e}{\hbar^2}(E - V)\psi = 0 \tag{9.55}$$

Starting from equation (9.55), the condition(s) that relate to entirely free electrons can be imposed. Essentially the condition for an entirely free electron is that binding potential for the electron be zero: $V = 0$. From this condition the SE becomes

$$\frac{d^2\psi}{dx^2} + \frac{2m_e}{\hbar^2}E\psi = 0 \tag{9.61}$$

The solution for this SE can be written down immediately when we realize that this time-independent SE is virtually identical to the classical wave equation when the time dependence is removed. Thus, from a purely mathematical point of view, we are dealing with the undamped vibrating string problem of classical physics. In this specific case the string is free to vibrate, since it is not held or constrained in any way. The solution for this kind

of vibration can be immediately rationalized. All types of vibrations are possible, and we can write this solution in 1-D as

$$\psi(x) = Ae^{i\alpha x} + Be^{-i\alpha x} \tag{9.62}$$

where A and B are constants to be determined and α is given as $\alpha = k$, at which equation (9.62) is the same as equation (9.40). This can be found by considering that the time-independent SE to be solved is a differential equation of the form

$$a \cdot \frac{d^2 f(x)}{dx^2} + bf(x) = 0 \tag{9.63}$$

This equation can be solved to yield

$$f(x) = Ae^{i\alpha x} + Be^{-i\alpha x} \tag{9.64}$$

with $\alpha = (b/a)^{1/2}$. From the SE above the ratio of the coefficients a and b is $b/a = 2m_e E/\hbar^2$. Recalling that $E = p^2/2m_e$, $p = h/\lambda$, and $\hbar = h/2\pi$, we obtain α as follows:

$$\alpha = \sqrt{\frac{2m_e E}{\hbar^2}} = \sqrt{\frac{2m_e p^2}{\hbar^2 2m_e}} = \frac{p}{\hbar} = \frac{2\pi p}{h} = \frac{2\pi}{\lambda} = k \tag{9.65}$$

Now considering the propagation of the wave in one direction, we obtain the final result:

$$\psi(x) = Ae^{ikx} \tag{9.66}$$

From equation (9.65) for α, where we found $\alpha = k$, we obtain for the energy

$$E = \frac{\hbar^2}{2m_e} k^2 \tag{9.67}$$

This result is revealing of the behavior of free electrons. First we see that $E \propto k^2$, which indicates that all energies are possible. Thus there exists a continuum of allowed electron states corresponding to values of k. Second, the allowed energies are arranged in a parabolic band, and this is shown in Figure 9.6. This parabolic band is known as the free electron band.
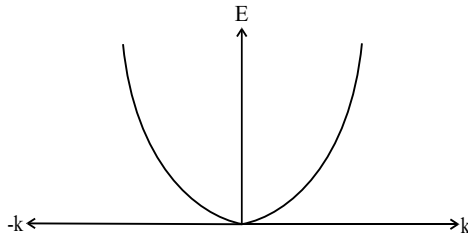
Figure 9.6 Prabolic electron energy band for free electrons.

9.3.4.2 *Strongly and Weakly Bound Electron Solution to the SE*   We can model

the situation for strongly bound electrons by assuming that the electrons are in a potential well of infinite depth. This is the so-called particle in a box formulation in which the walls of the 1-D box are infinitely high. Under this assumption of infinitely high barriers, there is zero probability for the electron to be outside the box. We first solve this problem, and then we relax the strict requirement on the infinitely high walls to finite walls and resolve the weakly bound electron problem. We commence with the 1-D time independent SE as before:

$$\frac{d^2\psi}{dx^2} + \frac{2m_e}{\hbar^2}(E - V)\psi = 0 \tag{9.55}$$

but in this case $V \neq 0$. This problem and the differential equation above are analogous to the classic vibrating string problem that we used for the free electrons, but rather than the string being unconstrained as above for the free electron problem, now the string is tightly (infinitely tightly!) held at two positions along its length. Thus, when the string is plucked in between the pinning points, it will vibrate and sustain those vibrations that do not destructively interfere. Furthermore, because of the infinitely tight pinning, the vibrations cannot propagate beyond the pinned points. Within the pinned points the waves can propagate to the left and right. As for the case above the general solution to the differential equation is as it was above:

$$\psi(x) = Ae^{i\alpha x} + Be^{-i\alpha x} \tag{9.62}$$

The barriers (pinning points) are shown in Figure 9.7. Now this general solution, equation (9.62), needs to be tailored to be in conformity with the specific conditions of the problem, the so-called boundary conditions. With reference to Figure 9.7 the boundary conditions are (1) at 0 and l (the pinning points), $V = \infty$; (2) $\psi = 0$ at $x \leq 0$, and $x \geq l$. These boundary conditions mean that the vibrations do not exist to the right or left of the barriers but do exist in the region 0 to l, so we adjust the general solution accordingly. This kind of problem is called a boundary value problem. The functions to be solved are called eigenfunctions, and the resultant values that solve the problem are called eigenvalues. The condition $\psi = 0$ at $x = 0$ in equation (9.62) yields the following:

$$Ae^{i\alpha 0} + Be^{-i\alpha 0} = 0 \tag{9.68}$$

The solution for equation (9.68) is that $A = -B$. The condition $\psi = 0$ at $x = l$ in equation (9.62) yields the following:
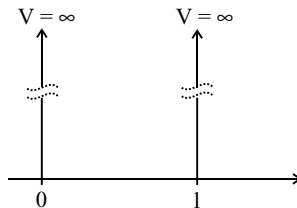
Figure 9.7   1-D potential well of length *l* and with infinite potential barriers.

$$Ae^{i\alpha 1} + Be^{-i\alpha 1} = 0 \tag{9.69}$$

To simplify the analysis of result (9.69), we can make use of an Euler relationship:

$$\sin\rho = \frac{1}{2i}\left(e^{i\rho} - e^{-i\rho}\right) \tag{9.70}$$

For the case of equation (9.69), $\rho = \alpha l$ and $A = -B$, we obtain

$$A\left(e^{i\alpha l} - e^{-i\alpha l}\right) = 2Ai\sin al = 0 \tag{9.71}$$

This equation will hold true if and only if (iff) $\alpha l = n\pi$, where n is an integer, $n = 0, 1, 2, 3. \ldots$. Therefore we can express this condition as

$$\alpha = \frac{n\pi}{l} = \sqrt{\frac{2m_e E}{\hbar^2}} \tag{9.72}$$

and solve for the energy:

$$E = \frac{\hbar^2}{2m_e} \cdot \frac{n^2\pi^2}{l^2} \tag{9.73}$$

The implications of this solution are quite important. Unlike the free electron ($V = 0$) solution (equation 9.67) where all energies were allowed, now as given by equation (9.73) for the tightly bound electrons, only certain values are permitted, specifically those values of E where n is an integer. Thus the energy is quantized in units of n^2. The quantization of the energy is a result of the boundary conditions imposed on the solution of the SE.

Returning to the plucked string analogy used above, we consider that the string is held infinitely tightly at two points, and plucked in between the pinning points. So we ask what wavelengths for the vibrations are permitted that do not suffer from destructive interference. It is clear from Figure 9.8a (for $n = 1, 2,$ and 3) that the wavelengths that survive are those that "fit" integrally in the length of the string (the length of the box, l). These allowed waves constructively interfere and lead to standing waves. All the others destructively interfere and die out. Of particular interest is Figure 9.8b, which shows $\psi\psi^*$, and for real ψ this is ψ^2. The probability is that for $n = 1$ the particle is near the center of the box; for $n = 2$ and higher, the particle has multiple points of high probability. This is a decidedly nonclassical result. For a classical particle in a box the particle would have uniform probability everywhere in the box. Interestingly we can extrapolate to high values of n (e.g., $n = 100$) so that the number of peaks are high, but as n increases, more uniform probability results. For high energies (high n) the particle then behaves classically. It is further useful to consider the allowed energies as a function of n as shown in Figure 9.9a. Recall that equation (9.73) shows that as n increases, E increases as n^2, and the energy levels are more and more separated. This result is different from what is expected for atoms bound in molecules that have a binding potential. For the case of bound atoms, as n increases, the energy levels become more closely spaced up to the ionization level where an electron has enough energy to escape from the atom. This case is shown in Figure 9.9b and is the result of the fact that the form for the binding potential V is decidedly different. Usually the binding energy is referenced to the energy needed for an electron to be removed from the atom, and this level is set as 0, so the binding
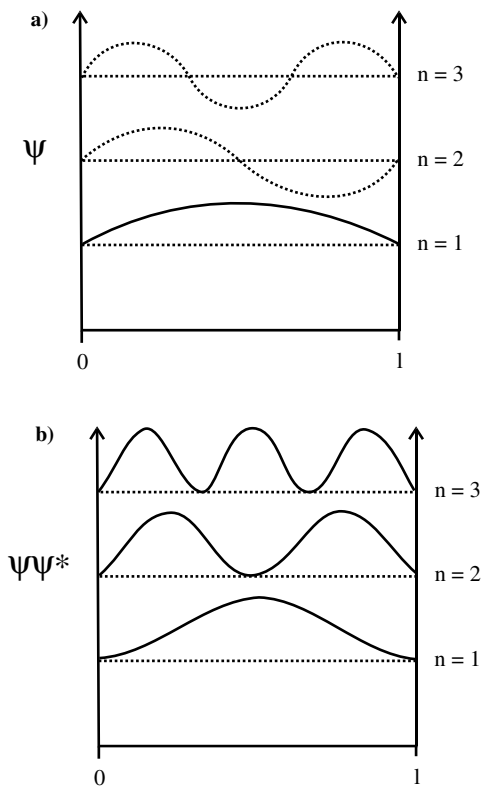
Figure 9.8 (*a*) Three solutions (*n* = 1, 2, and 3) for the bound electron problem; (*b*) probabilities for the three solutions in (*a*).

energies are negative as was shown in Chapter 7, Figure 7.4. Since in Figure 9.9*a*, $V = \infty$, while in Figure 9.9*b*, V is proportional to the reciprocal of the separation that is characteristic of a Coulombic or Morse potential that results in $E \propto 1/n^2$, and hence the closer E levels are at higher n.

We can now modify this picture slightly and release our infinitely strong grip on the string, in order to permit some of the vibrations to "leak" from the bound region to the adjacent regions of the string. This means that we lift the restriction that $V = \infty$ and set V at some finite value and resolve the SE. For this problem we consider an electron wave propagating in 1-D from left to right (+x direction) as is shown in Figure 9.10. Figure 9.10 also shows that the wave is traveling in two different regions with respect to the binding potential. In region 1 the electron is free and $V = 0$, and in region 2 the electron experiences a finite binding potential, $V = V_{\text{finit}}$. The finite binding potential is greater than the kinetic energy of the electron wave, $V_{\text{finit}} > E_{\text{kin}}$, but it is not infinite. The origin $x = 0$ is set at the border of regions 1 ($x < 0$) and 2 ($x > 0$). We need to write separate SE's for the two regions and solve the SE's so that the solutions match at $x = 0$.

The SE for region 1 is the same as the SE above for $V = 0$ and is as follows:
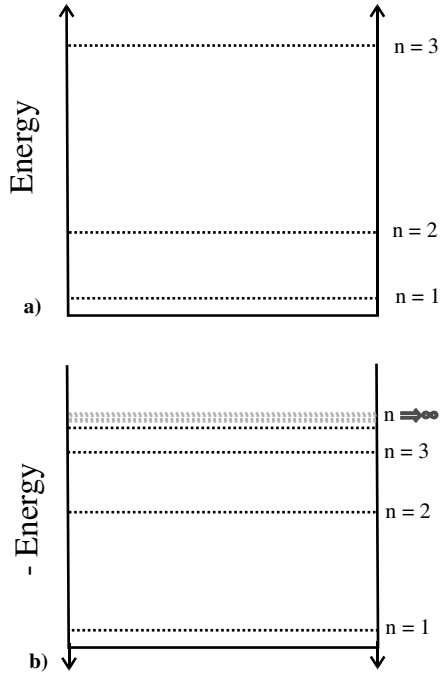
Figure 9.9 (*a*) Energy levels corresponding to equation (9.73) with energy spacing increasing with *n*; (*b*) energy levels for molecular potential with energy spacing decreasing with *n*.
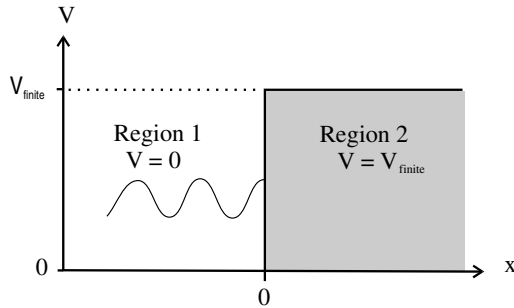
Figure 9.10 Electron in two regions: Region 1 is the free electron region, and region 2 has a finite binding potential.

$$\frac{d^2\psi}{dx^2} + \frac{2m_e}{\hbar^2}E\psi = 0 \qquad \text{(Region 1)} \tag{9.74}$$

And for region 2 where $V = V_{\text{fiinit}}$ the SE is as follows:

$$\frac{d^2\psi}{dx^2} + \frac{2m_e}{\hbar^2}(E - V_{\text{finite}})\psi = 0 \qquad \text{(Region 2)} \tag{9.75}$$

We have already solved these SE's above, so the solutions can be written immediately as follows:

$$\psi_1(x) = Ae^{i\alpha_1 x} + Be^{-i\alpha_1 x} \qquad \text{(Region 1)} \tag{9.76}$$

where

$$\alpha_1 = \sqrt{\frac{2m_e E}{\hbar^2}} \tag{9.77}$$

$$\psi_2(x) = Ce^{i\alpha_2 x} + De^{-i\alpha_2 x} \qquad \text{(Region 2)} \tag{9.78}$$

where

$$\alpha_2 = \sqrt{\frac{2m_e(E - V_{\text{finite}})}{\hbar^2}} \tag{9.79}$$

If $|V_{\text{fiinit}}| > |E|$, then α_2 will be imaginary. The complications arising from this can be avoided by making the following substitution:

$$\beta = i\alpha_2 \tag{9.80}$$

With this substitution, β can be written as follows:

$$\beta = \sqrt{\frac{2m_e(V_{\text{finite}} - E)}{\hbar^2}} \tag{9.81}$$

For the solution in region 2, we obtain the following:

$$\psi_2(x) = Ce^{\beta x} + De^{-\beta x} \tag{9.83}$$

It is now instructive to determine the constants A, B, C, and D for which the details of the problem need to be consulted. We consider the behavior of $\psi_2(x)$ as x approaches ∞ and impose the continuity conditions that $\psi_1(x) = \psi_2(x)$ at the border $x = 0$, and that the derivatives $d\psi_1/dx$ and $d\psi_2/dx$ are also equal at $x = 0$.

As $x \to \infty$, we obtain the following:

$$\psi_2(x) = Ce^{\beta\infty} + 0 \tag{9.84}$$

However, this result is not physical, because as was given by equation (9.43), the probability over all space must be 1:

$$\int_{-\infty}^{\infty} \Psi\Psi^* dV = 1 \tag{9.43}$$

The way to ensure this condition is for $C = 0$, and then $\psi_2(x)$ becomes

$$\psi_2(x) = De^{-\beta x} \tag{9.85}$$

For the condition $\psi_1(x) = \psi_2(x)$ at $x = 0$, we obtain

$$Ae^{i\alpha_1 x} + Be^{-i\alpha_1 x} = De^{-\beta x} \tag{9.86}$$

which at $x = 0$ becomes

$$A + B = D \tag{9.87}$$

For the condition $d\psi_1/dx$ and $d\psi_2/dx$ at $x = 0$, we obtain

$$A i \alpha_1 e^{i\alpha_1 x} - B i \alpha_1 e^{-i\alpha_1 x} = -D\beta e^{-\beta x} \tag{9.88}$$

which at $x = 0$ becomes

$$A i \alpha_1 = B i \alpha_1 = -D\beta \tag{9.89}$$

From the results in (9.89) and (9.87), A and B in terms of D can be obtained as follows:

$$A = \frac{D}{2}\left(1 + \frac{i\beta}{\alpha_1}\right) \quad \text{and} \quad B = \frac{D}{2}\left(1 - \frac{i\beta}{2}\right) \tag{9.90}$$

For our present purposes the most interesting result is the solution above in equation (9.85) where $\psi_2(x) = De^{-\beta x}$ and where $\psi_2(x)$ is shown versus x in Figure 9.11a in region
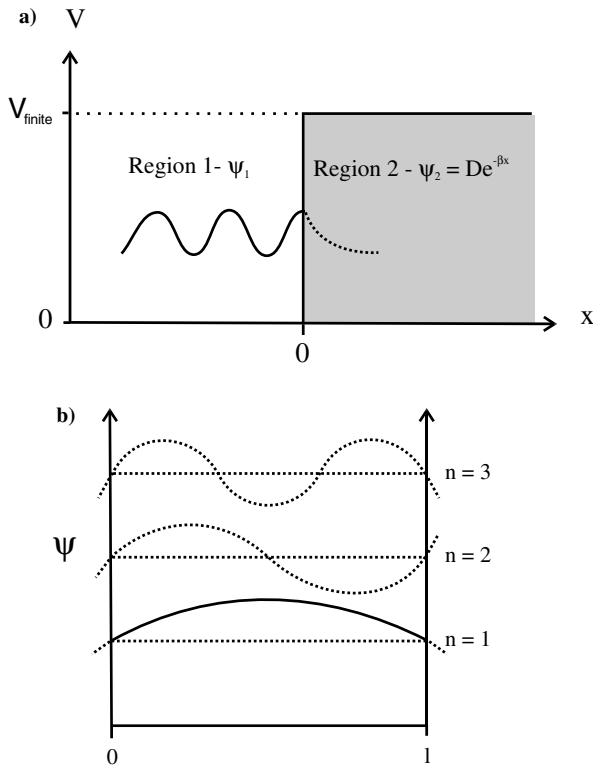
Figure 9.11 (*a*) Solutions to the SE in two regions from Figure 9.10 as ψ_1 and ψ_2; (*b*) other solutions with dashed portion indicating leaking into adjacent regions.

2. Figure 9.11*a* shows the shape of the wave function ψ_2 in region 2 to be an exponentially decaying wave function. Thus the wave function penetrates into region 2, but it decays rapidly. This penetration of the wave function into the barrier is called quantum mechanical tunneling, and the tunneling results from changing the barrier height from infinitely high to finite. Likewise Figure 9.11*b* shows the result for other wave functions corresponding to different *n*'s as was shown in Figure 9.8 above for the infinite barrier. Tunneling is an important result of quantum mechanics because it shows that it is probable for an electron to exist in a region that is classically disallowed. One consequence of tunneling is that if an electronic detector is placed close to the barrier in region 2, the electron is detected (but weakly), and this is the underlying idea for scanning tunneling microscopy that has become an important surface science tool. As will be discussed in Chapter 11, there are a number of electronic devices that operate based on tunneling. Alternatively, as modern devices have gotten smaller in the nanotechnology domain, electrodes in devices are more closely spaced and tunneling can cause excessively high leakage currents between the close electrodes that disrupt normal device operation. Some device researchers see tunneling as a size limitation criterion for electronic devices.

9.3.4.3 *Periodic Solid Solution to the SE–Kronig-Penney Model*   The SE solutions discussed above yield insights about the solution for a solid material. We proceed by first considering that there is an atomic binding potential for electrons that is characteristic of their atom type. Recall Figure 7.4, which showed the shape for an atomic potential. We now imagine an array of atoms, in particular, an ordered array for a single crystal of a particular material. Associated with each atom in the crystal is an atomic potential that characterizes the atom's interaction with adjacent atoms. The array of atomic potentials is this also ordered with the periodicity of the crystal lattice. An example of one kind of ordered potential is shown in Figure 9.12*a* where the real crystal potential is periodic and of a complex shape. The shape shown is that of an essentially Coulombic potential, which depends on the reciprocal of the separation as was mentioned previously. Thus the binding is stronger closer to the atomic core. Different atoms of course display different shaped potentials, and the solution of the SE for one such potential will require modification for another kind of atom. For materials with different kinds of atoms, different structures and morphologies, the situation becomes even more complicated. Yet in those cases in which potentials are available, precise calculations of the electronic structure have been made. The point is that procedures are presently known to solve complicated SE's. These procedures enable precise calculations that are mathematically intense and beyond the scope of this text. Nevertheless, it is important to understand the implications of the calculations and to gain a feel for electronic structure, whether obtained from experiment or theory. To this end a simple model for a solid that yields physically correct, though inaccurate, numerical results is useful. This idea was successfully pursued in the early 1930s and is called the Kronig-Penney model. This model will be outlined below.

The Kronig-Penney (KP) model commences with a simplification of the potential for a periodic solid. Rather than attempting to approximate complex shaped potentials, the KP model uses the conjoining of two regions and the repeating of these regions with crystal periodicity. Figure 9.12*b* shows the KP periodic potential as made up of a potential free region ($V = 0$): region 1 with a width *a* as the separation between atoms (similar to the lattice parameter), and region 2 having a finite binding potential (V_{fiinit}) and width

a)

b)
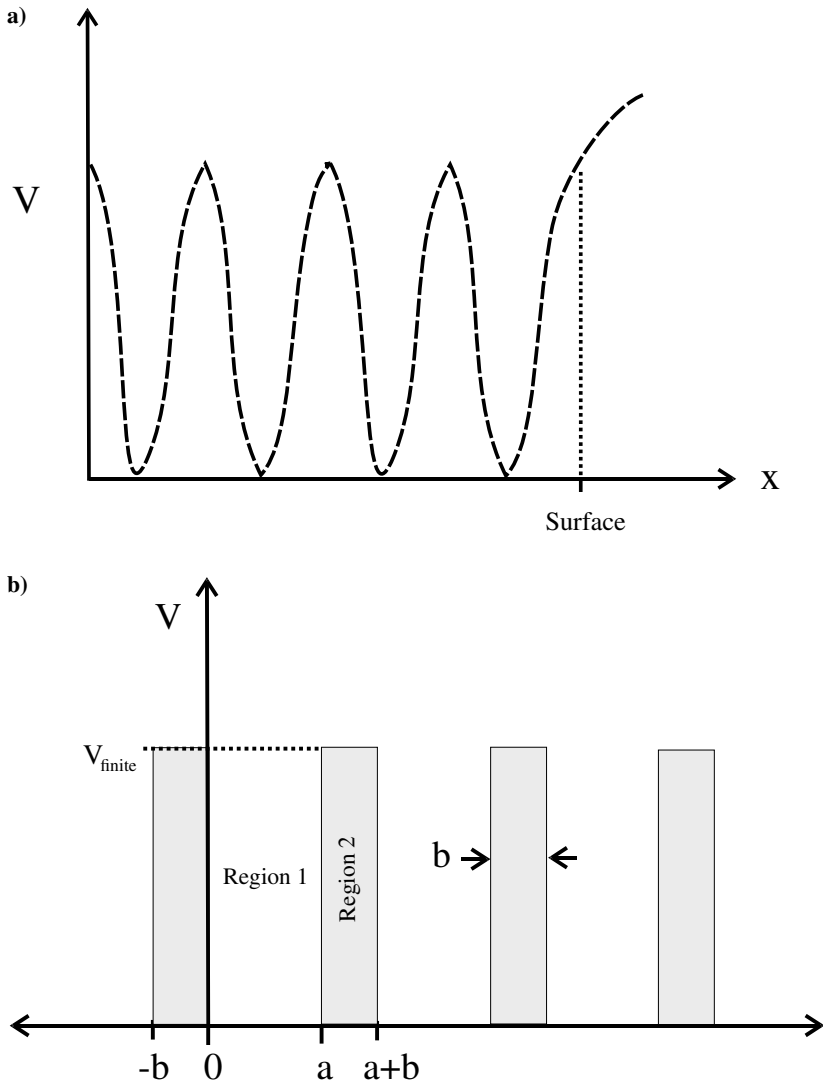
Figure 9.12 (*a*) 1-D periodic binding potential for a solid; (*b*) Kronig-Penney potential for a periodic solid.

b (similar to the barrier width). Although this simple potential cannot yield correct numerical results, as we will see below when the results from the KP model are compared with real electronic structures, the KP model will be found to reveal the correct overall physics, but without some details. Furthermore the KP model is amenable to algebraic analytical solution, and follows directly from the solutions above for the free and bound electrons.

The procedure to obtain a solution of the SE using the KP model is essentially the same procedure as was used above for the finite potential. First, for regions 1 and 2 in Figure 9.12*b*, the two relevant SE's are written

$$\frac{d^2\psi}{dx^2} + \frac{2m_e}{\hbar^2}E\psi = 0 \qquad \text{(Region 1)} \tag{9.91}$$

$$\frac{d^2\psi}{dx^2} + \frac{2m_e}{\hbar^2}(E - V_{finite})\psi = 0 \qquad \text{(Region 2)} \tag{9.92}$$

As before the energies for regions 1 and 2, which are equations (9.77) and (9.79), respectively, are expressed as

$$\alpha = \sqrt{\frac{2m_e E}{\hbar^2}} \quad \text{and} \quad \beta = \sqrt{\frac{2m_e(V_{finite} - E)}{\hbar^2}} \tag{9.93}$$

The difference with the KP model is that the potential is periodic throughout the crystal, and the solution for the two regions must therefore be a simultaneous solution extending over the entire crystal. Solutions for this kind of problem exist, and they are given by the so-called Bloch theorem to yield (in 1-D) the form

$$\psi(x) = u(x)e^{ikx} \tag{9.94}$$

In this expression $u(x)$ is a periodic function with the same period as the barriers $(a + b)$. Notice that from our previous discussion of waves in Section 9.2, this solution is essentially a modulated wave where $u(x)$ is the modulating function. These waves are often referred to as Bloch waves. Since our discussion will be only in 1-D, it should be understood that the periodicity is different in each crystal direction. Thus the solution in 3-D is for more complicated.

The Bloch theorem is used to provide the correct solution for our periodic potential problem. The appropriate derivatives of ψ are taken and the SE is reformatted with these derivatives. The first derivative of the solution $\psi(x)$ in equation (9.94) yields two terms:

$$\frac{d\psi(x)}{dx} = u(x)ike^{ikx} + e^{ikx}\frac{du(x)}{dx} \tag{9.95}$$

The second derivative of equation (9.94) yields four terms:

$$\frac{d^2\psi(x)}{dx^2} = u(x)i^2k^2e^{ikx} + ike^{ikx}\frac{du(x)}{dx} + e^{ikx}\frac{d^2u(x)}{dx^2} + ike^{ikx}\frac{du(x)}{dx} \tag{9.96}$$

With the terms collected, we obtain

$$\frac{d^2\psi(x)}{dx^2} = e^{ikx}\left(\frac{d^2u(x)}{dx^2} + 2ik\frac{du(x)}{dx} - k^2u(x)\right) \tag{9.97}$$

These derivatives are now substituted in regions 1 and 2 of the SE, and the following substitutions are also made from equation (9.93):

$$E = \frac{\hbar^2\alpha^2}{2m_e} \quad \text{and} \quad V_{finite} - E = \frac{\hbar^2\beta^2}{2m_e} \tag{9.98}$$

For region 1, the SE is

$$e^{ikx}\left(\frac{d^2u(x)}{dx^2} + 2ik\frac{du(x)}{dx} - k^2u(x)\right) + \alpha^2\psi(x) = 0 \tag{9.99}$$

We substitute in $\psi(x) = u(x)e^{ikx}$, and the final equation for region 1 becomes

$$\frac{d^2u(x)}{dx^2} + 2ik\frac{du(x)}{dx} - (k^2 - \alpha^2)u(x) = 0 \quad \text{(Region 1)} \tag{9.100}$$

For region 2, the SE is

$$e^{ikx}\left(\frac{d^2u(x)}{dx^2} + 2ik\frac{du(x)}{dx} - k^2u(x)\right) - \beta^2\psi(x) = 0 \tag{9.101}$$

After substituting $\psi(x) = u(x)e^{ikx}$, as was done for region 1 above, we obtain for region 2,

$$\frac{d^2u(x)}{dx^2} + 2ik\frac{du(x)}{dx} - (k^2 + \beta^2)u(x) = 0 \quad \text{(Region 2)} \tag{9.102}$$

Regions 1 and 2 SE's can be seen to have the same form as the differential equation for a damped vibration in classical physics. If $u(x)$ is defined as the displacement for the vibration, then the form is as follows:

$$\frac{d^2u(x)}{dx^2} + A\frac{du(x)}{dx} + Bu(x) = 0 \tag{9.103}$$

This equation has a general solution of the form:

$$u(x) = e^{-Ax/2}(Ce^{i\xi x} + De^{-i\xi x}) \quad \text{where } \xi = \sqrt{B - \frac{A^2}{4}} \tag{9.104}$$

For regions 1 and 2, $A = 2ik$; for region 1, $B = -(k^2 - \alpha^2)$; and for region 2, $B = -(k^2 - \beta^2)$. Putting this together obtains the following solutions:

$$u(x) = e^{-ikx}(Ce^{i\alpha x} + De^{-i\alpha x}) \quad \text{(Region 1)} \tag{9.105}$$

$$u(x) = E^{-ikx}(Ae^{i\beta x} + Be^{-i\beta x}) \quad \text{(Region 2)} \tag{9.106}$$

A, B, C, and D are constants that need to be determined by the specific conditions of the problem, namely that the solutions and their derivatives are continuous at the boundary ($x = 0$) and at all periods ($x = a + b$). Thus we have four equations in the four unknowns A, B, C, and D.

At $x = 0$ the solutions given by equations (9.105) and (9.106) and the derivatives are equated (subscripts 1 and 2 are used to indicate the regions) to yield the first two equations as follows:

$$\text{For } u_1(x) = u_2(x) \text{ at } x = 0: \quad C + D = A + B \tag{9.107}$$

$$\text{For } \frac{du_1(x)}{dx} = \frac{du_2(x)}{dx} \text{ at } x = 0:$$

$$Ci(\alpha - k) + Di(-\alpha - k) = Ai(\beta - k) + Bi(-\beta - k) \tag{9.108}$$

Also two equations can be obtained for the periodic boundary condition $x = a + b$. These equations are most easily obtained by using Figure 9.12b and realizing that $u_1(x)$ at a must equal $u_2(x)$ at $-b$, and likewise for the derivatives, and this requirement yields the following two equations:

$$\text{For } u_1(x) \text{ (at } x = a) = u_2(x) \text{ (at } x = -b):$$

$$Ce^{(i\alpha - ik)a} + De^{(-i\alpha - ik)a} = Ae^{(ik - i\beta)b} + Be^{(ik + i\beta)b} \tag{9.109}$$

$$\text{For } \frac{du_1(x)}{dx} \text{ (at } x = a) = \frac{du_2(x)}{dx} \text{ (at } x = -b):$$

$$Ci(\alpha - k)e^{ia(\alpha - k)} - Di(\alpha + k)e^{-ia(\alpha + k)} = Ai(\beta - k)e^{-ib(\beta - k)} - Bi(\beta + k)e^{ib(\beta + k)} \tag{9.110}$$

Now there are four independent equations in the four unknowns A, B, C, and D, namely equations (9.107) through (9.110). A simultaneous solution is obtained by forming a determinant of the coefficients C, D, A, B, and setting the determinant equal to zero. The determinant has the form:

$$\begin{vmatrix} 1 & 1 & 1 & 1 \\ \alpha - k & -(\alpha + k) & \beta - k & -(\beta + k) \\ e^{ia(\alpha - k)} & e^{-ia(\alpha + k)} & e^{-ib(\beta - k)} & e^{ib(\beta + k)} \\ (\alpha - k)e^{ia(\alpha - k)} & -(\alpha + k)e^{-ia(\alpha + k)} & (\beta - k)e^{-ib(\beta - k)} & -(\beta + k)e^{ib(\beta + k)} \end{vmatrix}$$

The result after setting the determinant equal to 0, considerable algebra and applying Euler's formulas is as follows:

$$\frac{\beta^2 - \alpha^2}{2\alpha\beta} \sinh(\beta b) \cdot \sin(\alpha a) + \cosh(\beta b) \cdot \cos(\alpha a) = \cos(k(a + b)) \tag{9.111}$$

This result can be further simplified using several physically relevant assumptions. The first is to assume that V_{finite} is large compared with the kinetic energy for the electron, so that from equation (9.93) the expression for β above becomes

$$\beta = \sqrt{\frac{2m_e V_{\text{finite}}}{\hbar^2}} \quad \text{and} \quad \beta b = \sqrt{\frac{2m_e V_{\text{finite}} b^2}{\hbar^2}} \tag{9.112}$$

Also, since $\beta \propto V_{\text{finite}}^{1/2}$ and $\alpha \propto E^{1/2}$, then $\beta > \alpha$ and $\beta^2 >> \alpha^2$. Thus α^2 can be removed from the first term in equation (9.111). Now we assume that the electron binding potential drops off sharply at the atomic cores. Effectively this assumes a narrow barrier. Thus, as $b \to 0$, βb becomes small. The cosh of a small argument is 1 and the sinh of a small argument is the argument. Including all these assertions simplifies the KP results in equation (9.111) to the following:

$$\frac{\beta^2}{2\alpha\beta}(\beta b) \cdot \sin(\alpha a) + \cos(\alpha a) = \cos(ka) \tag{9.113}$$

Upon substitution, using equation (9.112), this becomes

$$\frac{m_e b V_{\text{finite}}}{\hbar^{2\alpha}} \cdot \sin(\alpha a) + \cos(\alpha a) = \cos(ka) \tag{9.114}$$

It is typical to define a term P as follows:

$$P = \frac{a m_e b V_{\text{finite}}}{\hbar^2} \tag{9.115}$$

The final form from the KP analysis is then

$$\frac{P}{\alpha a} \cdot \sin(\alpha a) + \cos(\alpha a) = \cos(ka) \tag{9.116}$$

The final simplified form, equation (9.116), is called the KP formula. This equation is of great interest because from it important revelations about electronic structure of solids can be directly made, as we will see below. First, we should observe that the right-hand side of the KP formula is a cosine function with the argument (ka). The only permissible values for the left-hand side of the final form then must lie between 1 and −1. The first term on the left is essentially $\sin(\alpha a)/\alpha a$, and P is a composite constant defined above in equation (9.115). The $\sin(\alpha a)/\alpha a$ function is called a sinc function, and it is characterized as a periodic function with decreasing amplitude, as shown in Figure 9.13a. Recall from equation (9.93) that α is an expression of the energy for the electron. Thus the left-hand side of the final KP formula defines all the energies for electrons. When all these energies are bound by the right-hand side of equation (9.116), the allowed energies lie between values of +1 and −1 for the left-hand side of equation (9.116). These allowed energy regions are called allowed energy bands. Figure 9.13b shows a plot of the left-hand side of the KP formula versus αa (or energy) for two values of P; one value of P is 2.5 times the other for comparison. Also plotted are the horizontal lines at values +1 and −1 that form the boundaries for allowed electron energy solutions, as dictated by the right-hand side of the KP formula (equation 9.116). The scales are conveniently chosen, and do not reflect real numbers. However, the shape of the solution in terms of allowed energy bands is evident.

The KP formula plotted in Figure 9.13b shows several important features. From the zero of energy at the center the KP function, there is a sharp rise. Then the function decreases traversing the +1 to −1 region in its decent; it repeats with ever-decreasing amplitude until all the values lie between +1 and −1. The values of energy that the KP function (f_{KP}) takes in the region $+1 \geq f_{\text{KP}} \geq -1$ are the allowed energies. This region is called an allowed energy band. It is also seen in Figure 9.13b that the plot corresponding to higher P (dashed line) traverses the +1, −1 region more steeply then the lower P plot (solid line). This means that for higher P (and V) there is a narrower range of allowed energies. Note that both before and after an allowed band there is a region of energies that lies above or below the allowed region. These regions are obviously not allowed energies, so they are called energy band gaps. The electronic structure is therefore composed of allowed energy bands separated by disallowed energy gaps, or more simply of bands and gaps.
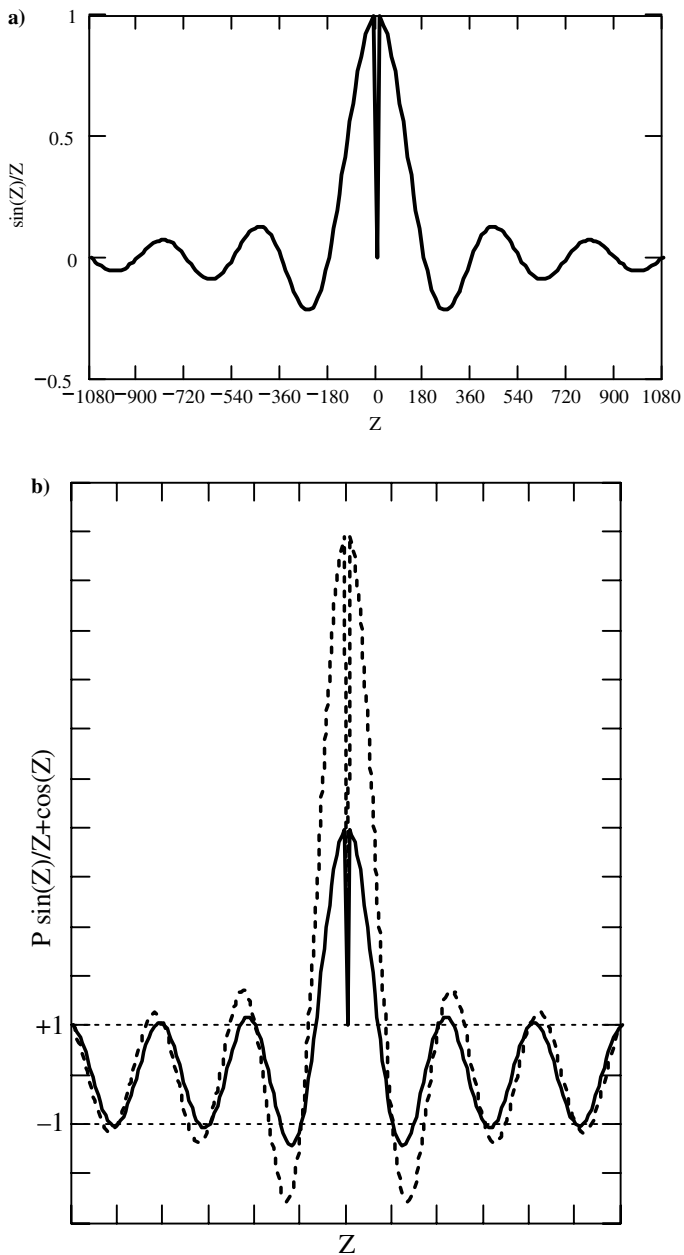
Figure 9.13 (*a*) Plot of sinc (*Z*) function where *Z* is an angle θ; (*b*) plot of left side of KP formula equation (9.116) where the argument *Z* is αa.

9.4   ELECTRON ENERGY BAND REPRESENTATIONS

9.4.1   Parallel Band Picture

The allowed bands and disallowed gaps derived from the KP model can be displayed in a parallel band picture, as shown in Figure 9.14. In this picture all information on how an energy band can vary with direction is lost. Nevertheless, the simple picture contributes to our understanding of how lower electron energy bands are dominated by the interatomic potential. Consequently these bands are narrow and widely separated. In contrast, the bands that house the higher energy electrons are wider and more closely separated. Exactly how these bands are occupied determines the overall electronic properties of the material. We will return to this important point of occupancy in Chapter 10. For now we consider that for an electron to move (i.e., for electronic conduction), quantum mechanics requires that electrons be in allowed states and that there be empty allowed quantum states accessible in energy for the electrons to move into. Another way to express this is that filled and empty allowed states are required. The KP model yields the allowed states for a given potential. Theoretically these bands of states extend to infinite energy, so there are infinite states available. However, the atomic number for an atom dictates how many electrons are apportioned to each atom in the solid. Thus there are a finite number of electrons that occupy the allowed energy bands starting from the lowest electron energies to the highest lying bands. This leaves the outermost bands sometimes filled and sometimes not filled. The last band to have any occupying electrons is called the valence band (VB). After the VB there is a gap and then another band that is empty. This next empty band is called the conduction band (CB). Three important situations arise and are depicted in Figure 9.15.

Figure 9.15*a* shows the situation with a completely filled VB and empty CB, and a relatively large band gap. Figure 9.15*b* shows the same situation as Figure 9.15*a* except that the band gap is relatively narrow. Figure 9.15*c* show a partially filled VB that also
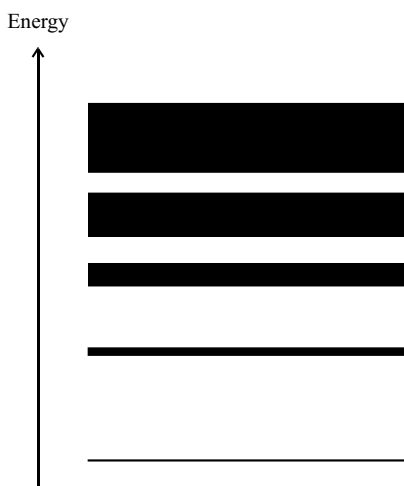
Figure 9.14   Parallel energy band representation with allowed electron energy bands (*shaded*) separated by disallowed energy gaps.
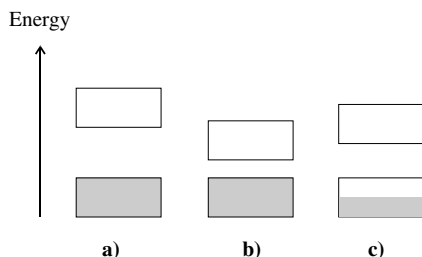
Figure 9.15   Valence band (*lower*) and conduction band with various amounts of electrons (*shaded*) for (*a*) a wide band gap material such as an insulator, (*b*) a narrow band gap material such as a semiconductor, and (*c*) a metallic material.

has empty states in the VB. Many variations are possible in a partially filled VB. Starting with Figure 9.15c, note that the VB has both filled and unfilled states in close energy proximity. Consequently the application of a small potential enables electron flow or conduction. The kinds of materials that have this energy band structure are good conductors and include metals. By contrast, Figure 9.15a shows no empty states in the VB and an energetically far away but empty CB. Because the CB is energetically unavailable, even the application of strong potential will not enable electronic conduction. So this kind of materials is a nonconductor or an insulator. In between these two kinds of behavior is Figure 9.15b with a filled VB and empty CB, but the CB is significantly closer in energy to the VB than in Figure 9.15a. Therefore there are methods to enable conduction such as the application of a reasonable external potential, and this kind of material is called a semiconductor.

In Chapter 10 these different kinds of materials will be discussed at some length, and the energies of the gaps that distinguish the materials will be quantified. However, in order to get a semiconductor to conduct, a special process called doping is usually required. This process actually adds either filled or empty states to the band gap, thereby narrowing the gap and enabling conduction. Before discussing the practical side of electronic structure, it is useful to first consider the energy band representations and their implications.

9.4.2   k Space Representations

In equation (9.115) we saw that when $P = 0$, the electrons are free ($V = 0$). The KP formula for this case is given by substituting $P = 0$ into equation (9.116). This yields

$$\cos(\alpha a) = \cos(ka) \tag{9.117}$$

From this we deduce that $\alpha = k$. Recall equation (9.65) for free electrons:

$$E = \frac{\hbar^2 \alpha^2}{2m_e} \tag{9.65}$$

With $\alpha = k$ this yields equation (9.67):

$$E = \frac{\hbar^2 k^2}{2m_e} \tag{9.67}$$

This equation shows the parabolic dependence of E with k, as was previously discussed and illustrated in Figure 9.6. From equation (9.117), which is the KP formula for $P = 0$, and further from the periodicity of the cosine function, the following is also true in 1-D:

$$\cos(\alpha a) = \cos(k_x a) = \cos(k_x a + n2\pi) \tag{9.118}$$

Equating the cosine arguments and substituting for α from the formula above, and then solving for E, we obtain

$$\alpha a = k_x a + 2n\pi \tag{9.119}$$

Thus

$$E = \frac{\hbar^2 \alpha^2}{2m_e} = \frac{\hbar^2}{2m_e}\left(k_x + \frac{2n\pi}{a}\right)^2 \tag{9.120}$$

This is a formula for a family of parabolas, each centered at $2n\pi/a$, where n is 0, 1, 2, . . . , or even multiples of π/a and part of this family is shown in Figure 9.16. Notice that the parabolas intersect at all multiples of π/a. Recall from the KP formula that the left-hand side, $\cos(ka)$, cannot exceed values of ± 1. Values of the right-hand side of the KP formula that exceed this limit represent disallowed electron energy states. This will occur when the argument of $\cos(k_x a)$, $k_x a = n\pi$ or $k_x = n\pi/a$, where $n = \pm 0, 1, 2, \ldots$ Thus exactly where the parabolas cross, the energy at the crossing is disallowed. In spectroscopy this is called a noncrossing rule. In order to represent the disallowed crossings, at the points of the crossings the parabolas are distorted to prevent crossing, and the result is shown in Figure 9.17 for the parabola centered at 0. This representation of allowed and disallowed energies is called the extended zone scheme. All the same information can be compressed into the first allowed energy zone, $\pm\pi/a$ by translating the pieces of the extended zone scheme by $n2\pi/a$, where $n = 1, 2, 3 \ldots$, and the result is shown along with the extended zone scheme in Figure 9.18 where the arrows indicate the translations. The region in between the vertical dashed lines represents the result of the translations of the parts of the extended zone scheme (as indicated by the arrows) into the first allowed energy zone, and this representation is called the reduced zone scheme.
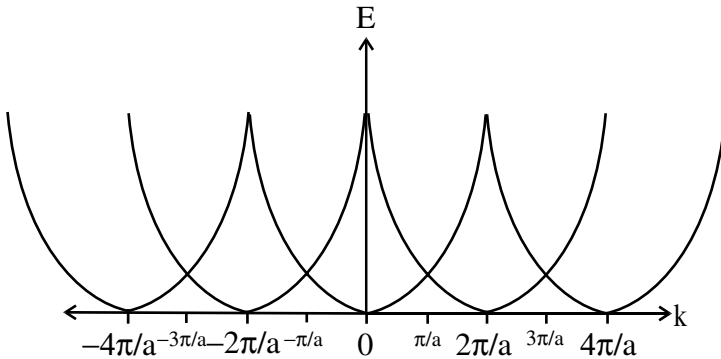
Figure 9.16   Free electron parabolic bands intersecting at $n\pi/a$, where n is 0, 1, 2. . . .
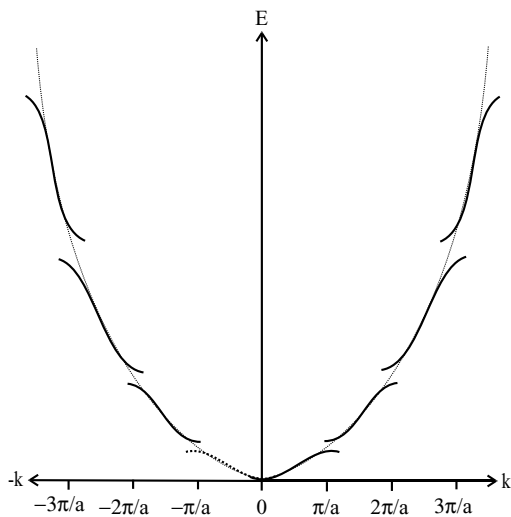
Figure 9.17 One free electron parabolic energy band showing distortions near the disallowed energies at $n\pi/a$, where n is an integer.
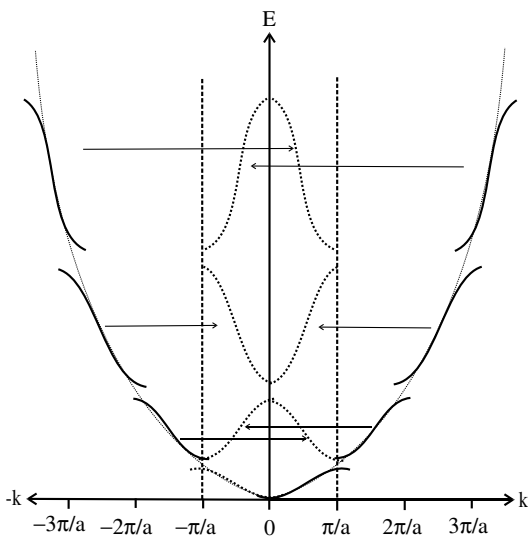
Figure 9.18 Reduced zone scheme where segments of the free electron parabola's are translated (*arrows*) into the first Brillouin zone.

The useful feature of the reduced zone scheme, as shown in Figure 9.19, is that the allowed electron energy bands and disallowed gaps are readily seen by scanning vertically in the representation. To the right in the figure the allowed energy bands are shown in cross hatch and separated by band gaps. The parabolic shape of the free electron bands
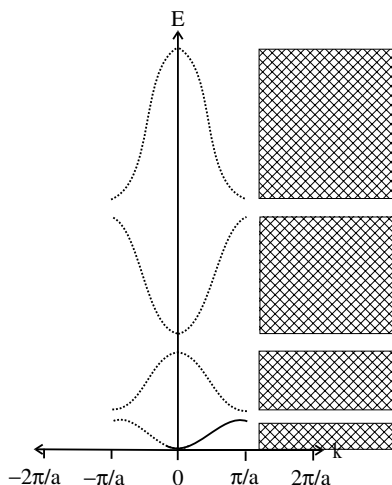
Figure 9.19   Reduced zone representation of electron energy band structure with allowed energy bands indicated by shaded areas at the right.

is seen near the mid part of the zone. The altered shape is near the band edges where the forbidden energies occur.

9.4.3   Brillouin Zones

At the end of Chapter 3 the Wigner-Seitz unit cell in **k** space was introduced, and the reciprocal space was called a Brillouin zone. In this chapter and up to this section we have dropped the vector (bold) notation for **k** that was appropriately used in Chapter 3. This is because thus far in this chapter we have mainly been interested in the magnitude of **k** and not its direction. Now we again take up the vector designation for **k** because the direction as well as the magnitude are being considered. Using Figure 3.18, we can see that waves propagating in a crystal that have wavelengths of the order of the size of the unit cell can diffract in certain appropriate directions where the diffraction conditions are met. Namely the wavelength or integral multiples of the wavelength fit in a space between scattering atoms or molecules. This is called the Bragg condition. Specifically the diffraction conditions are satisfied using equation (3.59) when the **k** space vector $\mathbf{G} = n\mathbf{a}^*$. With $\mathbf{a}^* = 2\pi/\mathbf{a}$ for an orthogonal system and recalling that in terms of **k**, diffraction occurs at $\mathbf{G}/2$, the diffraction condition in terms of **k** was found to be $\mathbf{k} = \pm n\pi/\mathbf{a}$, where $n = 1, 2, \ldots$ (see equation 3.60). Recall that Figure 3.18 gave the **k** space representation of diffraction where the first and second Brillouin zones are within $\mathbf{G}_1/2$ and $\mathbf{G}_2/2$, respectively, for **G** drawn to the nearest and next nearest neighbors. This figure from Chapter 3 is slightly modified to include the diffraction condition for the first Brillouin zone at $\pi/\mathbf{a}$ and displayed as Figure 9.20.

Here in Chapter 9 we have revisited diffraction from another vantage point. Specifically we found that for free electrons in a cubic lattice, the allowed electron energies extend in **k** space from 0 to $n\pi/\mathbf{a}$, at which position there exists a gap in the allowed elec-
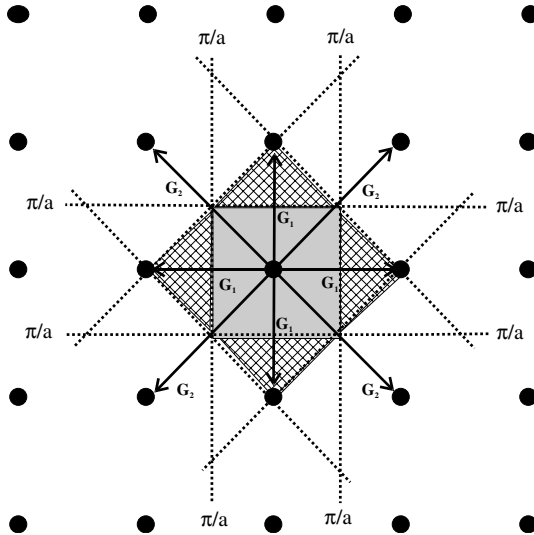
Figure 9.20 First (*shaded*) and second (*cross-hatched*) Brillouin zones in 2-D.

tron energies. This gap refers to those energies for electron waves that are not permitted to propagate in the crystal. In effect these energies (with corresponding wavelengths) are diffracted out of the crystal. This notion gives rise to another fully consistent approach to obtain electron energy band structure called the Ziman approach after its proponent.

Using Figure 9.20 and a little imagination, it is easy to construct a 3-D Brillouin zone representation for a simple (primitive) cubic by adding the other two dimensions and this is represented in Figure 9.21a with the boundaries indicated. It is more difficult to construct the Brillouin zones for more complicated structures, but the procedure is the same. Figure 9.21b and 9.21c show the first Brillouin zones for the BCC and FCC lattices. Also shown in Figure 9.21b and 9.21c are some useful term symbols that have been adopted. In these figures it is seen that the origin of **k** space is designated by the symbol Γ. From the figure the following directions from the origin Γ in the direction indicated are labeled as follows:

Γ → [100] is named Δ
Γ → [110] is named Σ
Γ → [111] is named Λ

The endpoints for the zone are also labeled. The endpoint for the first zone in the Δ direction is *H* for BCC and *X* for FCC; the endpoint for the first zone in the Σ direction is *N* for BCC and *K* for FCC; the endpoint for the first zone in the Λ direction is *P* for BCC and *L* for FCC. These symbols are used to denote a band in a 2-D representation to be introduced below.
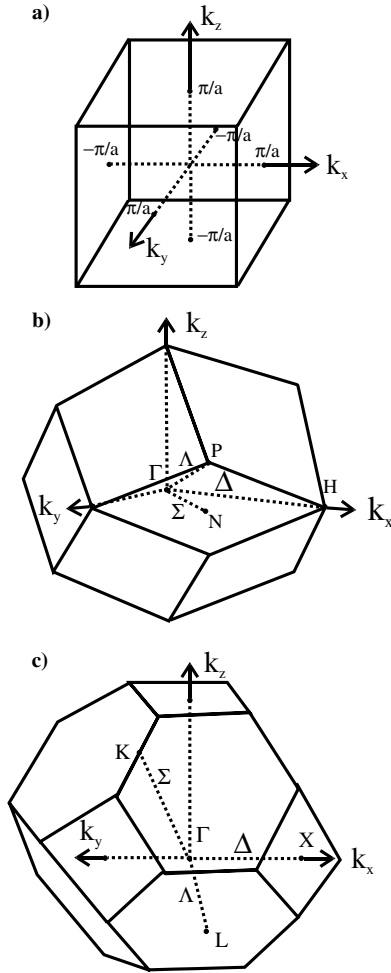
Figure 9.21 3-D Brillouin zones for (*a*) PC, (*b*) BCC, and (*c*) FCC structures with term symbols for important directions.

9.5 REAL ENERGY BAND STRUCTURES

Figure 9.22*a* shows the energy band structure for the metal Al in several of the important crystallographic directions in its FCC structure. A metal such as Al is chosen first because it would more closely approximate a $V = 0$ material. It is easily noticed that many of the bands and parts of the bands are parabolic in shape, thereby giving credence to the earlier assertion about the $V = 0$ approximation. It should also be noticed that several of the bands are not parabolic, an indication that in reality $V \neq 0$ even for a good conductor, although the approximation is useful for this material. In the $\Gamma \rightarrow X$ direction
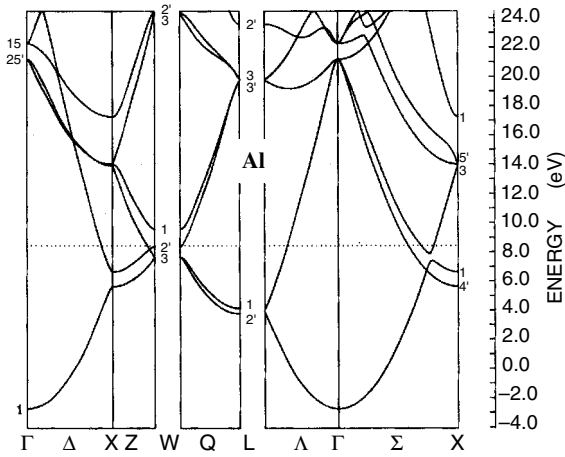
(*a*) **Energy band structure for Al. (Adapted from *Handbook of the Band Structure of Elemental Solids*, D. A. Papaconstantopoulos, Plenum Press, 1986)**
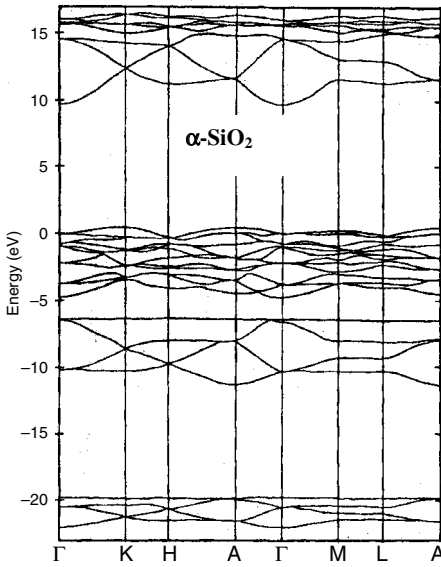
(*b*) **Energy bandstructure for α-quartz. (Adapted from *Physical Review B*, R. B. Laughlin, J. D. Joannopoulos, and D. J. Chadi, vol. 20, p. 5228, 1979)**

Figure 9.22   Electron energy band structure for (*a*) Al, (*b*) α-SiO$_2$, (*c*) Si, and (*d*) GaAS.
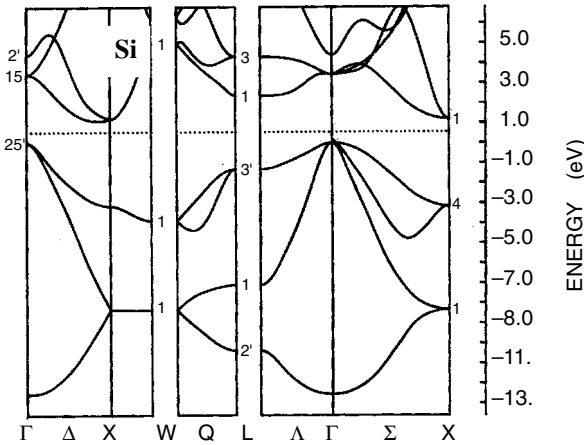
(*c*) **The energy band structure for Si. (Adapted from** *Handbook of the Band Structure of Elemental Solids*, **D. A. Papaconstantopoulos, Plenum Press, 1986)**
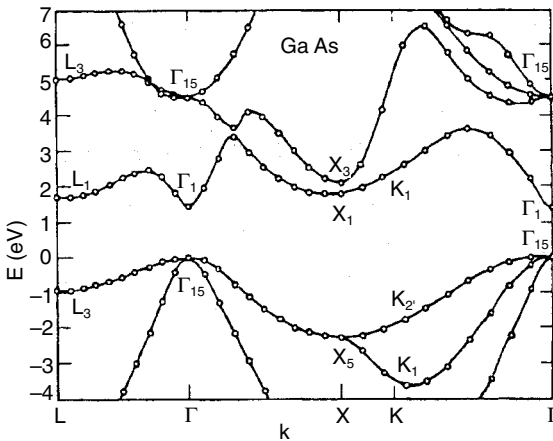
(*d*) **Energy band structure for GaAs. (Adapted from** *Physical Review*, **M. I. Cohen and T. K. Bergstresser, vol. 141, p. 789, 1966)**

Figure 9.22   *Continued*

the lower parabolic band and the upper band show a gap in between the bands. Thus in this particular direction an electron near the top of the lower band is precluded from attaining the upper band states unless the electron at the top of the lower band receives energy greater than the gap between the bands. However, if the electron changes direction, then there are directions in which there is no gap. In fact for Al at any energy on the diagram there is always a direction that can be found in which there is no

energy gap. The ability or process by which an electron can probabilistically "find" an appropriate direction is called percolation. The idea of percolation is derived from water percolating up through sand from a spring. The water finds paths around the grains of sand.

In contrast to a free electron kind of electron energy band structure is the energy band structure for SiO_2 that is shown in Figure 9.22b. Notice that with a good imagination one can see some parabolic regions in the band structure. However, there are large energy gaps, and there are no percolation directions that could enable an electron to percolate to the higher energy levels without a large input of energy. In between these extreme cases for electron energy band structure are the band structures for Si and GaAs shown in Figure 9.22c and 9.22d, respectively. For both of these so-called semiconductors, some band regions are parabolic. Like SiO_2 both of these materials have band gaps in all directions. However, the minimum gap for SiO_2 is around 9 eV, and the minimum gaps for Si and GaAs are about 1.1 eV and 1.4 eV, respectively. As we delve into semiconducting materials more deeply in Chapter 10, we will see that even a 1 eV gap is large relative to the energy available at room temperature (kT at room $T \approx 0.025$ eV). Thus the electrons aren't very "free" to migrate in these semiconducting materials as they do in the metals. In fact, unless these nearly 1 eV gap materials are doped to add appropriate energy levels, they are reasonably good electrical insulators.

In observing the band structure of Si and GaAs, in particular the gaps, the gap for GaAs of about 1.4 eV, is observed to be a vertical transition from a band that is concave downward to one that is concave upward. This transition takes place without a change in $\mathbf{k}$ (the horizontal axis in Figure 9.22) and is called an optical transition (or a $\mathbf{k} = 0$ transition), since none of the energy in the transition is lost to the GaAs lattice. In contrast, the minimum gap in Si of 1.1 eV takes place from $\Gamma \to X$ with a change in $\mathbf{k}$, a nonvertical transition. This means that momentum is transferred to the Si lattice. The main implication of this band feature, namely $\mathbf{k} = 0$ or $\mathbf{k} \neq 0$ transition, is that those materials with $\mathbf{k} = 0$ band gaps are more useful for optical devices where the optical transition, $\mathbf{k} = 0$, has a higher probability to absorb and emit a photon. For Si, for example, the minimum gap at 1.1 eV is a nonoptical transition, but there are optical transitions in Si at 3.4 and 4.3 eV resulting in peaks in the optical absorption at those energies.

9.6 OTHER ASPECTS OF ELECTRON ENERGY BAND STRUCTURE

Returning again to the simplest case for free electron energy bands in a simple cubic solid, the parabolic shape was the main characteristic of the bands in 1-D. We can now imagine the 1-D parabola in Figure 9.6 rotated around the energy axis to sweep out a funnel-like 3-D parabolic figure, as shown in Figure 9.23a. Horizontal cuts in this figure are circles that represent different energies. In Figure 9.23b the circular cuts are stacked on top of the first Brillouin zone with smaller energies being smaller circles. Now for a given Brillouin zone size, the largest energy circle that just fits inside the zone is shown at the border of the Brillouin zone. Larger energies are allowed in the zone, but only near the corners of the Brillouin zone. This is illustrated by the outermost circle that has four of its arcs within the first zone, and four arcs penetrating beyond the zone boundary and into the second zone. Also the higher energies in the first zone are found in the corners of the zone where there are fewer electron states for these higher energies. We will return to this point in Chapter 10 when we consider the density of allowed electron states where we will find a decrease in the density of states for higher energies.
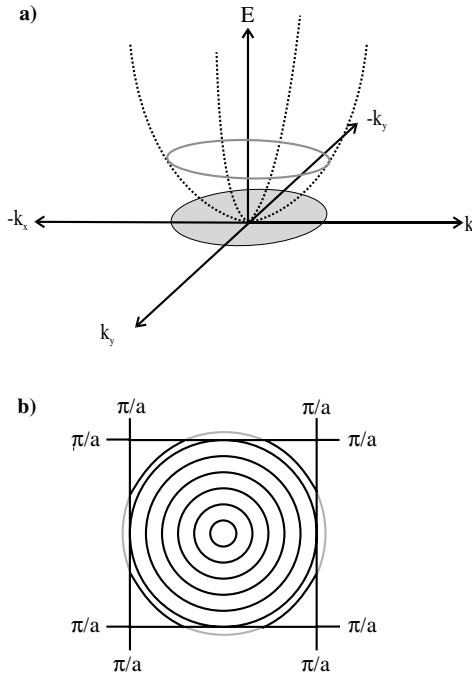
Figure 9.23 (*a*) Free electron bands in 3-D; (*b*) projection of vertical cuts in the bands in (*a*).

The last topic for this chapter is a discussion of electron mass. It will be shown below that the mass for an electron is not a constant in a material, but rather mass is different depending on which band the electron in question resides. Thus the term "effective mass," often written as m^*, is used to denote this quantum mechanical mass.

Recall the formula developed earlier for the group velocity for a wave packet, $\mathbf{v_g}$:

$$\mathbf{v_g} = \frac{d\omega}{d\mathbf{k}} \tag{9.38}$$

Substituting $2\pi\nu = \omega$ yields

$$\mathbf{v_g} = \frac{d(2\pi\nu)}{d\mathbf{k}} \tag{9.121}$$

Substitution $E = h\nu$, we obtain

$$\mathbf{v_g} = \frac{d(2\pi(E/h))}{d\mathbf{k}} = \frac{1}{\hbar}\frac{dE}{d\mathbf{k}} \tag{9.122}$$

The strategy is to make an analogy to the relationship $\mathbf{F} = m\mathbf{a}$ and solve for m using our development above. The acceleration $\mathbf{a}$ is given as

$$\mathbf{a} = \frac{d\mathbf{v_g}}{dt} = \frac{1}{\hbar}\frac{d}{dt}\left(\frac{dE}{d\mathbf{k}}\right) = \frac{1}{\hbar}\frac{d^2E}{d\mathbf{k}^2}\frac{d\mathbf{k}}{dt} \tag{9.123}$$

We can find an expression for $d\mathbf{k}/dt$ using $\mathbf{p} = \hbar\mathbf{k}$:

$$\frac{d\mathbf{p}}{dt} = \hbar\frac{d\mathbf{k}}{dt} \quad \text{and thus} \quad \frac{d\mathbf{k}}{dt} = \frac{1}{\hbar}\frac{d\mathbf{p}}{dt} \tag{9.124}$$

Then using $\mathbf{p} = m\mathbf{v}$, we can write for $\mathbf{a}$,

$$\mathbf{a} = \frac{1}{\hbar^2}\frac{d^2E}{d\mathbf{k}^2}\frac{d\mathbf{p}}{dt} = \frac{1}{\hbar^2}\frac{d^2E}{d\mathbf{k}^2}\frac{dm\mathbf{v}}{dt} \tag{9.125}$$

With substitution of $\mathbf{F} = m\mathbf{a} = md\mathbf{v}/dt$, the final relationship is now written

$$\mathbf{a} = \frac{1}{\hbar^2}\frac{d^2E}{d\mathbf{k}^2}\mathbf{F} \tag{9.126}$$

This means that the coefficient of $\mathbf{F}$ in this formula is $1/m$ or in terms of the effective mass, m^*:

$$m^* = \hbar^2\left(\frac{d^2E}{d\mathbf{k}^2}\right)^{-1} \tag{9.127}$$

Calculus teaches that the second derivative of a function $E(\mathbf{k})$ yields the curvature of the function. Thus m^* is proportional to the reciprocal of the curvature or $m^* \propto 1/\text{cur-}$ vature of a particular band. Furthermore the radius of curvature is given as the reciprocal of the curvature. The result is that the effective mass is then proportional to the radius of curvature of an energy band, the $E(\mathbf{k})$ versus $\mathbf{k}$ curve. For now we can apply this idea to the free electron bands seen in Figure 9.19. Near the bottom of the lowest energy band, the band is curved. Thus it has a large curvature at this position, and hence a small effective mass. Also near the zone edge there is similarly large curvature. However, in between these curved regions the curvature is quite small, yielding a large effective mass. Thus m^* is relatively large in the center of the band between 0 and $\pi/\mathbf{a}$ compared to adjacent end regions of the band that have higher curvature and thus a smaller m^*. Interestingly the curvature also changes from concave up (positive value) near 0 to concave down near $\pi/\mathbf{a}$ (negative value). The region of the band that is concave upward is called an electron band, and the concave downward band region is called a hole band. The regions of the bands that correspond to large negative effective masses are called heavy hole bands (for large positive effective masses, heavy electron bands), and those corresponding to small negative effective masses are called light hole bands (small positive effective masses, light electron bands). The masses can then be listed as m_e^* for electrons and m_h^* for holes. The other bands also show both kinds of curvature as well as different curvatures. In the next chapter holes will be defined more carefully, and it will become clearer how the concept of electrons and holes as carriers of current is central to understanding electronic properties of materials and electronic devices.

RELATED READING

D. A. Davies. 1978. *Waves Atoms and Solids*. Longman, London. A very well-written text covering many of the topics in Chapters 9, 10, and 11 with good insights.

R. E. Hummel. 1992. *Electronic Properties of Materials*. Springer-Verlag, New York. This text provides well-written coverage of the material in Chapters 9, 10, and 11 at the appropriate level. The author has used this book as a text for the electronic materials part of the materials science course.

J. P. McKelvey. 1993. *Solid State Physics for Engineering and Materials Science*. Krieger. A higher level text than Hummel, and also well written, readable, but for the topics covered is more complete.

M. A. Omar. 1993. *Elementary Solid State Physics*. Addison Wesley, Reading, MA. A text that covers many of the topics in Chapters 9, 10, and 11 and with many more topics not covered in the present text. A readable text in the subject.

EXERCISES

1. (a) Sketch a modulated wave form $(\cos\alpha \cdot \cos\beta)$ where the frequency of the high-requency component is $100\times$ the low-frequency component.
   (b) Show the group and phase velocities on the sketch.
   (c) Explain how this kind of wave form can be used to describe an electron that has the characteristics of a particle.

2. Calculate the energy for an electron and for photons with wavelengths of 0.1 nm, 1 nm, and 10 nm.

3. Show that the form for the Schrödinger equation derives from the duality principle of deBroglie.

4. For two adjacent regions 1 and 2 where the electron is free and bound, respectively, sketch the propagation of matter waves from the free electron side to the bound side (from region 1 to 2). Discuss what happens to the wave function when the binding potential and the length of sides are separately increased.

5. Starting with the KP formula, show how the KP model gives rise to the parallel energy band picture for materials by identifying the atomistic parameters that determine band widths and separations.

6. (a) Starting from a square 2-D lattice in RESP, construct the first two Brillouin zones.
   (b) Using your construction explain how energy states (in terms of both the energies and number) vary with different directions in the first Brillouin zone. From this result sketch a density of states versus energy curve.

7. Using the Si band diagram (Figure 9.22c), point out hole and electron bands, and for one of each kind of band, trace out and discuss the variation in the effective mass for the particle in **k** space.

8. Using the Si band diagram (Figure 9.22c) answer the following:
   (a) Identify one hole band with large (heavy hole band) and another with relatively smaller (light hole band) m_h^*. Include the basis for your selection.
   (b) Locate the Fermi level.
   (c) How would this diagram change if this material became amorphous.
   (d) Explain why many of the bands have a parabolic shape.

10

ELECTRONIC PROPERTIES OF MATERIALS

10.1 INTRODUCTION

In this chapter we deal with those aspects of electronic structure that affect and underlie the electronic properties of conductors, nonconductors or insulators, and semiconductors. We start with the electronic energy band structure of Chapter 9 that defines the allowed electronic states. Because band structure alone is not sufficient to define electronic properties, we need to look at how the allowed states are filled. This was briefly mentioned in Chapter 9 when we noted that the disposition of the last bands filled determines to a large extent the resulting properties. Thus the occupation of the allowed states is where we start our focus in the chapter. We will calculate the number of allowed states, the so-called density of states (DOS) function. In addition we will assess the probability that a given allowed state is filled. Since, for electrons and especially electrons with low energies, the Boltzmann probability function is not appropriate, a new distribution function will be used. This function is called the Fermi-Dirac distribution function. This function is readily rationalized, and it leads to a definition of the Fermi energy level that essentially is the electron energy at the probability, $P = \frac{1}{2}$ for the state to be occupied. At absolute zero temperature all states below the Fermi energy, E_F, are filled and all states above E_F are empty.

Once the basics mentioned above are established, then electronic conduction can be considered. This we do from both a classical perspective, so as to obtain the intuitive fundamental relationships, and a more correct quantum mechanical perspective, where we show that only electrons near E_F participate. In addition we will consider the important topic of superconductivity in order to gain some basic ideas about this modern subject. Free electron conduction is the simplest to understand and thus provides the starting point for electronic properties. This discussion is followed by a look at semiconductor properties and the main ideas that underlie most modern electronic

devices. Devices and other modern areas of electronic materials science are reserved for Chapter 11.

10.2   OCCUPATION OF ELECTRONIC STATES

The occupancy of allowed electronic states $N(E)$ depends on the number of allowed states, the so-called density of states (DOS), multiplied by the probability that a state will be occupied, the so-called Fermi-Dirac distribution function $F(E)$. All this is multiplied by 2, since each quantum state can have two electrons of opposite spin. We now proceed to develop the appropriate relationships for DOS, $F(E)$, and $N(E)$.

10.2.1   Density of States Function, DOS

Recall from Chapter 9 the energy formula for the bound electron problem:

$$E = \frac{\hbar^2 \pi^2}{2m_e l^2} \cdot n^2 \qquad (9.73)$$

This formula teaches that the electron energy in a well of length l is quantized in units of n (or n^2) where n is any positive integer. This set of integers defines the allowed states. In three dimensions n is given as

$$n^2 = n_x^2 + n_y^2 + n_z^2 \qquad (10.1)$$

To better visualize the development, it is helpful to define a spherical state space that contains all n in the orthogonal coordinate system defined by the states n_x, n_y, and n_z. This state space is illustrated in Figure 10.1, which shows that any integer n can be obtained by a vector from the origin to the integer n. With n as a positive integer, we restrict our interest to the first octant of this spherical space where all n are positive integers.
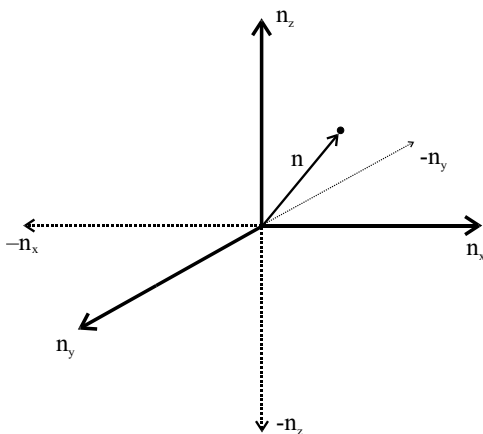
Figure 10.1   State space defined in 3-D by integers n_x, n_y, and n_z.

The objective is to calculate the number of states for an energy E that is less than some maximum energy in the system, E_n. We start with the number of states, η, in the first octant of positive n's. η is given by the volume of a sphere with radius n, but only in the first octant as

$$\eta = \frac{1}{8} \cdot \left(\frac{4}{3} \pi n^3 \right) \tag{10.2}$$

Using equation (9.73), we relate E to n and solve for n to obtain the following expression:

$$n = \left\{ \frac{2m_e l^2}{\hbar^2 \pi^2} E_n \right\}^{1/2} \tag{10.3}$$

This expression for n is substituted into equation (10.2) for η to obtain the result

$$\eta = \frac{\pi}{6} \left\{ \frac{2m_e l^2}{\hbar^2 \pi^2} \right\}^{3/2} E^{3/2} \tag{10.4}$$

The DOS, which is the number of states per unit energy, is obtained as the derivative of η with respect to energy, abbreviated as $Z(E)$, and given as

$$Z(E) = \frac{d\eta}{dE} = \frac{l^3}{4\pi^2} \left\{ \frac{2m_e}{\hbar^2} \right\}^{3/2} E^{1/2} = \frac{V}{4\pi^2} \left\{ \frac{2m_e}{\hbar^2} \right\}^{3/2} E^{1/2} \tag{10.5}$$

where $V = l^3$. The DOS function is plotted in Figure 10.2 as the solid line on $Z(E)$ versus E. It is seen that the density of electron states per unit energy increases as $E^{1/2}$. The dashed line at higher E will be considered later. The explanation for the decrease in $Z(E)$ at higher electron energies lies with a fact discussed near the end of Chapter 9, where we learned that there are fewer electron states at higher energies near the Brillouin zone edges (recall Figure 9.23b). Thus the DOS function decreases at higher energies because of the restrictions of the allotted area in each Brillouin zone.
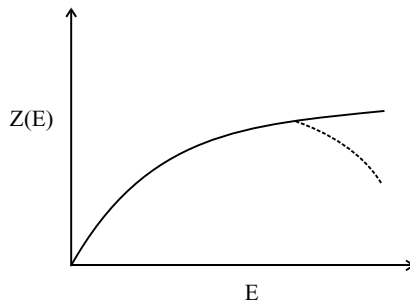
Figure 10.2   Density of states function $Z(E)$ as a function of energy. Solid line for parabolic increase in $Z(E)$, and dashed line indicates that $Z(E)$ decreases for higher energies in the Brillouin zone.

10.2.2 The Fermi-Dirac Distribution Function

The Fermi-Dirac distribution function, $F(E)$, yields the probability for an allowed state to be occupied by an electron. For large non-quantum mechanical particles such as molecules, dust, billiard balls, and trucks, the probability for occupancy of a particular energy state is given by the Boltzmann distribution, $P(E)$, and the probability is an exponential function of the energy as was discussed in Chapter 4 (e.g., see the development of equation 4.34):

$$P(E) = (\text{const})e^{-E/kT} \qquad (10.6)$$

It is important now to notice that there are no restrictions on how many particles can occupy any given energy state, and that states of higher energy are more sparsely populated. It is this probablistic notion that drives various systems to ultimately attain the lowest energy state possible. Also as discussed in Chapter 4, $P(E)$ derives from the assumed proportionality of the change in population, dn, of n states as being proportional to E as

$$\frac{dn}{n} \propto dE \qquad (10.7)$$

For electrons we must consider the quantum mechanical selection rules that do not permit electrons to have identical energy descriptions in terms of quantum numbers. Essentially we need to employ the reasoning that at most two electrons can occupy a quantum state and that these electrons differ only in spin. We present the result below in the form of a new, more appropriate distribution function, the Fermi-Dirac distribution function, $F(E)$, and define it in terms of a "new" energy term the Fermi energy, E_F:

$$F(E) = \frac{1}{1 + e^{(E - E_F)/kT}} \qquad (10.8)$$

One simple way to arrive at this Fermi-Dirac (FD) distribution function is to consider two allowed states, 1 and 2, as shown in Figure 10.3. As was stated above, we need to keep in mind the Pauli exclusion principle that permits two electrons per state, and the principle of detailed balancing that equates forward and reverse kinetically simple processes. A kinetically simple process is a process of a single step as it is written. This means that the written step is not a composite, composed of perhaps many steps. Also it is assumed that the probability of the ith state being occupied is $f_i = f(E_i)$, meaning the probability is a function of the energy. For example, the probability is lower for a higher energy state to be occupied than for a lower energy state.
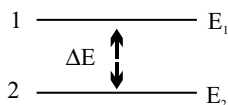
Figure 10.3 Two allowed electron states separated by ΔE.

Consider the following electron (e^-) interactions between the two allowed energy states in Figure 10.3: an e^- in state 1 goes to state 2 and an e^- in state 2 goes to state 1. This is illustrated by the double-ended arrow in the figure. The following energy balance obtains for the transitions:

$$E(1) - E(2) = E(2) - E(1) \tag{10.9}$$

For any transition to occur, it is required that the initial state be occupied and the final state empty, and that electrons do not occupy the same state. The probability for state 1 being occupied is f_1 and for state 2 being empty is $1 - f_2$. The probability for both events to occur at once is obtained from the product of probabilities:

$$P_{12} = f_1 \cdot (1 - f_2) \tag{10.10}$$

For the inverse process, that is, state 1 empty and state 2 filled, we can then write

$$P_{21} = f_2 \cdot (1 - f_1) \tag{10.11}$$

Detailed balance requires that $P = P_{\text{inverse}}$. Thus we obtain

$$f_1 \cdot (1 - f_2) = f_2 \cdot (1 - f_1) \tag{10.12}$$

In terms of ratios we can rewrite this result as

$$\frac{f_1}{1 - f_1} = \frac{f_2}{1 - f_2} \tag{10.13}$$

This equation in quotient form can be put into more useful difference form by taking the logarithms of both sides as follows:

$$G_1 = \ln\frac{f_1}{1 - f_1} \quad \text{and} \quad G_2 = \ln\frac{f_2}{1 - f_2} \tag{10.14}$$

Referring to Figure 10.3, we see a change in energy ΔE between the two states and the probabilities for occupancy of one state or another is a function of the energy difference. This is written $\Delta G \propto -\Delta E$. In using the differential form, we can convert the energy to small changes and obtain $dG = -L\,dE$, where L is the constant of proportionality. Integration will yield the probability G_1. Thus the integral of dG is evaluated from probability 0 to G_1. The energy integral is evaluated from the energy position where the probability corresponds to $G_1 = 0$ to E. $G_1 = 0$ when $f_1 = \frac{1}{2}$, and this energy is defined as the Fermi energy E_F. With these integration limits included, the integral to be evaluated is

$$\int_0^{G_1} dG_1 = -L\int_{E_F}^{E} dE \tag{10.15}$$

The result is

$$G_1 = L(E_F - E) = \ln \frac{f_1}{1 - f_1} \tag{10.16}$$

Rearrangement of this expression in terms of f_1, which is the probability that a state is occupied, and now renamed as $F(E)$ yields the following result:

$$F(E) = \frac{1}{1 + e^{(E - E_F)/kT}} \tag{10.8}$$

where L is the constant of integration and has been experimentally determined to be $L = 1/kT$.

Figure 10.4 displays a plot of the Fermi-Dirac distribution function $F(E)$ versus energy at absolute zero (solid line), and at some temperature above absolute zero (dashed line). The step function appearance of $F(E)$ at absolute zero indicates that the probability is unity that an electron will be at or below the Fermi energy, E_F. This means that all the electrons are found at or below E_F, and that E_F represents the highest energy of any electron in the material. However, at any temperature above 0 K there is a tail extending above E_F. In Figure 10.4 the tail is exaggerated and is typically about $5kT$ where kT is about 0.025 eV at 300 K. With E_F values for metals of about 5 eV, the tail represents only a small part of the electron concentration. Nevertheless, at $T > 0$ K a few electrons have a finite probability to inhabit states that have energies greater than E_F. Because this occurs, there is a finite probability that some states below E_F are unoccupied. These empty states in the valence band are called holes, and electrons in the valence band can move into the holes under an electric field. To distinguish this electron motion from that in the conduction band, it is usual to describe the motion of the holes in the direction opposite to electrons in the valence band, and thus speak of hole motion in describing conduction in the valence band.

It is useful to also consider the mathematical form of $F(E)$ in equation (10.8) above, and then probe various T and E regions. From equation (10.8) for $F(E)$, for $E < E_F$ as T approaches 0 K, the exponential term in the denominator approaches 0, so that $F(E)$ approaches 1. This is consistent with $F(E)$ at $T = 0$ K as is shown as the solid line in Figure 10.4. Likewise, for $E > E_F$ and T approaching 0 K, the exponential in the denominator grows large so that $F(E)$ goes to 0 as is also depicted in Figure 10.4 above E_F and
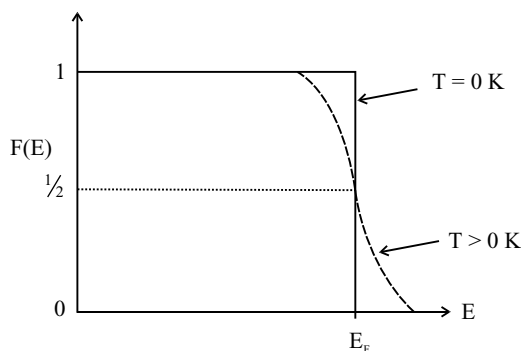
Figure 10.4 Fermi-Dirac distribution function, $F(E)$ at 0 K (*solid line*) and higher T (*dashed line*). The Fermi level E_F at $F(E) = \frac{1}{2}$.

for the solid line ($T = 0\,\mathrm{K}$). For large E, the exponential term in the denominator dominates, and $F(E) \approx e^{-E/kT}$, which is the Boltzmann distribution. So at large electron energies the Boltzmann distribution can be used to describe electron probabilities.

10.2.3   Occupancy of Electronic States

With the density of states $Z(E)$ and the Fermi-Dirac distribution function $F(E)$ obtained, we can calculate the number of occupied electron states $N(E)$ in a material:

$$N(E) = 2 \cdot Z(E) \cdot F(E) \tag{10.17}$$

Figure 10.5a displays the shape of $N(E)$ from the product above for parabolic $Z(E)$ as shown in the solid line of Figure 10.2. However, as was discussed in Chapter 9 (see Figure 9.23), there are fewer states near the Brillouin zone edges. Thus $Z(E)$ is not increasing in a parabolic manner with energy. Rather, as E increases, $Z(E)$ falls rapidly toward the zone boundary, and this is illustrated by the dashed line for $Z(E)$ in Figure 10.2. With this correct fall-off in $Z(E)$ in the product, the more correct shape for $N(E)$ is that appearing in Figure 10.5b, where a maximum is evident near the mid band region. This fact will be important later when electronic conduction is discussed.

The change in the number of occupied states per energy interval dN_o is given as $N(E)\,dE$. The integration of $N(E)dE$ from 0 energy to the Fermi level E_F at $T = 0\,\mathrm{K}$ ($F(E) = 1$) will then yield the number of electrons in allowed states as
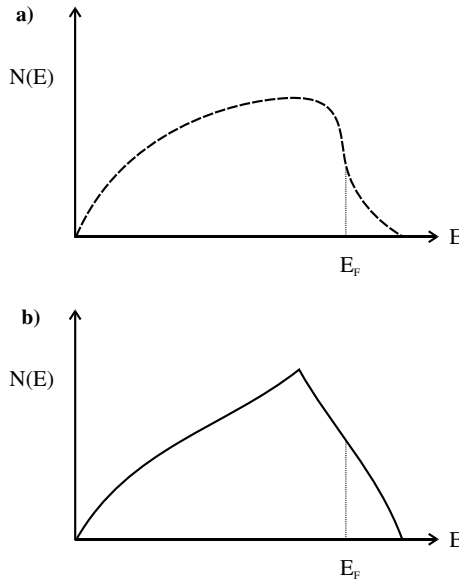
Figure 10.5   Number of occupied electron states $N(E) = 2Z(E)F(E)$. (a) Parabolic $Z(E)$; (b) falloff in $Z(E)$ at higher energies.

$$N_o = \int_0^{E_F} N(E)dE = \int_0^{E_F} 2Z(E)F(E)dE = \int_0^{E_F} 2Z(E)dE \qquad (10.18)$$

Then, with $F(E) = 1$, we substitute for $Z(E)$ using equation (10.5) to obtain:

$$Z(E) = \frac{V}{4\pi^2}\left\{\frac{2m_e}{\hbar^2}\right\}^{3/2} E^{1/2} \qquad (10.5)$$

Integrating over E yields the following:

$$N_o = \frac{V}{3\pi^2}\left\{\frac{2m_e}{\hbar^2}\right\}^{3/2} E_F^{3/2} \qquad (10.19)$$

If we let N_V be the number of electrons per volume, N_o/V, and rearrange solving for E_F, the following is obtained:

$$E_F = \left(3\pi^2 N_V\right)^{2/3} \frac{\hbar^2}{2m_e} \qquad (10.20)$$

This interesting result teaches that the Fermi level is a function of the number of electrons and the volume of the unit cell that depends on the crystal structure.

10.3   POSITION OF THE FERMI ENERGY

The Fermi energy E_F, also referred to as the Fermi level, is the highest energy occupied state at $T = 0\,K$ for an electron, and very close to that for $T > 0\,K$, which is within about $0.1\,eV$ at room temperature. As a practical matter E_F represents the highest energy electrons in an equilibrium solid notwithstanding the Fermi tail. Therefore any excitations of electrons imposed, for example, by an external electric field, or even optical excitation, will potentially lift the equilibrium energy levels to higher energies. Clearly, knowledge of the starting level, E_F, is relevant to understanding the excitation. As will be seen below, knowledge of the position of E_F is in fact crucial to understand virtually all electronic properties and devices. Therefore we now address the problem of the location of E_F in a material. We commence with a semantic argument based on what we already know about the energy band structure, and this argument will enable a very close estimation of the position of E_F that will suffice for most problems. Then a more precise analytic formulation of the problem is presented where the position of E_F is calculated.

We return to Figure 9.15. Recall that it displays the parallel band picture for three cases: Figure 9.15*a* is for a full valence band with a wide gap and an empty conduction band, Figure 9.15*b* shows a full valence band with a narrow gap and an empty conduction band, and Figure 9.15*c* shows a partially full valence band. In each case, with some additional logic, we can label the position of the Fermi level from the definitions above. We commence with the case of a partially filled valence band, and modify it slightly with the addition of the Fermi level E_F as shown in Figure 10.6*a*. In Figure 10.6*a* the most energetic electrons for a partially filled valence band are those at the top of the filled levels in the valence band. A dotted line is drawn, and this energy level is labeled as E_F
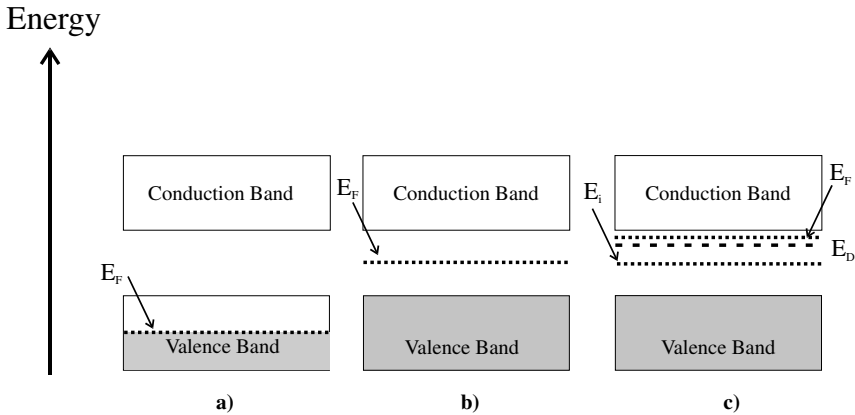
Energy

Figure 10.6 Parallel electron energy bands for (*a*) a metal, (*b*) a semiconductor (or insulator), and (*c*) a doped semiconductor.

the highest level for filled states. Of course, we keep in mind the caveat given above, that this is strictly true only at $T = 0\,\text{K}$ but close to true for room temperature.

Figure 10.6*b* displays a full valence band, a gap, and then an empty conduction band. Depending on the size of the gap, this case can represent either a semiconductor (narrow gap) or an insulator (wide gap). To rationalize the position for E_F in this situation, we recall the probabilistic definition for the Fermi energy as the probability of $\frac{1}{2}$ for an electron to go to the higher allowed level. The probability will be $\frac{1}{2}$ half way in energy between the filled valence band and the empty conduction band. Thus the dotted line is in the middle of the gap. This is an interesting case: while the Fermi level is near the middle of the gap, there are no allowed states at that position.

Figure 10.6*c* is the same case as Figure 10.6*b* except that there is an added level of states represented by a dashed lines labeled E_D near the conduction band. This level is typically called a doping level. It results from adding impurity atoms that yield a nearly monoenergetic level of states. If the impurity states added are filled with electrons, the level is called a donor level and labeled E_D, as shown in Figure 10.6*c*. If the states are empty, they are called acceptor states because they can accommodate or accept electrons from the material; these states are labeled E_A and not shown. Doping will be discussed in more detail later in this chapter, but for now we need only admit to its possibility, and try to locate E_F. In Figure 10.6*c* the donor level is very close to the conduction band. In fact the distance in energy is of the order of 1 or $2kT$ (0.025–0.050 eV). Thus the energy distance to the conduction band is about kT. This means that at room temperature the filled donor states can ionize the electrons to the conduction band (recall the earlier two-state problem and equation (10.6) in which the number in the upper state is nearly $e^{-E/kT}$, where E is the small energy distance between E_D and E_{CB}). In this case then the probability $\frac{1}{2}$ point, E_F, is in between the E_D level and the E_{CB}. This is indicated by the dotted E_F in the figure. In addition near mid gap is another dotted line labeled E_i. This is called the intrinsic Fermi level, and it indicates where E_F will be if the material were not doped, that is, if the material exhibits intrinsic properties rather than properties due to doping (extrinsic properties). More will be said about this later.

Thus, by the definition of E_F, its position on the energy band diagram can be deduced. In reality this exercise is only approximately correct, but it is close enough for most discussions about electronic behavior. It is also worthwhile to perform a more accurate calculation to uncover the assumptions implicitly made above, and thereby deepen understanding of this energy level. What we will find below is that if the valence and conduction bands do not have the same shape or symmetry, then a correction is required to the symmetry argument made above. Recall that previously our discussion of shape or curvature of electron energy bands in Chapter 9 revolved around the concept of effective mass. Effective mass will arise again in our discussion here and in the same context.

To calculate the position of E_F in an energy gap, it is necessary to first realize that as an electron moves from a filled valence band to an empty conduction band, a hole is left behind in the valence band. We calculate the number of electrons in the conduction band n and then equate n to the calculated number of holes in the valence band p. The equation is then solved for E_F, which appears in the equation. For electrons in the conduction band, n is obtained from the equation for $N(E)$, equation (10.18), as

$$n = N(E) = e \int_{E_{C1}}^{E_{C2}} Z(E)F(E)dE \tag{10.21}$$

E_{C1} and E_{C2} are the bottom and top of the conduction band, respectively. This expression can be simplified by realizing that at $T > 0$, $E > E_F$ in the conduction band then $E - E_F \gg kT$. Thus in the conduction band $F(E)$ can be approximated by the Boltzmann distribution $P(E)$:

$$F(E) = \frac{1}{1 + e^{(E-E_F)/kT}} \quad \text{goes to} \quad e^{-(E-E_F)/kT} \tag{10.22}$$

For the case where $l = 1$, then $V = l^3 = 1$. Substituting for V in equation (10.5) obtains the following form for $Z(E)$:

$$Z(E) = \frac{\pi}{4} \left[\frac{8m_e}{h^2} \right]^{3/2} E^{1/2} \tag{10.23}$$

The assumption about $V = 1$ will drop out below when the result is equated with that for holes in the same volume. The integration limits are set for both convenience and consistency. For electrons, we set the energy at the top of the valence band to 0. Then the lower integration limit for the conduction band becomes $0 + E_g$ or E_g, and the upper is set conveniently at ∞. Of course, the conduction band does not extend to infinity, but the exponentially decreasing functions will take care of this. The infinity limit simplifies the math, as will be seen below. With E_g as the bottom of the conduction band, any energy E in the conduction band is now changed to $E - E_g$ for consistency. With equation (10.22) for $F(E)$ and equation (10.23) for $Z(E)$, the integral for n or $N(E)$ in equation (10.21) can be reformulated as

$$n = N(E) = \int_{E_g}^{\infty} \frac{\pi}{2} \left[\frac{8m_e}{h^2} \right]^{3/2} (E - E_g)^{1/2} e^{-(E-E_F)/kT} dE \tag{10.24}$$

To perform the integration, we note first that

$$-\frac{E-E_\text{F}}{kT} = -\left[\frac{E-E_\text{g}}{kT} + \frac{E_\text{g}-E_\text{F}}{kT}\right] \tag{10.25}$$

If we let $x = (E - E_\text{g})/kT$, then the first term above becomes $-x$ and $dE = kTdx$. This yields, for the integral,

$$n = N(E) = \int_{E_\text{g}}^{\infty} \frac{\pi}{2}\left[\frac{8m_\text{e}}{h^2}\right]^{3/2} kT^{3/2}\, x^{1/2}e^{-x}dx \tag{10.26}$$

The integral, now in a standard definite integral form, can be readily solved using the following:

$$\int_0^{\infty} x^{1/2}e^{-x}dx = \frac{\pi^{1/2}}{2} \tag{10.27}$$

The result for electrons in the conduction band is

$$n = N(E) = \frac{1}{4}\left[\frac{8\pi m_\text{e}kT}{h^2}\right]^{3/2} e^{(E_\text{F}-E_\text{g})/kT} \tag{10.28}$$

We can proceed to calculate the number of holes p left in the valence band. From the Fermi-Dirac distribution function we know the probability for electrons. To find the holes (using the subscript "h" to denote holes), or the absence of electrons in the valence band, we use $F_\text{h}(E) = 1 - F(E)$ and obtain:

$$F_\text{h}(E) = 1 - \frac{1}{1+e^{(E-E_\text{F})/kT}} = \frac{e^{(E-E_\text{F})/kT}}{1+e^{(E-E_\text{F})/kT}} \tag{10.29}$$

For $E < E_\text{F}$, equation (10.29) reduces to a Boltzmann-like distribution:

$$F_\text{h}(E) = e^{\,E-E_\text{F}/kT} \tag{10.30}$$

Remember that since $E = 0$ at top of VB, E is $-E$ below that point. Then

$$p = N_\text{h}(E) = 2\int_{-\infty}^{0} F_\text{h}(E)Z_\text{h}(E)dE = -2\int_0^{-\infty} F_\text{h}(E)Z_\text{h}(E)dE \tag{10.31}$$

This yields

$$p = N_\text{h(E)} = -2\int_0^{\infty} \frac{\pi}{4}\left[\frac{8m_\text{h}}{h^2}\right]^{3/2}(-E)^{1/2}\, e^{\frac{E}{kT}} e^{\frac{E_\text{F}}{kT}} dE \tag{10.32}$$

With $x = -E/kT$ and $kTdx = -dE$, we obtain

$$p = N_\text{h}(E) = \int_0^{\infty} \frac{\pi}{2}\left[\frac{8m_\text{h}}{h^2}\right]^{3/2} kT^{3/2}(x)^{1/2}\, e^{x}e^{E_\text{F}/kT} dx \tag{10.33}$$

Finally p is obtained:

$$p = N_h(E) = \frac{1}{4}\left(\frac{8m_h\pi}{h^2}\right)^{3/2} kT^{3/2} e^{E_F/kT} \tag{10.34}$$

We equate $n = p$, using equations (10.28) and (10.34) to obtain

$$e^{(E_F - E_g)/kT} = \left(\frac{m_h}{m_e}\right)^{3/2} e^{E_F/kT} \tag{10.35}$$

Note that the expression for n has the mass of the electron m_e while expressions for p uses the hole mass m_h. Equation (10.35) is solved for E_F by taking the logarithm of both sides:

$$E_F = \frac{E_g}{2} + \frac{3}{4}kT \ln\left(\frac{m_h}{m_e}\right) \tag{10.36}$$

The intuitive result above that $E_F = E_g/2$ is observed when the effective masses for holes and electrons are approximately equal, since $kT \ll E_g$ ($0.025 \ll 1$). In contrast, the effective mass difference for electrons and holes can be substantially different. For a 10× difference, which is larger than usual, the difference from the center of the gap is less than 0.05 eV. Thus for most purposes the center of the gap is sufficient. Recall from equation (9.127) that the effective mass is related to the band curvature. When the curvature or shape of the band in which the electron and hole reside is the same, then the masses are equal and the second term in equation (10.36) drops out.

Before proceeding to the electronic properties of materials, particularly semiconductors, we will summarize our discussion of holes. As we have seen, a hole is formed in a material with a filled valence band and a gap when an electron in the valence band is removed to the conduction band. In a metal at $T > 0$ K, for every electron above E_F a hole is left below E_F. Thus a hole is an occurrence in the valence band of materials. Once a hole is created in the valence band of a semiconductor, electrical conduction can take place not only in the conduction band by the motion of electrons through the empty allowed states in the conduction band, but also via electrons moving in the now vacated states in the valence band, the holes. So it is always electron motion, but in the valence band it is called hole motion (opposite in direction to electron motion) to distinguish it from electron motion in the conduction band. Also, because the motion in the different bands is via electrons with different energies and in different environments, the ease of electron motion can be quite different, and holes usually have slower motion. Basically hole motion is similar to vacancy motion, which we encountered previously as Nabarro-Herring creep. The idea of holes and vacancies moving provides a convenient and descriptive terminology.

10.4   ELECTRONIC PROPERTIES OF METALS: CONDUCTION AND SUPERCONDUCTIVITY

10.4.1   Free Electron Theory for Electrical Conduction

The classical theory of electrical conduction in metals is called the Drude theory. It is based on the idea that many electrons in metals are nearly free and therefore can migrate

easily with a modest applied potential V. The motion of electrons per unit time is called electron current I, and the flux of electrons $\mathbf{J}_e$ is the current per area, I/A. The electric field $\mathbf{E}$ is given as the potential per distance L, $E = V/L$. We recall the flux equations (5.1), from which Ohm's law is for the electron flux being proportional to the applied electric field as

$$\mathbf{J}_e \propto \mathbf{E} \tag{10.37}$$

The constant of proportionality is the conductivity σ as

$$\mathbf{J}_e = \sigma\mathbf{E} \tag{10.37}$$

The electric field can be expressed as the gradient in potential, as was done in Chapter 5. The conductivity σ is the reciprocal of the resistivity ρ as

$$\sigma = \frac{1}{\rho} \tag{10.38}$$

Ohm's law is often expressed as

$$I = \frac{V}{R} \tag{10.39}$$

Where R is the resistance and is expressed in terms of ρ as

$$R = \frac{\rho L}{A} \tag{10.40}$$

Where L and A are the length and cross-sectional area of the conductor. Recall that the unit for flux, and in particular, electron flux, is number of electrons/$A \cdot t$. Thus $\mathbf{J}_e$ can also be expressed as

$$\mathbf{J}_e = \sigma\mathbf{E} = N \cdot \mathbf{v}_d \cdot e \tag{10.41}$$

where N is the number of charges/volume, $\mathbf{v}_d$ is the drift velocity of electrons, and e is the unit electronic charge. Then solving equation (10.41) for the conductivity, we have the following equation for σ:

$$\sigma = \frac{N_e \mathbf{v}_d e}{\mathbf{E}} = N_e \mathbf{\mu}_e e \tag{10.42}$$

where μ_e is the electron drift velocity per electric field, called the electron mobility, and is an important device quantity that determines operation speed of many electronic devices.

In the free electron theory, the free electrons are considered to be in random motion and thus providing no net current. However, if an electric field is imposed, the resulting electromotive force eE causes a concerted motion of the electrons. The electromotive force can be written as equal to the classical Newtonian force on the electrons, $\mathbf{F} = m\mathbf{a}$ as

$$m\mathbf{a} = m\frac{d\mathbf{v}}{dt} = -e\mathbf{E} \tag{10.43}$$

The negative sign indicates that the direction of the electron motion and the electric field are opposite. This equation can be integrated to yield

$$m\mathbf{v} = e\mathbf{E}t \tag{10.44}$$

Equation (10.44) implies that for a constant applied field $\mathbf{E}$, the velocity of an electron and the electron flux increases indefinitely with time. Rather, what is observed is that for a given material a constant current or electron flux is obtained with the application of a constant field. Therefore equation (10.44) embodies a nonphysical result, and the model is obviously in need of correction. We can correct the differential equation above by imposing a viscous force on the electrons with the form $m\mathbf{v}/\tau$, where τ is the characteristic time for the decay of the electron velocity when the electric field is removed; τ is called a relaxation time. The viscous force is derived from the fact that as electrons move through the solid, they collide with nuclei, and these collisions tend to reduce the velocity in the forward direction. τ has units of time and is the time in between collisions. If the time is short, then a larger correction term is obtained. This correction to equation (10.44) yields

$$m\frac{d\mathbf{v}}{dt} = -e\mathbf{E} - \frac{m\mathbf{v}}{\tau} \tag{10.45}$$

Now suppose that an electric field $\mathbf{E}$ is applied to establish some electron velocity and then $\mathbf{E}$ is turned off ($\mathbf{E} = 0$). Imposing these conditions on equation (10.45), we obtain the differential equation for the relaxation of the velocity as follows:

$$\frac{d\mathbf{v}}{\mathbf{v}} = -\frac{dt}{\tau} \tag{10.46}$$

This integrates to the following expression for velocity:

$$\int_{v_0}^{v_d} \frac{d\mathbf{v}}{\mathbf{v}} = -\int_0^t \frac{dt}{\tau} \tag{10.47}$$

The result is

$$\mathbf{v}_d = \mathbf{v}_0 e^{-t/\tau} \tag{10.48}$$

Equation (10.48) indicates that when the field is removed, the velocity exponentially decays over time with a shape determined by the relaxation time τ, as is shown in Figure 10.7a. Also the viscous drag-corrected differential equation (10.45) can be explored when the velocity reaches a final velocity $\mathbf{v}_f$ in the applied field. This velocity increases when the field is applied, but it cannot increase indefinitely (as in equation 10.44, the uncorrected form). So the velocity levels off at some final velocity $\mathbf{v}_f$. The end point is evident when the velocity no longer changes, or at $d\mathbf{v}/dt = 0$:
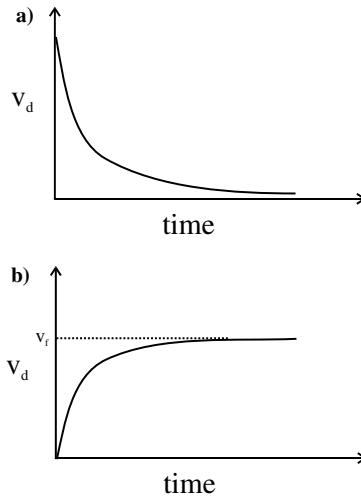
Figure 10.7 Electron drift velocity $\mathbf{v}_d$ versus time (*a*) decreases exponentially when the electric field is removed and (*b*) reaches a final velocity $\mathbf{v}_f$ under constant field.

$$\frac{m\mathbf{v}_f}{\tau} = e\mathbf{E} \quad \text{and} \quad \mathbf{v}_f = \frac{e\mathbf{E}}{m}\tau \tag{10.49}$$

The change in velocity resulting from the application of the applied field until the velocity no longer changes is illustrated in Figure 10.7*b*. For consistency the units for $\mathbf{v}_f$ can be checked: e has units of amps·time, $\mathbf{E}$ has units of (mass·length²)/(amps· time³·length), and τ has units of time. The result is that $\mathbf{v}_f$ correctly has units of distance/time.

Now we can pull all this together and obtain the classical expression for the conductivity σ, so that we can compare it later with the quantum mechanical result. From equation (10.41) for $\mathbf{J}_e$ above we obtain the starting formula for σ as

$$\sigma\mathbf{E} = N \cdot \mathbf{v}_d \cdot e \tag{10.50}$$

Then we substitute for $\mathbf{v}_f$ from equation (10.49) and solve for σ:

$$\sigma = \frac{Ne^2}{m}\tau \tag{10.51}$$

Thus the classical result illustrates the dependence of the conductivity on the number of electrons N and the relaxation time in between collisions, τ.

The free electron theory provides good correspondence with the observations for metals. However, this theory provides no understanding about insulator and semiconductor materials that also have large numbers of valence electrons. Thus a more complete theory is needed that explains all kinds of materials' electronic behavior, and the quantum theory does this very well.

10.4.2   Quantum Theory of Electronic Conduction

The main thrust of the quantum theory is that unlike the free electron theory, all the electrons in a material are not equal with respect to the conduction process. Recall from electron band theory in Chapter 9 that allowed energy states are based on the chemical bonding that determines the strength of the binding potentials and the ordering (short and long range) of the material. These allowed quantum states are filled with the available electrons for the particular material. When all the available electrons are in place, the highest filled level for metals closely determines the position of the Fermi level, E_F. For materials with a completely filled valence band E_F is near mid gap. The electrons are undergoing random motions with random velocities in the absence of an electric field. These random velocities give rise to random momenta as well. Therefore momentum or **k** space can be useful in interpreting the sequence of events that occur when an electric field is applied to the electrons that originally have random momenta. Figure 10.8a displays 2-D **k** space for a material with a Fermi level that is the same in all directions. Recall from Chapter 9 that the electron energy is proportional to **k** and for free electrons is given by the formula

$$E = \frac{\hbar^2}{2m_e}\mathbf{k}^2 \tag{9.67}$$
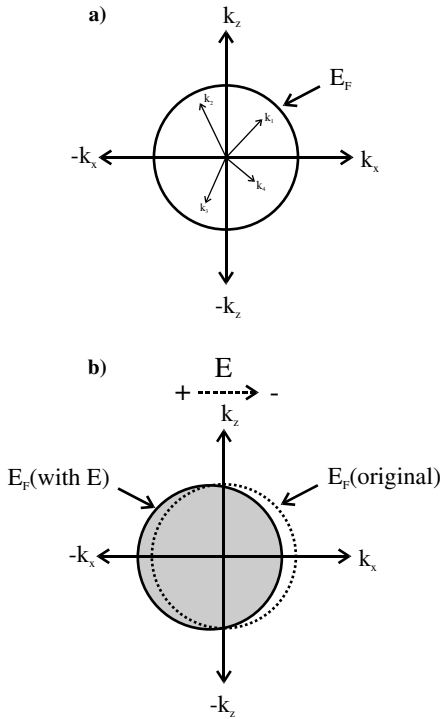
Figure 10.8   **k** space in 2-D for a cubic material showing the Fermi energy E_F for (a) without an applied field and (b) with an applied field **E**.

In Figure 10.8a any electron energy where $E \leq E_F$ can be identified by a vector to the appropriate **k** value within the circle defined by E_F. When an electric field **E** is imposed in the direction shown by the arrow in Figure 10.8b, the total randomness of the electron velocities is lost. Some electrons traveling opposite to the direction of the field **E** shown in Figure 10.8b receive maximum increase in energy from the electric field, while those electrons traveling in the opposite direction are reduced in energy by the field. This is indicated by the formation of a new Fermi circle shown as the shaded circle in Figure 10.8b. Those electrons in the shaded crescent on the left of Figure 10.8b receive energy from the electric field, while those electrons that were in the unshaded crescent on the right are reduced in energy by virtue of the applied electric field. It is the electrons in the shaded crescent that contribute to electronic conduction; the electrons in the overlap region of the two Fermi circles do not contribute to electronic conduction. This is understood by considering that all the electrons with **k** values in the overlap region have a corresponding electron with the opposite **k**. The momenta for all the electrons in any direction in the overlap region compensate, and the effects cancel. However, electrons with energies and momenta in the shaded crescent to the left do not have compensating momenta, and consequently there is a net electron momentum to the left for Figure 10.8b for the given electric field direction.

Thus, from a quantum mechanical point of view, not all electrons participate in electronic conduction. Rather, it is only those electrons that have energies near E_F that participate. This is the first deviation from the free electron theory. The momentum diagrams shown in Figure 10.8a and 10.8b can readily be converted to velocities for the electrons using equation (9.44). The result is shown in Figure 10.9a. Thus the velocities corresponding to the Fermi velocity are the uncompensated velocities that are included in electronic conduction, driven by the applied electric field. Figure 10.9b shows the parabolic shape for the number of occupied states $N(E)$ that obtain for purely free electrons, as was discussed earlier. In addition this figure shows E_F and that with an applied electric field there is a range of energies (and velocities) near $E_F(v_F)$ that participate in electronic conduction. The shaded bar in the figure shows the large density of occupied electronic states in the same region as the participating electron velocities.

We can put these ideas together, starting with equation (10.41) for the electron flux:

$$\mathbf{J}_e = N \cdot \mathbf{v} \cdot e \tag{10.52}$$

where classically **v** was the drift velocity $\mathbf{v}_d$. Now we substitute $\mathbf{v}_F$ for **v** and for N we substitute $N(E_F)dE$, since we are now concerned with $N(E)$ near E_F, to obtain

$$\mathbf{J}_e = \mathbf{v}_F e N(E_F) dE \tag{10.53}$$

Since by equation (9.67), E is related to **k**, the expression for $\mathbf{J}_e$ can be written as

$$\mathbf{J}_e = \mathbf{v}_F e N(E_F) \frac{dE}{d\mathbf{k}} d\mathbf{k} \tag{10.54}$$

Then, using equation (9.67) for $E(k)$, we can write the following:

$$\frac{dE}{d\mathbf{k}} = \frac{\hbar^2}{m_e} \mathbf{k} \tag{10.55}$$
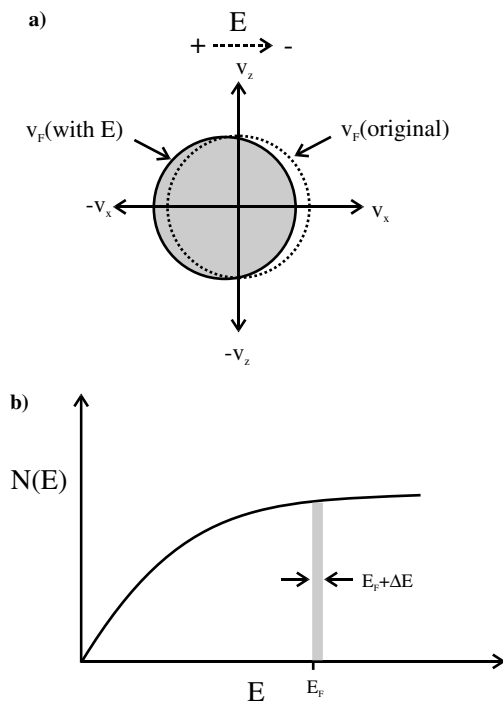
Figure 10.9 (*a*) 2-D velocity space for a cubic material with an applied electric field *E*; (*b*) the density of occupied electron states *N*(*E*). The states affected by the electric field are shaded.

Using this expression, and recalling that $\mathbf{p} = \hbar\mathbf{k}$ (equation 9.44), we can write

$$\frac{dE}{d\mathbf{k}} = \frac{\hbar^2}{m_e}\frac{\mathbf{p}}{\hbar} = \frac{\hbar^2}{m_e}\frac{m_e\mathbf{v}_F}{\hbar} = \hbar\mathbf{v}_F \tag{10.56}$$

Then we substitute this result in the expression for $\mathbf{J}_e$ above to obtain

$$\mathbf{J}_e = \hbar\mathbf{v}_F^2\, eN(E_F)\, d\mathbf{k} \tag{10.57}$$

Now an expression for $d\mathbf{k}$ is needed we can find it as follows: Starting from $\mathbf{F} = d(m\mathbf{v})/dt$, we apply the fact that the force $\mathbf{F} = e\mathbf{E}$:

$$\mathbf{F} = \frac{d(m\mathbf{v})}{dt} = \frac{d\mathbf{p}}{dt} = \hbar\frac{d\mathbf{k}}{dt} = e\mathbf{E} \tag{10.58}$$

Then from equation (10.58), we obtain $d\mathbf{k}$ as

$$d\mathbf{k} = \frac{e\mathbf{E}}{\hbar}\, dt = \frac{e\,\mathbf{E}}{\hbar}\tau \tag{10.59}$$

where as before the time interval τ (substituted for dt) is the time between collisions. Substituting this expression for $d\mathbf{k}$ into equation (10.57) for $\mathbf{J}_e$ yields

$$\mathbf{J}_e = \mathbf{v}_F^2 e^2 N(E_F)\mathbf{E}\tau \tag{10.60}$$

This expression is for the 1-D case in the x direction. In 3-D for a spherical Fermi surface,

$$\mathbf{v}_F^2(x) = \frac{1}{3}\mathbf{v}_F^2 \tag{10.61}$$

So the 3-D expression for $\mathbf{J}$ becomes

$$\mathbf{J}_e = \frac{1}{3}\mathbf{v}_F^2 e^2 N(E_F)\mathbf{E}\tau \tag{10.62}$$

From this we find $\sigma = \mathbf{J}/\mathbf{E}$ and obtain the quantum mechanical σ that can be compared with the classical result:

$$\sigma = \frac{\mathbf{J}_e}{\mathbf{E}} = \frac{1}{3}\mathbf{v}_F^2 e^2 N(E_F)\tau \tag{10.63}$$

This quantum mechanical result in equation (10.63) can be compared with the classical result in equation (10.51). The quantum mechanical $\mathbf{J}_e$ depends on $\mathbf{v}_F$ and not merely on any $\mathbf{v}$, and also on the number of electrons at the Fermi energy level but not all the electrons.

The quantum mechanical treatment has a broader applicability. For example, for metals that are materials with partially filled valence bands, we recall Figure 10.5b, which shows that there are the most electrons near the middle of a band. Thus metals should be the materials with the highest electronic conductivity. For insulators and semiconductors, there is a filled valence band and then a gap. With the Fermi level in the gap, the number of occupied states near the Fermi level is small, and we should anticipate low conductivity.

10.4.3 Superconductivity

As was noted above, normal electrical conductivity is characterized by electrons experiencing a resistance when electronic current flows through a material. However, under some circumstances, typically low temperature, some materials (27 elements, many alloys and compounds) exhibit zero resistance to current flow. As a result of zero resistance the current that flows in a superconductor does not decay in time, and as such is called a super current. It is obvious that superconductors with the lossless currents can revolutionize many areas of electronics, electrical transmission, and magnetics. Consequently there has been great interest in superconductive materials, especially since the discovery of superconductors that exist at relatively high temperatures. These materials will be discussed below, but before that we focus on the basics and consider the phenomenon of zero electrical resistance that is called superconductivity.

Figure 10.10a shows an ideal resistivity versus temperature characteristic for a superconductor. At temperature $T = T_c$ there is a transition from normal behavior where resistivity decreases with decreasing T. The normal behavior at $T > T_c$ is attributed to reduced
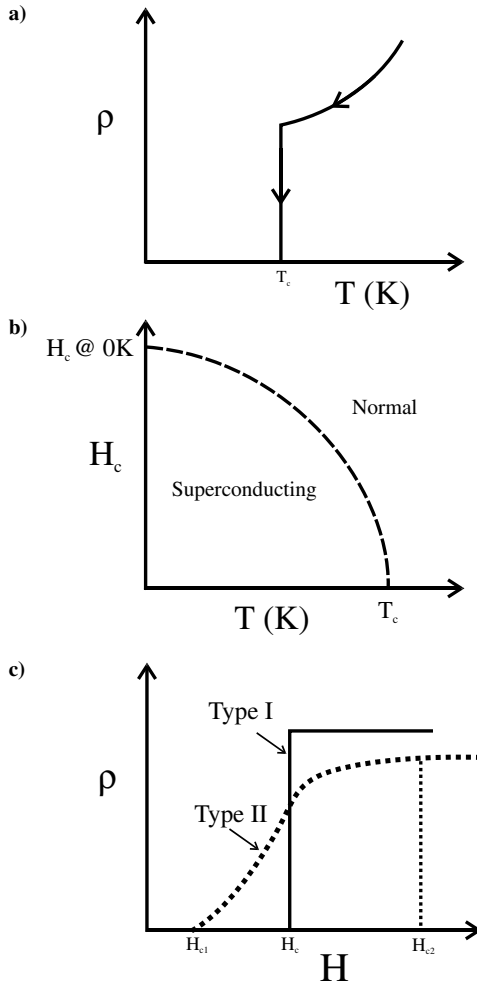
Figure 10.10 (*a*) Resistivity versus *T* for an ideal superconductor; (*b*) critical magnetic field versus *T* for a superconductor; (*c*) resistivity versus magnetic field for type I and II superconductors.

scattering of electrons as the atoms in the lattice vibrate less at lower *T*. At T_c, the critical superconducting transition temperature, the electrical resistance disappears. This figure displays ideal behavior while materials with defects and impurities might display a less sharp transition at T_c and/or a less steep approach to zero resistivity. Table 10.1 shows a listing of some common superconductors with their associated T_c.

It is seen in the table that all the metal and alloy superconductors have low T_c's that will require continual and extensive cooling for the material to operate in the superconducting state. The requirement for cooling to *T* near absolute zero reduces the technological importance of superconductors because such cooling is expensive and cumbersome. However, a remarkable discovery made by Bednorz and Mueller in 1986 has started a renewal of interest. These workers discovered that certain oxide compounds

Table 10.1   Some superconducting materials

Material	T_c (K)
W	0.01
Hg	4.15
Al	1.2
Nb_3Ge	23.1
LaBaCuO	40
$YBa_2Cu_3O_7$	92

(two of many now discovered are listed above) exhibit superconductivity at considerably higher temperatures than metals and alloys. This discovery may lead the way for applications without the expensive cooling apparatus required for the low-temperature superconductor materials. Research with the new high-temperature superconductor materials that are typically complex oxides is presently an active and fertile area of electronic materials research.

In the early days of superconduction research, it was found that the application of a magnetic field could destroy superconduction. For Al the critical field at which superconductivity is destroyed is around 100 Gauss and for Hg around 400 Gauss. For reasons not fully understood, the critical field is lower for the high-temperature oxide superconductors, and hence this is a limitation on these materials at this time. The magnetic field that destroys superconductivity is a function of temperature. The temperature at which the critical magnetic field $\mathbf{H}_c$ exists that negates the super current is plotted in Figure 10.10b. It is seen that at $T = 0\,K$, the material can withstand the highest magnetic field before it loses superconductivity, but at higher temperatures, a lower magnetic field will force the material back into the normal state. A critical magnetic field can come not only from an externally applied field but also from the super current itself.

When a material is brought into the superconducting state the super current (or any flowing current) gives rise to a magnetic field. If the current is sufficiently high, the produced magnetic field may exceed $\mathbf{H}_c$ and superconductivity will be destroyed. Also, as in any conductors, the flowing current induces a magnetic field with a magnetic field intensity $\mathbf{H}$. The magnetic field causes a current that opposes the super current and eventually causes the material to revert from the superconducting state back to the normal state. Figure 10.10c displays critical magnetic field behavior for two types of superconductors, referred to as type I and type II superconductors. Type I superconductors are also called soft superconductors, and they are characterized by a sharp onset of superconduction at $\mathbf{H}_c$ (solid line). Type II superconductors, often referred to as hard superconductors, display a less pronounced onset of superconduction. Type II superconductors are characterized by two values for the critical magnetic field $\mathbf{H}_{c1}$ and $\mathbf{H}_{c2}$. In between there are regions in the material that remain superconducting and regions that are normal. The mixed superconducting and normal region in between $\mathbf{H}_{c1}$ and $\mathbf{H}_{c2}$ is referred to as the vortex state. Type II superconductors are usually the choice for fabricating superconducting magnets, since these materials can withstand higher magnetic fields before reverting to the normal state.

As was shown above using Figure 10.10a, the most interesting characteristic of the superconducting state is that ρ goes to 0 at T_c. However, as a practical matter, it is not easy to know whether ρ is actually at or merely very close to zero resistance—that is, when the meter is really at or very near zero. Thus, in order to definitively determine the

superconducting state, a measurement of ρ alone may be insufficient, or at least by itself untrustworthy. A more definitive assessment can be made using the Meisner effect. As was mentioned above when a superconductor is in the superconducting state, the super current can flow. If at the same time a magnetic field is imposed on the superconductor in the superconducting state, the resulting super current will cause a magnetic field to occur that opposes an applied magnetic field, and expels the magnetic flux **B** due to the applied field **H**. The total expulsion of the magnetic flux that occurs in superconductors is called the Meisner effect. The magnetic flux is given as

$$\mathbf{B} = \mu_0(\mathbf{H} + \mathbf{M}) \tag{10.64}$$

where μ_o is the permeability of free space ($4\pi \times 10^{-7}\,\text{Hz/m}$). **M** is called the magnetization, and it represents the magnetization that occurs within the material. Because of a super current within a superconductor **B** = 0, and thus **H** = −**M**. The ratio of **M/H** is called the magnetic susceptibility χ, and the magnitude χ = −1 for a superconductor is considered perfect diamagnetism. For metals χ ≈ −10^{-5}. The combination of the two measurements above, namely ρ going to zero and the expulsion of magnetic flux, provides convincing evidence that a material is truly a superconductor.

A quantum mechanical theory was proposed to explain superconductivity, in 1957, by Bardeen, Cooper, and Schrieffer, thus called BCS theory. BCS theory explains virtually all the superconductor phenomena related to the low-temperature superconductors, namely the metals and alloys, but there is some dispute about its applicability to all aspects of the high-temperature superconductors, the oxide superconductors. Nevertheless, BCS theory provides important insights into superconductivity, and thus a brief conceptual outline will be given herein.

A key aspect of BCS theory is that the participating electrons, the super electrons, pair up into so-called Cooper pairs. Figure 10.11a shows a 2-D Fermi circle with two electrons *a* and *b* near the Fermi surface. Because of Coulombic forces, electrons repel each other. However, for electrons *a* and *b* the repulsion should be minimum because they are maximally screened from each other by the electrons in between (in the shaded region), and they are the greatest distance apart in **k** space having **k** and −**k** vectors near the Fermi surface. In BCS theory it is found that with the repulsions at a minimum and with spins opposite, two electrons can attract each other with a small attractive potential V. Many pairs can form at or very near E_F. As we have previously seen in Chapter 9, this binding potential can give rise to an energy band structure and a gap, the superconducting gap illustrated in Figure 10.11b. From the details of BCS theory this superconducting gap E_{scg} is given as

$$E_{scg} = 8\hbar\omega_D e^{-2/N(E_F)V} \tag{10.65}$$

where ω_D is the Debye frequency and found to vary with the reciprocal of mass as

$$\omega_D \propto \frac{1}{\sqrt{M}} \tag{10.66}$$

Since E_{scg} is proportional to ω_D, the gap is also proportional to $M^{-1/2}$, and this is known as the isotope effect. For example, if a heavier isotope is substituted for a lighter one, then the gap will decrease and T_c will also decrease. The isotope effect has also been used along with the Meisner effect to confirm superconductivity behavior. E_{scg} measured values
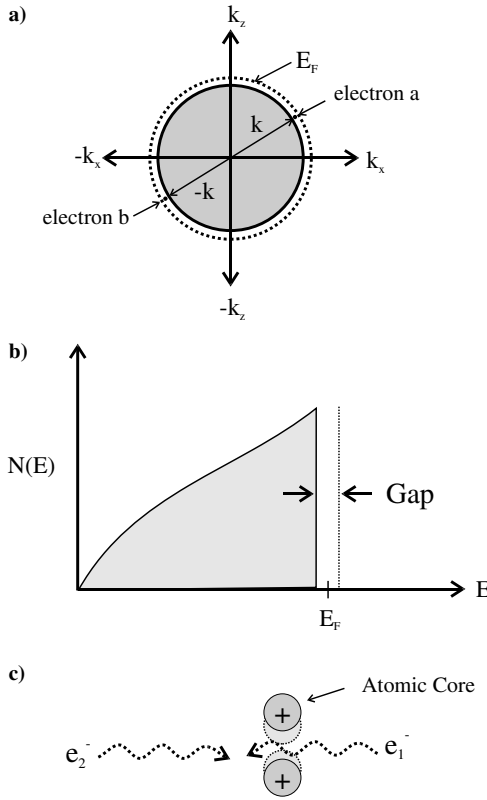
Figure 10.11   (*a*) 2-D **k** space for a superconductor showing electrons *a* and *b* near E_F with opposite **k**; (*b*) density of occupied electron states for a superconductor showing the superconducting gap; (*c*) phonon mechanism for forming Cooper pairs.

are about 10^{-4} eV or in the infrared range ($\approx 1.2\mu$). It is also interesting that as E_{scg} increases because of larger electron binding energies, T_c rises. The oxide high-temperature superconductors all have large potentials and are insulators at room temperature, yet they are excellent high T_c superconductors.

The mechanism for the origin of the electron pairing in the BCS theory is attributed to a phonon effect, namely a quantized atomic vibration. This phonon effect can be somewhat simplistically visualized using Figure 10.11*c*. Consider two atomic core positions indicated by the shaded circles with + charges. An electron (e_1^-) traveling from the right to left has a path between the two cores. The rapidly moving electron e_1^- causes a displacement of positive cores, as is indicated by the lighter shaded cores. Another electron e_2^- traveling in the opposite direction but at an instant later, so as not to be affected by e_1^-, "sees" a different environment with respect to the cores. In fact the slow displacement of the cores cannot relax back to the equilibrium position before the arrival of e_2^-. The greater center of positive charge due to the displacement of the cores causes e_2^- to accelerate. Thus the electrons e_1^- and e_2^- are thought of as bound by the vibration of the cores, the phonon.
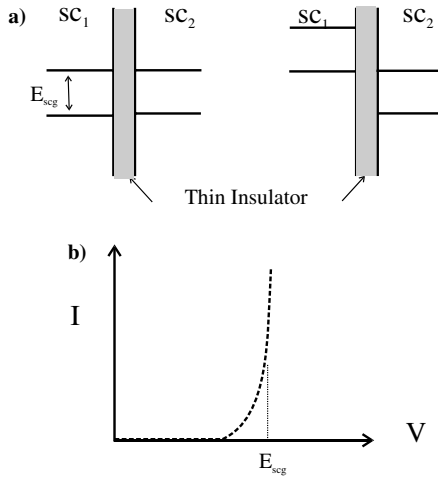
Figure 10.12 Josephson tunneling (*a*) between two identical superconductors and (*b*) when a field is applied tunneling of Cooper pairs can occur yielding (*c*) a sharp current rise.

Among the interesting applications for superconductors is superconducting tunneling, or the so-called Josephson tunneling. Tunneling in superconductors occurs similarly to tunneling in normal materials, and this is illustrated in Figure 10.12*a*. The left panel shows two nearly identical superconductors separated by a thin insulating barrier, typically less than 2 nm thick. If a potential is applied equivalent to or more than the gap E_{scg}, then tunneling will take place from occupied states in one superconductor to empty states in the other as is indicated in the right-hand panel. The current *I* versus voltage *V* characteristic is seen in Figure 10.12*b*. The sharp rise is in the picosecond time range, which makes device switching very fast with superconductors. What is different in superconductors is that for low currents, Cooper pairs migrate across the tunnel barrier, and consequently the insulator becomes a superconductor where current can flow without an accompanying voltage. For larger currents, the Cooper pairs can radiate energy as they drop to lower energy states as follows:

$$2eV_{12} = \hbar\omega \qquad (10.67)$$

where V_{12} is the dc voltage across the junction. Thus an applied dc voltage produces electromagnetic radiation typically at microwave frequencies. This effect is due to the fact that across the barrier the incident wave function receives a phase shift. This phase shift can be expressed in terms of the current flux **J** as follows:

$$\mathbf{J}_2 = \mathbf{J}_1 \sin\phi \qquad (10.68)$$

With a voltage *V* applied the phase shift is increased:

$$\mathbf{J}_2 = \mathbf{J}_1 \sin(\phi + \Delta\phi) \qquad (10.69)$$

The phase shift due to the applied potential $\Delta\phi$ can be expressed as follows:

$$\Delta\phi = \frac{Et}{\hbar} = \frac{eVt}{\hbar} \tag{10.70}$$

where the energy is expressed as eV. For the two electrons of a Cooper pair, this phase is multiplied by 2 and reinserted into the flux equation to yield

$$\mathbf{J}_2 = \mathbf{J}_1 \sin\left(\phi + 2\frac{eVt}{\hbar}\right) \tag{10.71}$$

Thus an applied dc potential yields a time-dependent current. The factor $Et/\hbar$ can be expressed in terms of a frequency as follows:

$$\frac{Et}{\hbar} = \frac{h\nu t}{\hbar} = \omega t = \frac{2eV}{\hbar}t \tag{10.72}$$

The resultant frequency is given by $\nu = 484V$, which is in the GHz range.

10.5 SEMICONDUCTORS

Semiconductors are those materials that have a filled valence band followed in energy by a band gap, and then an empty conduction band, as is shown in Figure 10.13*a*. The band gap is large relative to kT, but not very large and typically less than 2 eV. For example, Si, the most utilized semiconductor in the world, has a band gap of 1.1 eV while Ge has a gap of 0.7 eV and GaAs, an important compound semiconductor, has a gap of 1.4 eV. For comparison, SiO_2, a common insulator used in microelectronics has a band gap of about 9 eV, and diamond, another good insulator, has a gap of about 5.5 eV. Below we will use the Fermi-Dirac distribution function $F(E)$ and the density of states function $Z(E)$ to calculate the number of electrons in the conduction band for semiconductors and compare that number with insulators. For Si, we will find about 10^9 electrons/cm^3 in the conduction band at room temperature, and for insulators with wider gaps, several orders of magnitude fewer electrons are present. However, for typical device operation, such as the devices in a desktop PC, about three orders of magnitude more electronic carriers are needed for device operation than available in the pure Si semiconductor. The pure semiconductor exhibits intrinsic electronic properties dictated by the number of electronic carriers available, and thus this kind of semiconductor is called an intrinsic semiconductor. Intrinsic semiconductors are rarely useful because of the low number of current carriers. Thus the semiconductors require additional substances that can alter the number of carriers. These substances are called dopants, and the process involved is called doping. The result is a semiconductor whose electronic properties are dominated by the dopants, and this kind of semiconductor is called an extrinsic semiconductor. In the following sections we will explore both intrinsic and extrinsic semiconductors.

10.5.1 Intrinsic Semiconductors

A pure semiconductor such as Si, which has a band gap of about 1.1 eV, is shown in Figure 10.13*a*. As was discussed above in Section 10.3, this intrinsic semiconductor as well as other intrinsic semiconductors have E_F close to the center of the gap, and because

a) Energy

Conduction Band

E_g ⟷ .... E_F

Valence Band

b) Energy

CB ← Electrons

E_g ⟷ .... E_F

← Holes

VB

0   1

E → F(E)
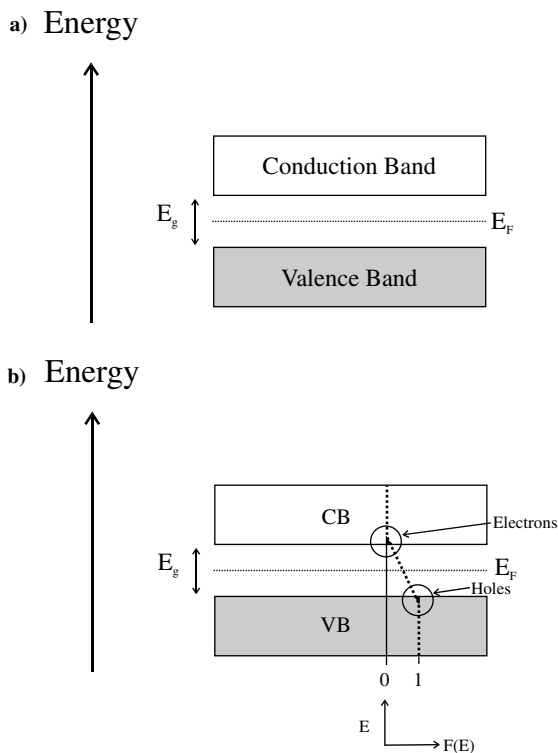
Figure 10.13 (*a*) Parallel band scheme for an intrinsic semiconductor; (*b*) band structure with the Fermi-Dirac function indicating holes in the valence band and electrons in the conduction band.

this is an energy gap, there is no density of allowed states in the gap. Thus the only way to effect electronic conduction is to somehow enable electrons to access the allowed empty states in the conduction band. Among the many ways to do this are to promote valence band (VB) electrons to the conduction band (CB) via a photon process, and/or via a thermal process, and/or by adding allowed electron states. This latter notion of adding states is called doping and will be discussed below for extrinsic semiconductors. The photon process is possible but impractical for device operation, except for devices that are used for detecting photons.

Since the usual electronic devices are operated at room temperature, we will fully characterize intrinsic semiconductors at room temperature, and then explore changes that may occur at other temperatures. At room temperature we recall that the Fermi-Dirac (*FD*) distribution (equation 10.8) yields band tailing to energies above E_F. Figure 10.13*b* shows the band structure as in Figure 10.13*a* but superimposed is the FD distribution function for room temperature. Note that when low in energy in the VB, $FD = 1$. Likewise, when high in energy in the CB, $FD = 0$. These regions of the VB and CB are filled and empty, respectively. However, in the region of the upper band edge for the VB and lower band edge for the CB, the situation is different from the 0 K picture and different from regions in the bulk of the bands described above. Specifically, the very top of the VB has empty states, holes, and the bottom of the CB has occupied electronic states.

These regions are shaded in Figure 10.13*b* in the circled regions. As was discussed above, based on the tail of *FD* distribution function, although there are electrons and holes available for conduction, the number of available carriers is too low for practical devices. Below we will calculate this number, and examine the temperature dependence. It is important to note that for intrinsic semiconductors, the number of electrons and holes are equal, meaning that for every electron in the CB there is a hole in the VB.

In order to calculate the number of electrons N_e in the CB (which is equal to the number of holes in the VB) for an intrinsic semiconductor, we need integrate $N(E)$ over the CB. Recall that

$$N(E) = 2Z(E)F(E) \tag{10.17}$$

Also recall that

$$Z(E) = \frac{V}{4\pi^2}\left\{\frac{2m_e}{\hbar^2}\right\}^{3/2} E^{1/2} \quad \text{and} \quad F(E) = \frac{1}{1+e^{(E-E_F)/kT}} \tag{10.5 and 10.8}$$

$F(E)$ can be simplified by realizing that $E - E_F$ for this problem is $E_{CB} - E_F$ and is about 0.5 eV (for Si with $E_g = 1.1$ eV). kT at room T is about 0.025 eV. Thus the exponential term in the denominator is

$$e^{(E_{CB}-E_F)/kT} = e^{0.5/0.025} \gg 1 \tag{10.73}$$

Therefore $F(E)$ is simplified to a Boltzmann-like form as follows:

$$F(E) = e^{-(E_{CB}-E_F)/kT} \tag{10.74}$$

The integration limits are set for convenience with the zero of energy at the bottom of CB and integrate to ∞. Once again, this upper limit is convenient yet fictitous, but since $F(E)$ goes to 0, the convenience does not impose inaccuracy. The integral to evaluate is as follows:

$$\int_0^\infty \frac{V}{2\pi^2}\left\{\frac{2m_e}{\hbar^2}\right\}^{3/2} E^{1/2} \cdot e^{-(E-E_F)/kT} dE \tag{10.75}$$

This integral can be put into a form that can be readily integrated. First, notice that E_F is a constant. Then by grouping constants and energy variables, we obtain for the number of electrons in the CB (N_e) the following integral:

$$N_e = \frac{V}{2\pi^2}\left\{\frac{2m_e}{\hbar^2}\right\}^{3/2} e^{E_F/kT} \int_0^\infty E^{1/2} \cdot e^{-E/kT} dE \tag{10.76}$$

The form of the integral in equation (10.76) can be recognized as the definite integral:

$$\int_0^\infty x^{1/2} e^{-nx} dx = \frac{1}{2n}\sqrt{\frac{\pi}{n}} \tag{10.77}$$

where $n = 1/kT$ and $x = E$. This yields the following result:

$$N_{e} = \frac{V}{2\pi^2}\left\{\frac{2m_e}{\hbar^2}\right\}^{3/2} e^{E_F/kT} \cdot \frac{kT}{2} \cdot (\pi kT)^{1/2} = \frac{V}{4}\left\{\frac{2m_e kT}{\pi\hbar^2}\right\}^{3/2} e^{E_F/kT} \qquad (10.78)$$

This expression for N_e can be rewritten for convenience with several substitutions. The number of electrons in the CB per volume is $n = N_e/V$, $E_F = -E_g/2$, and for m_e the effective electron mass ratio to the rest mass, m_e^*/m_o, is used to obtain the final result:

$$n = \frac{1}{4}\left\{\frac{2m_o}{\pi\hbar^2}\right\}^{3/2}\left(\frac{m_e^*}{m_o}\right)^{3/2} T^{3/2} e^{E_g/2kT} = 4.82\times10^{21}\left(m^{-3}K^{-3/2}\right)\left(\frac{m_e^*}{m_o}\right)^{3/2} T^{3/2} e^{-E_g/2kT}$$

$$n = 4.82\times10^{15}\left(cm^{-3}K^{-3/2}\right)\left(\frac{m_e^*}{m_o}\right)^{3/2} T^{3/2} e^{-E_g/2kT} \qquad (10.79)$$

In equation (10.79) there are two temperature-dependent terms: an exponential term and a pre-exponential term, with the former being dominant and yielding an exponential increase in the number of electrons n found in the CB of an intrinsic semiconductor. At 25°C there are about 10^{16} electrons/m^3 in Si with a gap of 1.1 eV, assuming that the effective mass ration is near 1. In 1 m^3 of Si there are 10^{28} Si atoms/m^3. Thus only 1 in 10^{12} Si atoms contribute an electron to the CB. Consequently Si is a good insulator at room temperature. The same formula can be used to calculate the number of holes in the VB with only the substitution of the effective mass for holes m_h^* for that for electrons.

The conductivity for an intrinsic semiconductor should include the conductivity of both electrons and holes as follows:

$$\sigma = ne\mu_e + pe\mu_h = 4.82\times10^{21}\left(m^{-3}K^{-3/2}\right)\left(\frac{m_e^*}{m_o}\right)^{3/2} T^{3/2} e(\mu_e + \mu_h)e^{-E_g/2kT} \qquad (10.80)$$

for the case where $m_e^* = m_h^*$; otherwise, another ratio term is needed.

As was previously discussed for metals, the electron mobility generally decreases as temperature increases because the atomic vibrations increase with the temperature, and then so does the electron scattering. Consequently the conductivity for metals decreases with temperature. The opposite is true for semiconductors. It is seen in equation (10.79) above that n increases exponentially with temperature, since the exponential term has T in the denominator of the negative exponent and $T^{3/2}$ also appears in the pre-exponential. The conductivity, which is a function of n (and p), must then also display an exponential increase with temperature; this is shown schematically in Figure 10.14 along with n. These metal and semiconductor conductivity values need to be kept in perspective. In fact they differ by some 8 to 10 orders of magnitude at room temperature with good metals near 10^8 S/m and semiconductors near 10^{-2} S/m, depending strongly on the size of the band gap. Conduction in metals can decrease by an order of magnitude or two when heated to more than several hundred degrees, and semiconductor conduction increases by the same amount. However, these materials will remain far apart in the absolute values of the conductivity. The number of free electrons in metals is around 10^{28-29} m^{-3}, while for intrinsic semiconductors the number of carriers is around 10^{15-16} m^{-3}. For heavily doped Si, for example, the number of electrons can go as high as 10^{21} m^{-3}. Thus the number of available carriers for metals is typically considerably more than 10^6 the number in semiconductors.
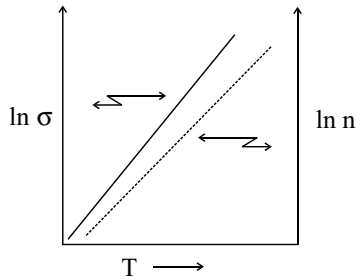
Figure 10.14 Logarithmic plot of the conductivity σ and electron concentration *n* for an intrinsic semiconductor versus temperature.

10.5.2 Extrinsic Semiconductors

Extrinsic semiconductors derive their electronic properties largely from the addition of impurity atoms called dopants that effect profound changes in the electronic behavior of the semiconductor, and yet are present in low concentrations, typically less than 0.1%. Usually doping takes place in inorganic semiconductors simply by the addition of atoms with different electronic configurations. These atom configurations must substitute (fit) for atoms on the semiconductor lattice. However, other methods of doping are possible, such as the ion implantation of chemically inert moieties that create localized damage in the semiconductor, and create electronic states that act as dopants.

Essentially dopants create localized electronic states in the band gap of the semiconductor. Previously in Chapter 9, when electronic energy bands were discussed, we saw that delocalized or extended states arise from the intermixing of wave functions from all the atoms in a solid. For example, for Si with a density of $2.33\,g/cm^3$ and an atomic weight of $28\,g/mol$, there are about 5×10^{22} atoms/cm^3, and the atoms are tenths of a nm apart. Thus wave functions of order of $10^{22\text{–}23}$ mix for Si of one cm^3 to form the Si electron bands, the so-called extended allowed electron states for the material. The resistivity of ultra pure Si is of the order of $\rho = 10^5\,\Omega$-cm, compared with Cu where $\rho = 10^{-6}\,\Omega$-cm. The maximum phosphorus (P) dopant that can be added to dope Si with an excess of electrons, called N-type doping, is the solubility limit for P in Si, which is about $10^{20}\,cm^{-3}$. This is less than 1% maximum concentration of P in Si or less than one P in 100 Si's, and it yields a resistivity for Si of about $\rho = 10^{-3}\,\Omega$-cm. For so-called metal oxide semiconductor field effect transistors (MOSFET's), the doping in the active device region of Si is about $10^{15\text{–}16}\,cm^{-3}$ P's in Si, and thus much lower than the solubility limit, yielding a resistivity of about $\rho = 5\,\Omega$-cm. At this doping level there is one P for every 10^7 Si atoms, and the dopant atoms are about $50\,nm$ apart. These are normal doping ranges that represent low concentrations relative to the number of semiconductor atoms. Consequently the dopant atoms are far apart and do not have overlapping wave functions, so they do not form bands. Rather, the dopants typically give rise to new localized electron states.

Figure 10.15 shows the distortion created if a larger and a smaller atom are inserted in an otherwise uniform lattice. First, the local order is altered, in that a number of surrounding atoms have altered coordinates. The distortion does not extend throughout the Si and damps out after several lattice spacings. Second, the binding potential in a region surrounding the substitutional atom is altered. From the Kronig-Penney (KP) model discussed in Chapter 9, we know that *a*, *b* and the potential barrier *V* (or *P* in the KP model)
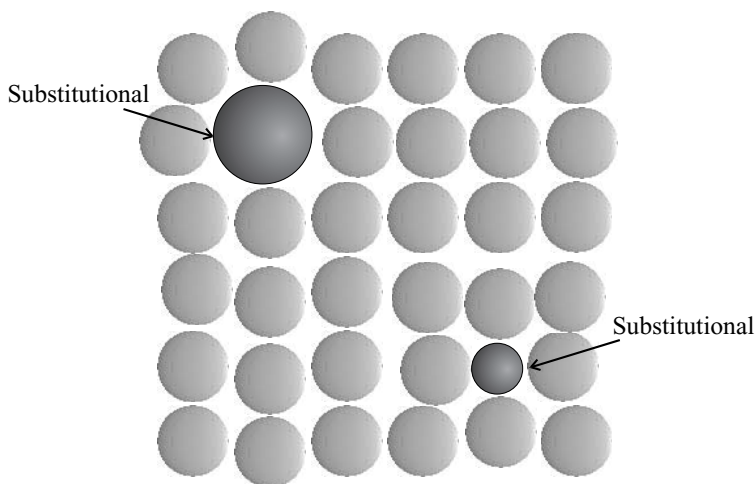
Figure 10.15 Substitutional impurities on a 2-D lattice.

are modified in the vicinity of the substitutional dopant atom. These modifications undoubtedly lead to different allowed electron states. However, because of the sparseness of the dopants, the new states are localized in that the wave functions do not overlap. Also, if we consider the electron configuration of Si, which lies in group IV of the Periodic Table, the outer electron shell, the M shell, has a s^2p^2 electron configuration that hybridizes in solid Si to sp^3, and this leads to tetrahedral covalent bonding of each Si to four nearest neighbor Si atoms. Two common dopants for Si are P, which dopes Si with excess electrons to create N-type Si, and B, which dopes Si with excess holes to create P-type Si. P is in group V of the Periodic Table with one more electron in the M shell with an s^2p^3 electronic configuration; B is in group III with one less outer electron than Si in its L shell with an s^2p^1 configuration. If we imagine P and B substituting on the tetrahedral bonded sp^3 Si lattice, then relative to a Si atom, P has an extra electron and B has one too few electrons. The result is that P produces new electron states that are filled with electrons and within a few hundredths of an eV below the CB of Si, while B substituted on the Si lattice produces new electron states that are initially empty and that are hundredths of an eV above the Si VB. These new states are shown in Figure 10.16. In Figure 10.16a the level of filled electron states near the CB is labeled E_D, where the D signifies that each of the filled states at this filled level can donate its electron to the CB, so the D subscript signifies that this level is a donor level. In Figure 10.16b the level of empty acceptor states is near the valence band with the subscript A. Table 10.2 shows several dopants that can be used to dope Si and Ge. It is clear that the levels of allowed donor and acceptor states are close to the CB and VB, respectively. With room temperature yielding about 0.025 eV energy, most all of the donor-occupied states will have sufficient kinetic energy available to the electrons to raise the energy to that of the nearby CB. Thus most of the donor states are ionized and yield their electrons to the CB. Likewise the empty acceptor level are close enough in energy from the filled VB that electrons from the top of the VB can occupy the empty acceptor states and leave holes behind in the VB.

It is useful to compare the conductivity for an intrinsic and extrinsic semiconductor. Recall the formula describes conductivity:

Energy
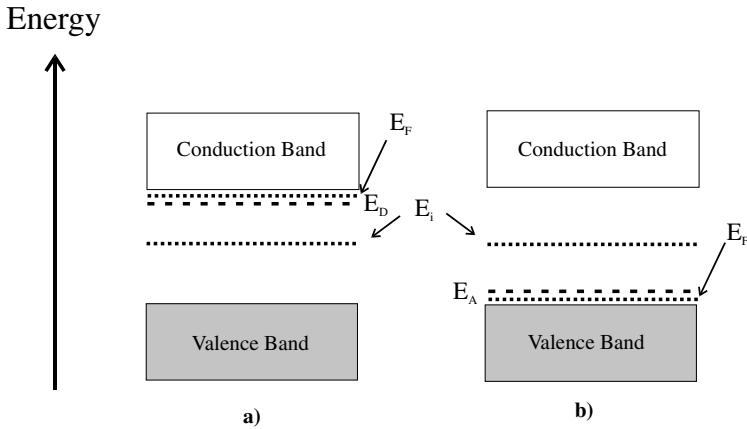
Figure 10.16 Parallel band scheme for (*a*) an N-type and (*b*) a P-type semiconductor showing the donor (E_D) and acceptor (E_A) levels, as well as the Fermi levels (E_F) and the intrinsic Fermi levels (E_i).

Table 10.2   Common dopants for Si and Ge and the levels produced

	Dopant	Si Level (eV)	Ge Level (eV)
Donors (below CB)	P	0.045	0.012
	Sb	0.039	0.0096
	As	0.049	0.0127
Acceptors (above VB)	B	0.045	0.0104
	In	0.160	0.0112
	Ga	0.065	0.0108

$$\sigma = N_c \mu_c e \tag{10.42}$$

where N_e is the number of carriers ($N(E_F)$) and μ_c the carrier mobility. We need to first compare the number of carriers. As was indicated above for intrinsic Si at room T, there are about 10^{10} electrons/cm^3 while for doped Si that number can rise to about 10^{20} electrons (or holes)/cm^3. Thus at room T we would anticipate a higher conductivity for doped Si.

In Figure 10.14 we saw that the temperature dependence of an intrinsic semiconductor exponentially increases in carrier concentration with a rise in T. The T dependence for an extrinsic semiconductor is more complicated with a typical dependence, shown in Figure 10.17. At very low temperatures when kT is less than the ionization energy for the dopant states (less than the energy difference between the dopant level and nearest band edge), the carrier concentration is the same as for an intrinsic semiconductor. This region is sometimes referred to as the "freeze-out" region where the thermal energy is insufficient to ionize all of the dopants. Depending on T, some fraction of the dopants can contribute carriers. This is followed by a relatively T independent carrier concentration where all the possible carriers are generated by the dopant, and these extrinsic carriers greatly exceed the intrinsic number of carriers. However, at still higher T, the number
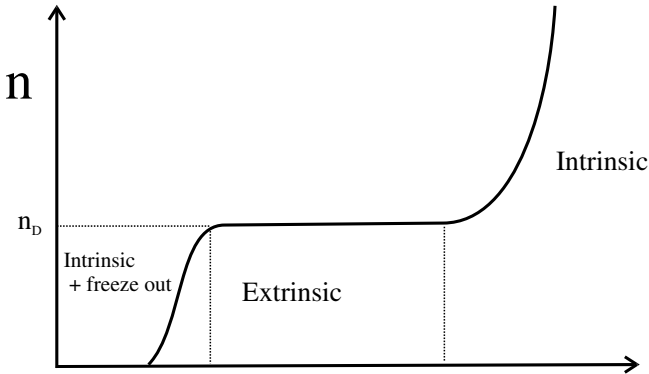
Figure 10.17 Temperature dependence of the carrier concentration for an N-type semiconductor.
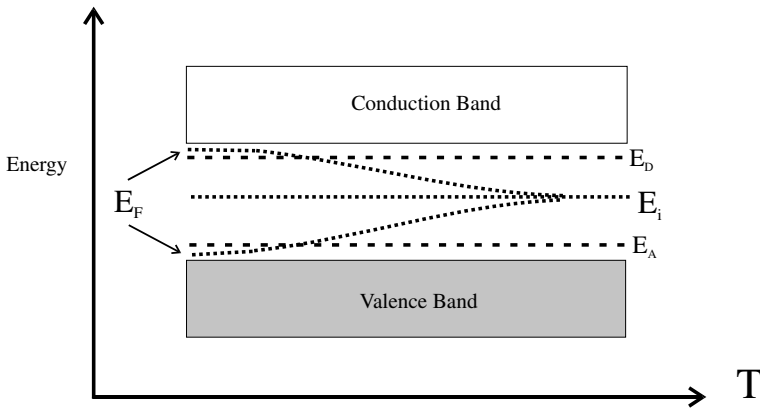
Figure 10.18 Temperature dependence of the Fermi levels for both N- and P-type semiconductors. The extrinsic Fermi levels (E_F) move toward the intrinsic Fermi level (E_i).

of intrinsic carriers produced exceeds the extrinsic number, and the total number rises exponentially with T.

As the carriers gain energy, more and more of both the intrinsic and extrinsic carriers will enter the conduction band. In the N-type Si with E_F, shown between the CB edge and E_D in Figure 10.16a, as T increases, E_F necessarily decreases in energy as electrons are first exhausted from the donor level. Ultimately E_F drops to E_i when all the donor electrons are exhausted and the material is dominated by intrinsic carriers. The phenomenon is similar for holes, in that the E_F starts out at low T in between the VB edge and E_A shown in Figure 10.16b. As T rises more, the electrons are able to reach the donor level raising E_F. Ultimately, as T continues to rise, the number of holes produced will outnumber the extrinsic carriers, and E_F will rise to the intrinsic level E_i. The cases for both N- and P-type Si (or other semiconductors) are shown in Figure 10.18.

Carrier mobility generally decreases with T in semiconductors for the same reason as for metals, namely electron scattering from the atoms. We can now proceed to relate all this information about carriers in terms of their quantity and mobility to determine how
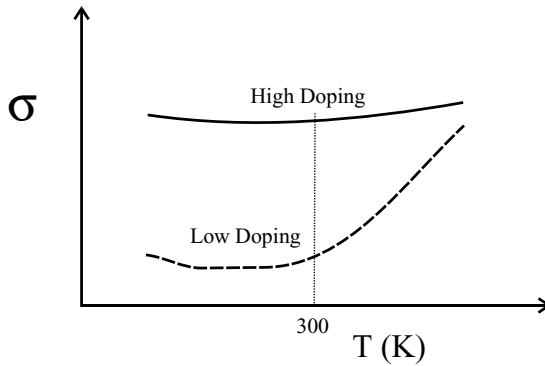
Figure 10.19 Temperature dependence of the conductivity (σ) for heavily and lightly doped semiconductors.

the conductivity varies with T, and this is shown in Figure 10.19 for both lightly and heavily doped semiconductors. As is expected, the heavily doped semiconductor displays a higher conductivity. As the temperature increases well beyond room temperature, both samples display higher conductivity, which approaches intrinsic conductivity as the intrinsic carriers dominate conduction (see Figure 10.17). Both samples are limited in conductivity by decreased mobility at higher T, and this is noticeable at lower T where a small dip due to scattering is seen as T rises.

10.5.3 Semiconductor Measurements

It is useful to consider the principles that underlie several important semiconductor measurements, in that the measurements are useful in themselves but further support and confirm the models used here to describe electronic conduction in materials. In this section three measurements are considered. The first is the Hall effect measurement that uses the simultaneous application of electric and magnetic fields, and can access the majority carrier type (electron or hole) and the majority carrier concentration. The second measurement is the determination of the effective mass using electron cyclotron resonance, and the third is the four-point probe measurement of resistivity.

For the Hall measurement we refer to Figure 10.20 where an electric field, $\mathbf{E}_x$, is applied to an N-type semiconductor solid in the $+x$ direction. This field gives rise to a current flux $-\mathbf{J}_x$. With this field applied and current flowing, a magnetic field is simultaneously applied in the z direction as is shown in this figure as $\mathbf{B}_z$. This magnetic field causes a deflection in the electric current that can be predicted by the so-called right-hand rule. With the thumb of the right hand pointed in the direction of $\mathbf{E}_x$, point the index finger of the right hand in the direction of $\mathbf{B}_z$. The middle finger of the right hand now points in the direction of the deflection, which is the $+y$ direction. This deflection of the electron flow in the $+y$ direction gives rise to a new electric field called the Hall field also in $+y$. The force from this Hall field is given as

$$\mathbf{F}_H = e\mathbf{E}_y \tag{10.81}$$

This force is in equilibrium with the Lorentz force, $\mathbf{F}_L$, that results from $\mathbf{B}_z$, and is given as
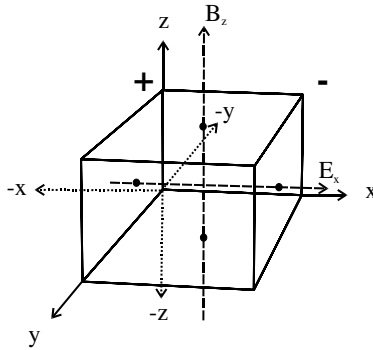
Figure 10.20   Hall measurement geometry with electric field applied (E_x) and magnetic field (B_z) causing a potential in y, the Hall potential.

$$\mathbf{F}_L = \mathbf{v}_{-x}\mathbf{B}_z e \tag{10.82}$$

where $\mathbf{v}_{-x}$ is the electron velocity in the $-x$ direction of the current. Equilibrium requires that $\mathbf{F}_H + \mathbf{F}_L = 0$, and this yields the Hall field E_y as

$$\mathbf{E}_y = -\mathbf{v}_{-x}\mathbf{B}_z \tag{10.83}$$

This equation along with equation (10.52) can be used to express the current flow:

$$\mathbf{J}_{-x} = N\mathbf{v}_{-x}e = \frac{-N\mathbf{E}_y}{\mathbf{B}_z e} \tag{10.84}$$

Equation (10.84) can be solved for the carrier concentration N as follows:

$$N = -\frac{\mathbf{J}_{-x}\mathbf{B}_z}{e\mathbf{E}_y} \tag{10.85}$$

A positive N is indicative of electrons serving as majority carriers. With the current $\mathbf{J}$ and a magnetic field applied, and therefore known, the potential that yields the field $\mathbf{E}_y$ is measured and N is calculated. Usually a Hall coefficient R_H is reported and given as

$$R_H = -\frac{1}{eN} \tag{10.86}$$

The sign of R_H indicates the carrier type, where negative ($-$) is for electrons and positive ($+$) for holes.

The effective mass for electrons can be measured using the electron cyclotron resonance effect, as is shown in Figure 10.21a. A slab of semiconductor is placed in a magnetic field and the field causes a precession of the electron in a circular or spiral motion as shown in this figure. The angular frequency of the precession ω_c is given as follows:

$$\omega_c = \frac{e\mathbf{B}}{m^*} \tag{10.87}$$

For a free electron the frequency is

$$v_c = \frac{\omega_c}{2\pi} = 2.8\mathbf{B}\,(\text{GHz}) \quad \text{for } B \text{ in kiloGauss} \tag{10.88}$$

Thus for $\mathbf{B} = 1\,\text{kGauss}$ $v_c = 2.8\,\text{GHz}$, which is in the microwave range. Experimentally one applies a magnetic field in the kiloGauss range, and parallel to the magnetic field an *RF* field in the microwave region is scanned in frequency. When the precession frequency is reached, resonance occurs and energy from the *RF* field is absorbed. To measure the frequency, the *RF* absorption versus applied *RF* frequency is measured, as illustrated in Figure 10.21. From v_c and $\mathbf{B}$, m* is calculated. This measurement is typically performed at low *T* near liquid He or N_2 temperatures to reduce collisions during precession.

The four-point probe method is often used to measure σ or ρ. This method avoids the problem of having very high resistance contacts that overshadow the measurement of the material properties. These so-called Schottky contacts, or junctions, will be addressed in Chapter 11. Figure 10.22 shows four sharp probes *A*, *B*, *C*, *D* that are typically equally spaced (e.g., at typically about 1 mm). A current *I* from a current source is applied to flow between *A* and *D*. The voltage *V* across *B* and *C* is measured. By Ohm's law, the relationship obtains

$$V = IR = I\rho\frac{l}{A} = \frac{I}{\sigma}\frac{l}{A} \tag{10.89}$$

where *R* the resistance is expressed in terms of the resistivity ρ or conductivity σ and the length *l* and area *A*. For the specific geometry shown in Figure 10.22, $l = dr$ the distance between probes, and *A* is $2\pi r^2$, or the surface area of a hemisphere through which the current passes. Then the formula above in differential form becomes
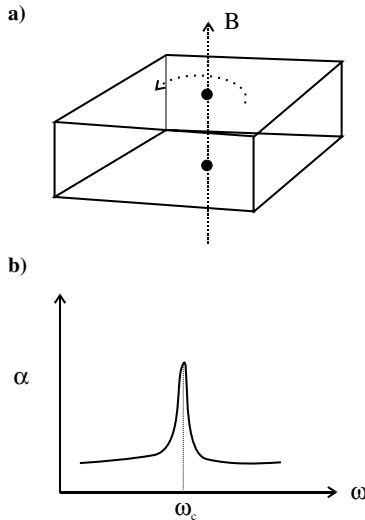
Figure 10.21 (*a*) Magnetic field **B** causes precession of electrons; (*b*) microwave absorption peak at precession frequency.
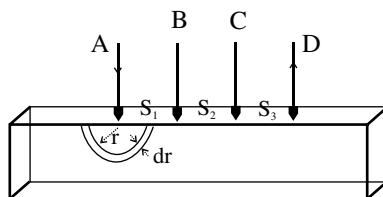
Figure 10.22   Four-point probe (*A, B, C, D*) geometry with equal separations (*S*).

$$dV = \frac{I}{2\pi\sigma r^2} dr \qquad (10.90)$$

To obtain the voltage drop from *B* to *C*, we integrate as follows:

$$\int_B^C dV = \int_{S_1}^{S_1+S_2} \frac{I}{2\pi\sigma r^2} dr \qquad (10.91)$$

The integration of equation (10.91) yields the result

$$V_C - V_B = \frac{I}{2\pi\sigma}\left(\frac{1}{S_1} - \frac{1}{S_1+S_2}\right) \qquad (10.92)$$

This formula is useful, but to be more accurate, the voltage drops at *A* and *D* must be taken into account because the current passes through those contacts. The corrected formula is as follows:

$$V_C - V_B = \frac{I}{2\pi\sigma}\left(\frac{1}{S_1} + \frac{1}{S_3} - \frac{1}{S_1+S_2} - \frac{1}{S_2+S_3}\right) \qquad (10.93)$$

For all the spacing equal to *S*, we obtain the following result for σ:

$$\sigma = \frac{1}{2\pi S V_{BC}} \qquad (10.94)$$

10.6   ELECTRICAL BEHAVIOR OF ORGANIC MATERIALS

The basis for the electronic properties of materials has been the electronic energy band structure, which is based on the periodic structure and the consequent periodic potential for crystalline materials. It is now worthwhile to reconsider the long and short range order, as was discussed in Chapter 2. A poignant example of the distinction between long and short range ordering is SiO_2. Recall that SiO_2 is composed of SiO_4 tetrahedra, as shown in Figure 2.2*a*, that have sp^3 bonding within a tetrahedron, and the tetrahedra are joined via the bridging O's at each apex of the tetrahedra. If the tetrahedra are arranged in a periodic manner, then SiO_2 will have long range order and be deemed to be crystalline, as shown in Figure 2.2*b*. Alternatively, the tetrahedra can be arranged with
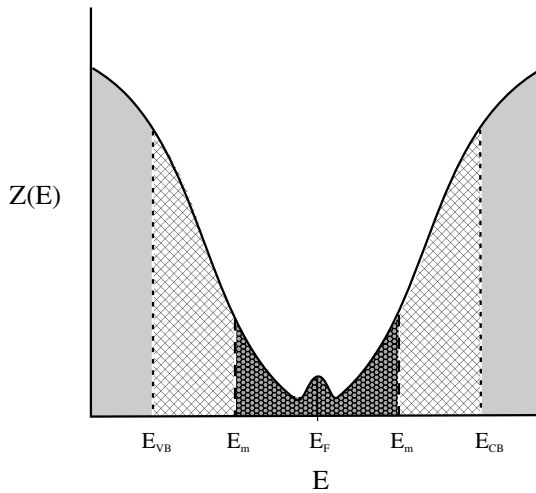
Figure 10.23 Electron energy band structure in terms of the density of electron states, $Z(E)$, versus electron energy. Valence and conduction band edges are shown with band tailing and localized states in the band gap.

random angles and thus the material will be amorphous, as shown in Figure 2.2*c*. The lack of long range order, but the presence of the same chemical bonding or short range order, gives rise to a more complicated electron energy band structure. Of course, the precise nature of the energy band structure will depend on the specific material, and the specific nature of the ordering or lack thereof. However, Figure 10.23 shows a typical electronic energy band structure for an amorphous material that contains many of the features of importance. Due to the chemical bonding and the short range ordering there is a VB and CB with extended states, as indicated by the energy at the band edges, E_{VB} and E_{CB}. In crystalline materials, the density of states goes to zero for both the VB and CB. However, because of the long range disorder this does not occur. Instead, there are allowed states penetrating into the energy gap, and this is called "band tailing." In between E_{VB} and E_m and in between E_{CB} and E_m are localized states due to the lack of long range order. The use of the "m" subscript refers to the mobility of these localized states, which appears similar to that of the delocalized states but, because they are localized, electron transport is via hopping from state to state. This conduction mechanism typically displays a significantly lower mobility, as compared with electron transport through the quasi-continuous states in the allowed energy bands in crystalline materials. Thus, the difference in energy between the two E_m energies indicates a mobility gap. In the mobility gap there is also the possibility of localized states due to defects, so-called defect states. In many cases the states have a peak and a mid-gap peak, as indicated in Figure 10.23. The example shown in Figure 10.23 is only shown to indicate the kinds of states that can occur, and indeed for which experimental evidence exists. A specific material with a specific level of disorder and defects can greatly alter the level and positions of the localized states.

RELATED READING

D. A. Davies. *Waves Atoms and Solids.* Longman, London. A well-written text covering many of the topics in Chapters 9, 10, and 11 with good insights.

R. E. Hummel. *Electronic Properties of Materials.* Springer-Verlag, New York. This text provides a well-written coverage of the material in Chapters 9, 10, and 11 at the appropriate level. The author has used this book as a text for the electronic materials part of the materials science course.

J. P. McKelvey. *Solid State Physics for Engineering and Materials Science.* Krieger. 1993. Higher level than Hummel, well-written, readable, and for the topics covered more complete.

M. A. Omar. *Elementary Solid State Physics.* Addison Wesley, Reading, MA. A text that covers many of the topics in Chapters 9, 10, and 11 and also many more topics not covered in the present text. A readable text on the subject.

EXERCISES

1. Calculate for a semiconductor, with $E_g = 0.1$, 1 and 10 eV, the temperature at which there is a 10% probability that the electron to be in the CB.

2. Explain why a half-filled valence band yields a material with the highest conductivity. In your response explain why the conductivity is not higher for, say, a $\frac{3}{4}$ filled VB.

3. Explain the difference between electron and hole conduction.

4. Why are electrons and holes in equal numbers in an intrinsic semiconductor. How can this balance become altered.

5. Calculate the number of electrons in the conduction band of a semiconductor with a band gap of 0.1, and 1 and 10 eV at room T. What assumptions have you made about the mass of the electrons.

6. Repeat problem 5 for 500 K. Discuss the differences in the results from exercises 5 and 6.

7. Discuss when the classical and QM theories for electrical conduction yield the same and different results.

8. Explain why dopant-produced states are localized states while the allowed states in the VB or CB are extended or delocalized states.

9. Explain how a substitutional impurity can act as a dopant.

10. Explain how a Cooper pair forms and leads to a superconducting band gap.

11. Explain how a Cooper pair can lead to superconductivity.

12. Explain why as a practical matter it is insufficient to measure only ρ when determining T_c. What other measurements can be done to confirm superconductivity.

13. Explain why amorphous materials exhibit an energy band structure similar to crystalline materials in some features but different in other features. List the similarities and differences.

14. Explain hopping conduction in amorphous materials.

15. (a) Using sketches of velocity space for electrons that include the Fermi velocity, explain how electronic conduction occurs in the presence of an electric field. (b) Using the sketches in (a) discuss the difference(s) from the classical model (Drude) for conduction.

16. (a) Sketch the parallel band structure for a p-type extrinsic semiconductor showing both the intrinsic Fermi level (E_i) and the real Fermi level (E_F). (b) On this sketch show the evolution of both Fermi levels with increasing temperature and discuss briefly why these change(s) if any are occurring. (c) Briefly discuss the difference(s) in the temperature behavior of semiconductor and metal conductivity.

17. Calculate the probability for an allowed electron state to be occupied below the Fermi energy at $0\,K$.

18. Discuss two ways to promote an e^- into the conduction band of an insulator.

19. Discuss what information is missing from the parallel energy band picture.

11

JUNCTIONS AND DEVICES AND THE NANOSCALE

11.1 INTRODUCTION

The work in the field of electronic devices has covered vast territory. Consequently the intent in this chapter is to use the basic ideas of electronic structure and properties from Chapters 9 and 10 and provide a first level of understanding about how the selected, important devices operate. Among these will be rectifiers, solar cells, and transistors, including those kinds of transistors that presently dominate computer technology. A majority of these important devices operate based on forming junctions between materials. Therefore it is useful to first consider the electronic consequences of joining materials to form junctions. It will be seen that the devices discussed in this chapter operate on the principles developed for junctions.

Electronic devices have evolved considerably since the vacuum tubes up to the 1960s, through transistors in the 1970s, to chips that contain huge numbers of devices that are integrated to a common purpose such as memory or logic in computers. All this evolution has occurred through the downsizing of device size that has been driven largely by the need for faster more capable computers. The technology has spilled out from computers to other areas such as biotechnology, medicine, sensors, communications, and photonics. Now we stand on another threshold in size that is at the scale of molecules and small groups of atoms, the nanoscale. Recent materials discoveries of nanoscale materials such as bucky balls (C_{60}), carbon (and other) nanotubes, quantum dots, and quantum wires have led researchers to think about new devices where the small size dictates the device characteristics. The nanoscale area has just begun, so the outflow of consumer products is modest at this time. However, for an electronics materials scientist it is a time for considering how the nanoscale device will impact electronics technology. To this end an introductory section on nanotechnology and nanodevices is included.

11.2   JUNCTIONS

In this section the characteristics of metal–metal, metal–semiconductor, and semiconductor junctions are discussed. The focus is on the behavior of the electrons when two materials are brought together, in particular, we consider the electrons in the valence and conduction bands that determine the electronic properties.

11.2.1   Metal–Metal Junctions

To understand junctions and the events that occur when materials are brought together, we commence with the parallel band picture as shown in Figure 11.1a for two metals. In this picture the metals display a partially filled valence band, with E_F showing the approximate position of the highest energy electrons. On the left is one metal M_1 and on the right another metal M_2, and each has a different E_F. The energy required to move an electron from E_F to infinite distance from the metal is called the work function ϕ_M, and also called the ionization energy. The zero of energy at infinite distance is oftentimes called the vacuum level. Thus the electrons bound in a material are at negative energies with respect to the vacuum level. When these two metals are joined, electrons from the metal with the higher E_F, metal M_2, flow to M_1, which has a lower E_F. Electrons endeavor to achieve the lowest allowed energies. Thus metal M_2 that was neutral before joining to metal M_1, as was M_1 before joining, now becomes charged. M_1, which gains electrons,
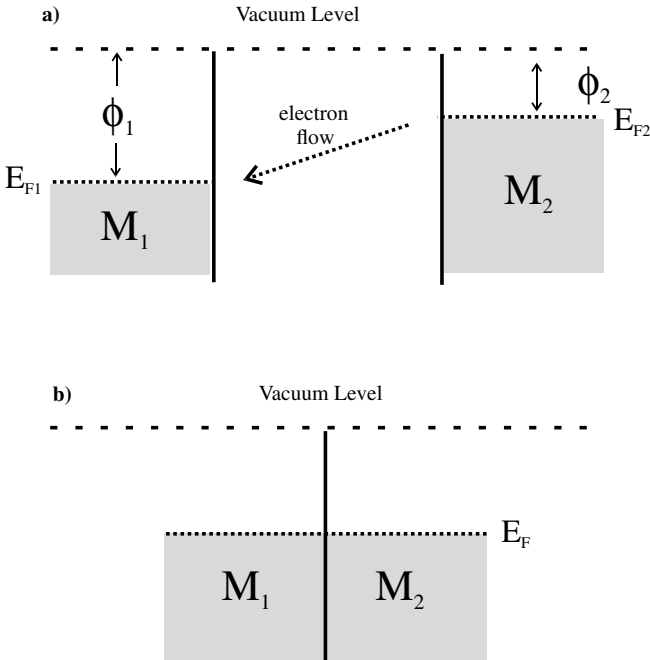
Figure 11.1   (a) Two separated different metals with Fermi levels (E_F) and work functions (ϕ) indicated; (b) the same metals as in (a) but after joining and at equilibrium.

becomes negative while M_2, which loses electrons, becomes positively charged. Likewise E_{F1} for M_1 rises while E_{F2} for M_2 drops. When all the electrons flow from higher to lower energy, equilibrium will result where the E_F's will level at $E_{F1} = E_{F2} = E_F$ as is shown in Figure 11.1b. The difference in potential generated between M_1 and M_2 as a result of the equilibration of the Fermi levels is called the contact potential ϕ_{1-2} and is given as

$$\phi_{1-2} = \phi_1 - \phi_2 \tag{11.1}$$

Because the contact potential results in no difference in Fermi levels, it is difficult to measure experimentally as a potential difference in an external circuit. Also the charge on the metals cannot create an electric field in the metals. However, one way to measure the contact potential is to first bring the metals in contact, so as to enable charge exchange, and then separate the bars as shown in Figure 11.2. Then with the separated bars held in vacuum, an electron beam is provided to move through the gap in between the separated metals. The contact potential ϕ_{1-2} will cause the electron beam to deviate in proportion to the magnitude of ϕ_{M1-M2}.

Another more practical method, called the Kelvin method, is to bring two different metals in contact and then separate them, but with close proximity. One of the metals is shaped as a sharp tip and has a known work function. The contact will enable charge exchange, and then when the metals are separated, the metal with the sharp tip is mechanically vibrated. The charge motion causes an ac current that can be reduced to zero by the application of an external dc potential that is exactly equal to the contact potential. Then, with a value for the contact potential obtained from the current nulling, and with one metal work function known, the other metal work function can be calculated using equation (11.1).

11.2.2 Metal–Semiconductor Junctions

The essential difference between metal–metal and metal–semiconductor junctions is that the semiconductor (and also for an insulator) unlike a metal can support an internal electric field as a result of extra charge being present. The region in the semiconductor where the field is generated is called the space charge region, and the evolution of the space charge region can be understood as shown in Figure 11.3. Figure 11.3a gives a parallel band picture of a metal on the left and a N-type semiconductor on the right. The semiconductor has a work function ϕ_S defined as the distance in energy from E_F to the vacuum level as for a metal. In this case for an N-type semiconductor, E_F is between the donor level and the conduction band (labeled as E_{FS}). The distance between the conduction
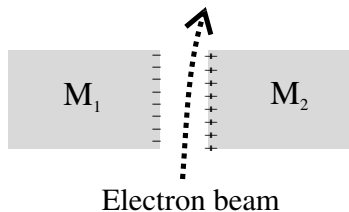
Electron beam

Figure 11.2 Two different metals previously in contact and equilibrated, and now separated and with a directed electron beam.
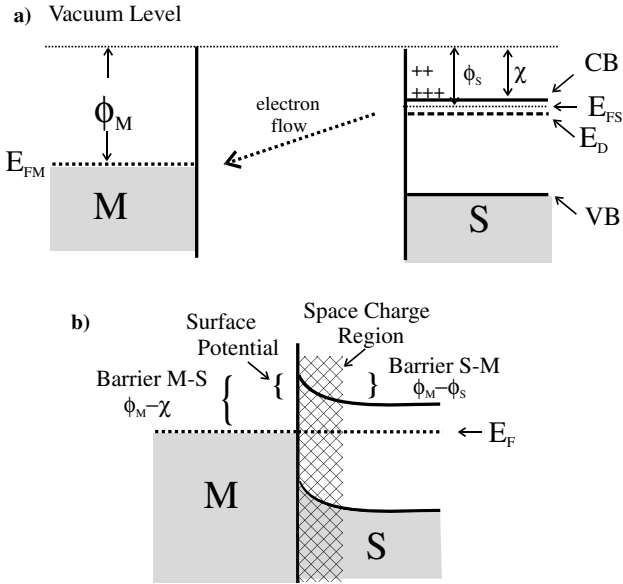
Figure 11.3 (*a*) A separated metal (*M*) and N-type semiconductor (*S*) with with Fermi levels (E_F), work functions (ϕ), electron affinity (χ), band edges (E_{vb} and E_{cb}), and donor level (E_D) with $E_{FM} < E_{FS}$; (*b*) *M* and *S* after contact and equilibration with band bending and the development of a space charge region (cross-hatched) and a surface potential.

band and the vacuum level is called the electron affinity χ. It is seen in the figure that for an N-type semiconductor near room temperature $\chi \cong \phi_S$. With the given E_F's (E_{FM} is the original Fermi level for the metal) the arrow indicates the direction for electron flow from the CB of the semiconductor to the metal, in order to reach equilibrium. The positive (+) charges in the semiconductor indicate the establishment of the space charge layer in the semiconductor. As electrons flow from the semiconductor to the metal, the E_F's will tend to equilibrate as before for metal–metal junctions. Thus the E_F for the semiconductor drops and the metal E_F rises until equilibrium is achieved, and the result at equilibrium is shown in Figure 11.3*b*. The electrons that flow to the metal come from the ionized donors in the semiconductor that are left behind, primarily in the region of the semiconductor near the junction. This cross-hatched region in Figure 11.3*b* becomes depleted of the majority carrier electrons in the N-type semiconductor, and consequently this region assumes a positive charge. The residual positive charge in the region depleted of majority carriers, the depletion region, is so indicated on the resulting energy band diagram as an upward hill toward the junction. This indicates that it becomes more difficult for more electrons to leave the semiconductor, and its magnitude is equivalent to the semiconductor surface potential ψ_s reached at equilibrium. Also this surface potential provides a barrier to electron flow from the semiconductor to the metal. In the opposite direction, a somewhat larger barrier for electron flow is seen. From what we have learned about the occupancy of electronic levels from the two-state model discussed earlier, it is straightforward to conclude that the electron current flow is proportional to e^{-E_B}, where E_B is the barrier energy to electron flow to an

allowed state. Therefore we can write for current flow, where current is the change in charge per time,

$$I = Ae^{-E_B/kT} \tag{11.2}$$

It can be seen from Figure 11.3b that there are nearly equal barriers in both directions. Later it will be seen that the application of an applied potential can alter this symmetry and permit current to flow preferentially in one direction. This is called rectification.

Figure 11.3 depicts the case of a metal–N-type semiconductor junction where $\phi_M > \phi_S$. There are three other cases to consider: $\phi_M < \phi_S$ with an N-type semiconductor, and $\phi_M > \phi_S$ and $\phi_M < \phi_S$ for P-type semiconductors.

Figure 11.4 displays the case for $\phi_M < \phi_S$ with an N-type semiconductor. It is seen that electrons achieve equilibrium by flowing from M to S, thereby leaving a space charge region in the semiconductor that is electron rich and thus negatively charged. The electrons in this region can easily migrate elsewhere to reduce the space charge, so the bands in the semiconductor bend downward, indicating a "down hill" migration toward the metal. In the other direction, namely M to S, there is a barrier whose height depends on the specific values for the ϕ's but is nevertheless uphill. In this case electron current can flow more readily in one direction.

Figures 11.5 and 11.6 display the cases of $\phi_M > \phi_S$ and $\phi_M < \phi_S$, respectively, for P-type semiconductors. Figure 11.5 for $\phi_M > \phi_S$ shows the electron energy bands in the space
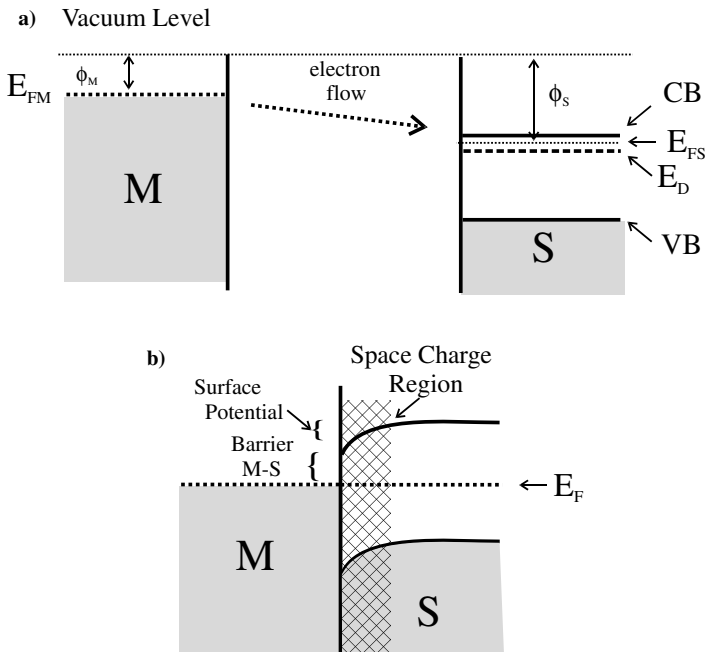
Figure 11.4 (*a*) Separated metal (*M*) and N-type semiconductor (*S*) with Fermi levels (E_F), work functions (ϕ), band edges (E_{vb} and E_{cb}), and donor level (E_D) with $E_{FM} > E_{FS}$; (*b*) *M* and *S* after contact and equilibration with band bending and the development of a space charge region (cross-hatched) and a surface potential.

a) Vacuum Level
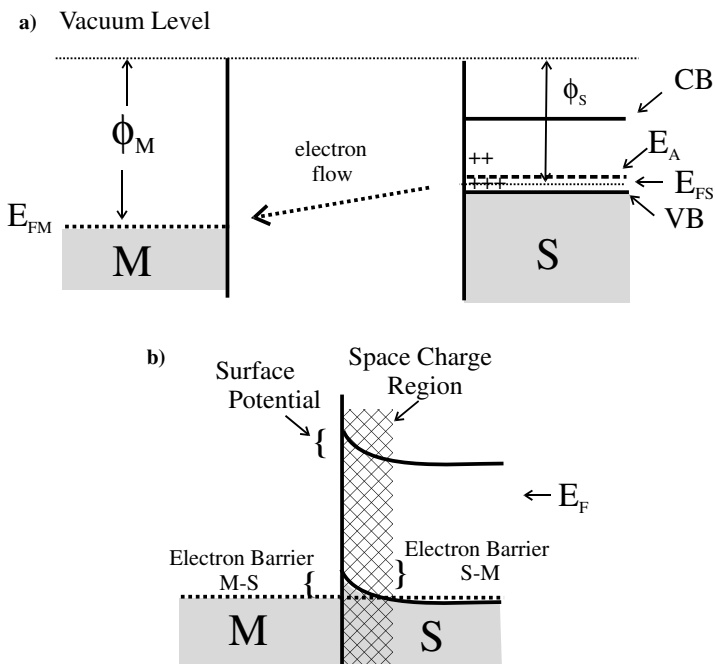
Figure 11.5 (a) Separated metal (M) and P-type semiconductor (S) with Fermi levels (E_F), work functions (ϕ), band edges (E_{vb} and E_{cb}), and acceptor level (E_A) with $E_{FM} < E_{FS}$; (b) M and S after contact and equilibration with band bending and the development of a space charge region (cross-hatched) and a surface potential.

charge region bending upward, indicating a barrier for electron flow from *S* to *M*. However, the majority carriers in the P-type semiconductor are holes. Thus the upward bent bands for electrons are actually down hill for holes. Therefore holes can flow readily from *S* to *M*, but electrons from the metal are impeded by the barrier from *M* to *S*. In Figure 11.6, for $\phi_M < \phi_S$, the electron energy bands in the space charge region of the semiconductor are bending downward, and hence this is actually a barrier for the majority carrier holes in the semiconductor. Electrons from the metal can easily pass into the semiconductor.

Those junctions where majority carriers can pass easily are called "ohmic" contacts where Ohm's law is obeyed. Those junctions where majority carriers must overcome exponential barriers are called "Schottky" contacts.

11.2.3 Semiconductor–Semiconductor PN Junctions

A PN junction is made by contacting a P-type semiconductor with an N-type semiconductor. As before, with metals electrons will flow so as to achieve the lowest energy configuration consistent with the appropriate physics. Thus, if we join an N- and P-type semiconductor, say, Si, then the donor level will be higher than the acceptor levels as is shown in Figure 11.7a. Electrons flowing from N-type to P-type would reduce the majority carriers on the N side, but because the electrons would fall into holes on the P-type

a) Vacuum Level

b)
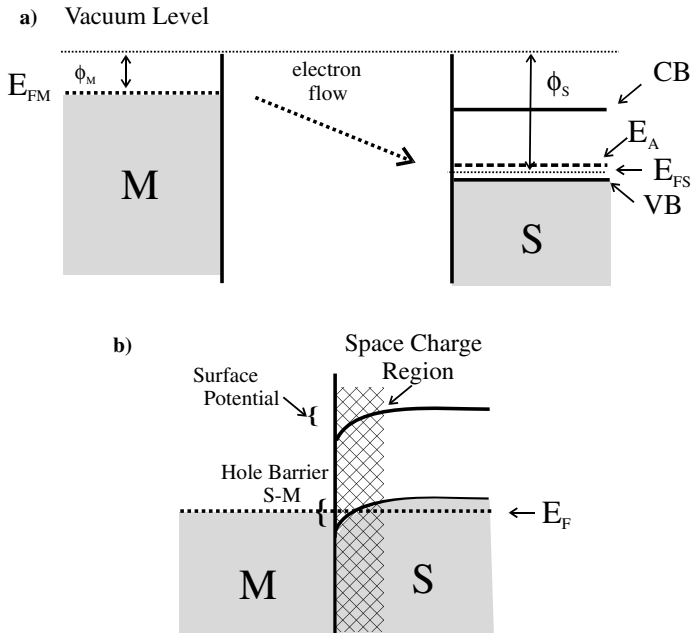
Figure 11.6 (*a*) Separated metal (*M*) and P-type semiconductor (*S*) with Fermi levels (E_F), work functions (ϕ), band edges (E_{vb} and E_{cb}), and acceptor level (E_A) with $E_{FM} > E_{FS}$; (*b*) *M* and *S* after contact and equilibration with band bending and the development of a space charge region (cross-hatched) and a surface potential.

side, the majority carriers on the P side are also reduced. Thus the region in between the P and N semiconductors, the junction region, is devoid or nearly devoid of carriers. This region is called the depletion region. Electrons from the N-type side have a large barrier to overcome, in order to traverse the barrier, as do holes from the P-type side. Thus it would be difficult for current flow in either direction. At first glance a PN junction doesn't seem very useful, but actually this kind of junction is at the heart of the present-day microelectronics industry, and the PN junction appears in many devices. However, to render this junction useful, typically an external potential is applied that is called a "bias." Figure 11.8*a* shows the same PN junction as in Figure 11.7 but with a forward bias, that is, a positive (+) potential on the P-type and a negative (−) potential on the N-type. The effect of the bias is to push the majority carriers together, reduce the depletion width, and reduce the barriers. The opposite occurs with reverse bias as is shown in Figure 11.8*b*. As we will see below for devices, the bias gives rise to rectifier devices and when combined with another PN junction results in a transistor.

11.3 SELECTED DEVICES

Simply expressed, electronic devices are devices that do something useful using electrical current or potential. Some devices perform a useful task simply by the output of a useful potential. Thermocouples that are metal–metal junctions fall into this category of passive
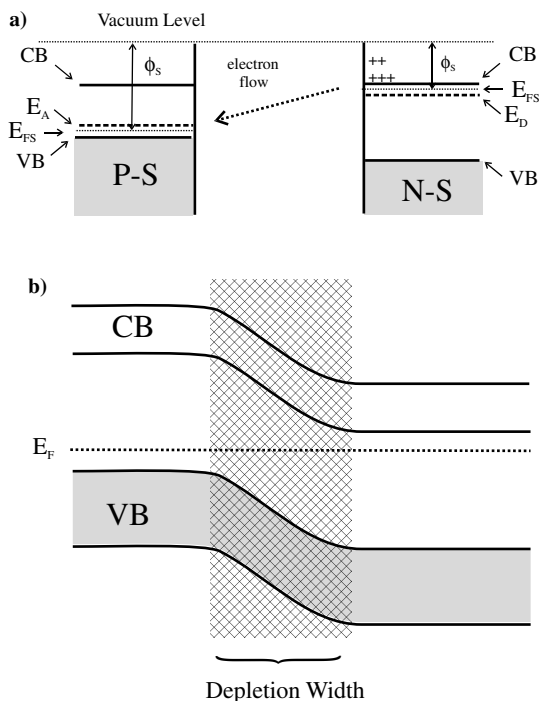
Figure 11.7   (*a*) Separated P- and N-type semiconductors with Fermi levels (E_F), work functions (ϕ), band edges (E_{vb} and E_{cb}), and doping level (E_A and E_D); (*b*) P-S and N-S after contact and equilibration with the development of a space charge region (cross-hatched), that is, depleted of carriers.

devices. As will be described below, thermocouples take advantage of the temperature dependence of the contact potential. Thermoelectric devices are another type of passive device, that uses a current to perform cooling, in which case the device embodies a thermoelectric refrigerator. On the other hand, there are active devices that switch a current, called transistors, and there are active devices that enhance a potential, called amplifiers, and that limit current to flow in one direction, so-called rectifiers. Further there are devices that convert one form of energy into another. For example, a device that converts photon energy into an electrical current is called a photocell. All of the solid state versions of electronic devices both passive and active take advantage of materials properties. A limited sample of active and passive devices is included in the following sections, in order to demonstrate the clever ways that technology makes use of the solid state properties.

11.3.1   Passive Devices

A thermocouple consists of metal–metal junctions. As was discussed above, when two metals are joined a contact potential develops. The contact potential changes with temperature; that is, the contact potential displays temperature dependence that is a manifestation of the Seebeck effect. Essentially, if a bar of metal (or semiconductor) initially at thermal equilibrium is heated at one end, the electrons in the hot end will attain a

a)

b)
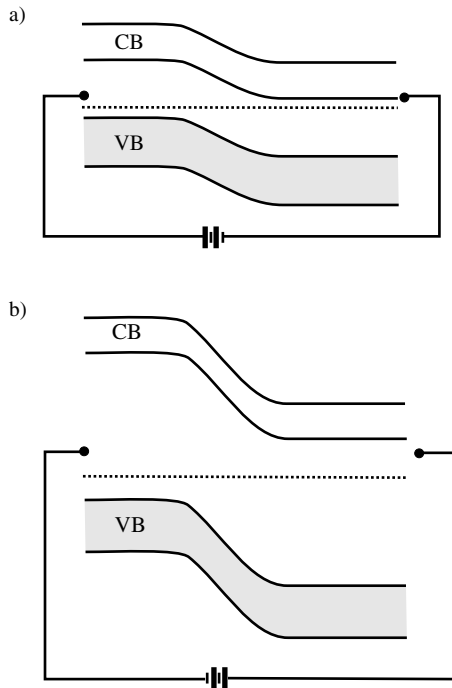
Figure 11.8 (*a*) Joined P- and N-type semiconductors with forward bias; (*b*) joined P- and N-type semiconductors with reverse bias.

higher kinetic energy and higher velocity than those in the cool end. Consequently electrons will flow from the hot to cold ends on average, causing a Seebeck potential V_S to develop that opposes electron flow. The size of the potential will vary with temperature and material, and the constant of proportionality between the change in temperature ΔT and the change in contact potential ΔV is called the Seebeck coefficient for a given material.

To produce a measurable output, two junctions between different metals *A* and *B* are formed as shown in Figure 11.9. The thermocouple junction is indicated by a bead that is often the result of spot welding wires of the appropriate metals. One junction is heated to T_h, and the other is held at a fixed temperature for reference, T_{ref}. At the heated junction each metal will exhibit a different Seebeck potential. The difference in potential between the hot and cold junction is a measure of the relative junction temperature, provided that the junctions can be calibrated against a known temperature standard. Tables of potential values versus temperature exist for many common metals and alloys from which thermocouples can be fabricated by spot-welding wires together. A set of two junctions is called a thermocouple, and is commonly used to measure temperature. The typical output for thermocouples is in the mV range and is easy to measure accurately. One way to use a thermocouple is to place one junction into an oven for which the temperature is desired, and the other is immersed in a constant temperature bath such as ice water to provide a stable reference potential. Modern thermocouples generate the potential of the ice bath for different combination of metals by simply using
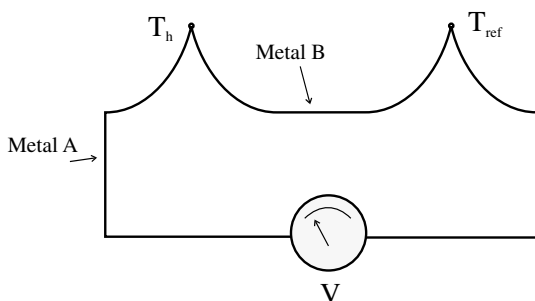
Figure 11.9 Thermocouple formed with two dissimilar metals *A* and *B* with two *AB* junctions (·). One junction is at a higher temperature (T_h) than the other (T_{ref}).
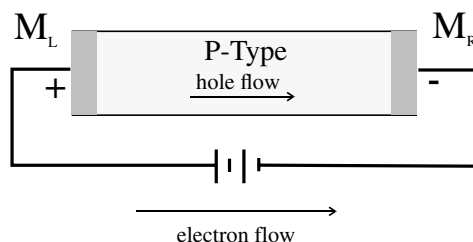
Figure 11.10 Peltier effect illustrated using two metal (M_L and M_R) junctions to a P-type semiconductor, with current flow in the semiconductor and external circuit indicated.

a precision high-impedance voltage source. Thus only one real metal–metal junction is necessary.

The Seebeck effect yields an electron flow as a result of a temperature difference or gradient. The inverse of that is to provide an external current source from a power supply to sustain a current that pushes electrons from the cold to the hot end of a bar of metal. This way cooling of the heated end of the bar will occur. Also the potential could be reversed so that the current flows in addition to the Seebeck current to produce accelerated heating of the cool end of the bar. The heating and cooling associated with current flow is called the Thompson effect. When a junction is involved the heating or cooling effect can be much larger, and is called the Peltier effect. Figure 11.10 illustrates the Peltier effect with metal–semiconductor junctions. This figure shows a P-type semiconductor where the majority current carriers are holes and there are two metal semiconductor junctions M_R and M_L. In the external circuit the current is carried by electrons with a flow as shown in the figure, and that derives from the direction of the dc power supply. For current to flow, there must be a change in carriers near the contact electrodes and a recombination of electrons from the metal with holes from the semiconductor. Electrons at the power supply potential are injected into the semiconductor at the right side contact, and these electrons combine with holes and lose their energy to the semiconductor. Thus the semiconductor gets hotter near the right-hand contact M_R. At the left-hand contact M_L electrons and holes recombine, and the energy of the holes is removed by the current flow in the external circuit. Hence cooling occurs at the left side contact. Thus the Peltier effect gives rise to the construction of a thermoelectric refrigerator and/or heater.

11.3.2   Active Devices

Devices that control and adjust current in a circuit are herein termed "active" devices. As will be discussed below, the devices chosen for discussion are among the most commonly found devices, and they are all based on junctions for their operation.

11.3.2.1   Rectifiers   Rectifiers are devices that permit current to flow predominantly in one direction. In this way an ac current can be transformed into a current that flows in one direction or a dc current. Typical solid state rectifiers can be designed using metal–semiconductor and semiconductor–semiconductor PN junctions. Many rectifiers are two-terminal devices with one terminal at the P and the other on the N type semiconductor or on the metal, and are consequently called diodes.

For metal–semiconductor junctions we refer back to Figure 11.3, and the fact that the current in either direction is given by exponential equation (11.2). Starting from equation (11.2), we can write an expression for the current from M to S as follows:

$$I_{MS} = Ae^{-(\phi_M - \phi_S)/kT} \tag{11.3}$$

In equation (11.3) we ignored any difference between ϕ_S and χ. Likewise we can write for the reverse direction,

$$I_{SM} = Ae^{-(\phi_S - \phi_M)/kT} \tag{11.4}$$

Thus the net current $I_{net} = I_{SM} - I_{MS} = 0$. This situation is markedly changed with the application of a bias voltage. First we apply a forward bias to the metal–semiconductor junction shown in Figure 11.3, which lowers the barrier from S to M. Since this is an N-type semiconductor, for a forward bias we apply a negative potential to the semiconductor, and the current I_{SM} is modified as

$$I_{SM} = Ae^{-(\phi_S - \phi_M - eV)/kT} \tag{11.5}$$

The net current is no longer 0 and can be written as the difference between forward and reverse currents:

$$I_{net} = I_0\left(e^{eV/kT} - 1\right) \tag{11.6}$$

where I_0 is called the saturation current. This formula is often referred to as the rectifier formula. Thus it is seen that for the bias potential $V = 0$, $I_{net} = 0$ as shown before. For $+V$, I_{net} rises exponentially with V and for $-V$, I_{net} is essentially I_0 or the saturation current. These conditions are shown for an ideal rectifier in Figure 11.11a.

Similarly a PN junction can also be used for rectification. As was shown in Figures 11.7 and 11.8, the depletion region provides a barrier, and forward and reverse bias can alter the barrier height. Also we can anticipate the use of a rectifier formula such as equation (11.6) developed above. However, in a PN junction we have to keep track of both majority and minority carriers in both the P and N sides of the junction. If a forward bias V is applied to the PN junction as shown in Figure 11.8a with a positive (+) potential on P and a negative (−) on N, the majority carriers that were in equilibrium on their respective sides are pushed into the junction area and combine with their opposite, while minority carriers on each side of the depletion region of the junction will rise in an expo-

a)

b)
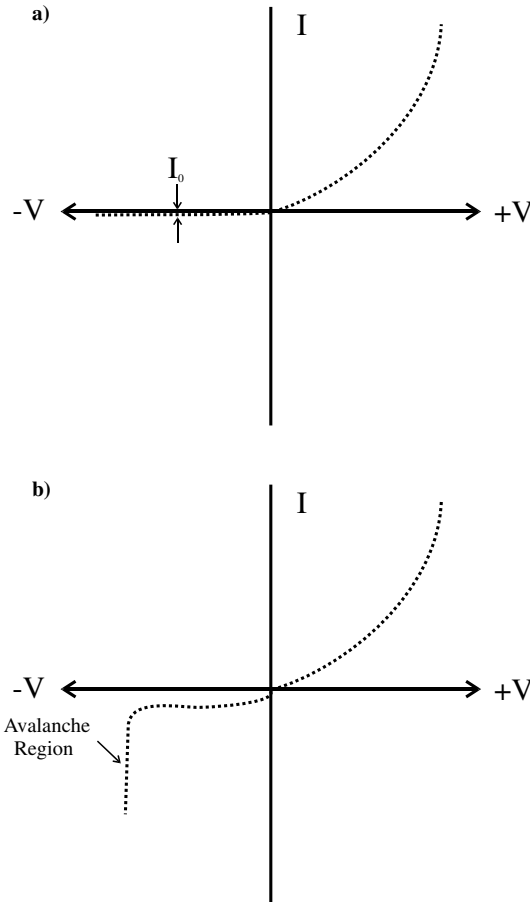
Figure 11.11 (a) Ideal diode current (I) versus applied voltage (V) characteristic showing current in one direction; (b) a more realistic diode characteristic showing some reverse current and avalanche breakdown in the reverse current region.

nential way as a function of V. The holes on the N side rise to $p_N e^{eV/kT}$ and electrons on the P side rise to $n_P e^{eV/kT}$. These are nonequilibrium values and are greater than the values in the bulk of the semiconductor on either side. Consequently these excess minority carriers will diffuse into the bulk. Recall from Chapter 5 on diffusion that the minority carriers diffuse with a characteristic length L, as is given by a slightly modified equation (5.113):

$$L_p = (D_p \tau_p)^{1/2} \quad \text{and} \quad L_n = (D_n \tau_n)^{1/2} \tag{11.7}$$

Where rather than a diffusion time t being used, τ represents the lifetime for the minority carrier. The excess of minority carriers is the difference in the number with and without an applied potential V, and the values are expressed as

$$p_{\text{excess}} = p_N^{eV/kT} - p_N \quad \text{and} \quad n_{\text{excess}} = n_P e^{eV/kT} - n_P \quad (11.8)$$

The hole (minority carrier) quantities at a distance x in the N and P sides of the junction can be expressed as

$$p(x) = \left(e_N^{eV/kT} - p_N\right)\left(e_p^{-x/L} + p_N\right) \quad \text{and} \quad n(x) = \left(n_P e^{eV/kT} - n_P\right)\left(e_n^{-x/L} + n_P\right) \quad (11.9)$$

The current density J for holes and electrons can be calculated from a diffusion formula (equation 5.1 or 5.11) as

$$J_p = -eD_P\frac{dp}{dx} = \frac{eD_P p_n}{L_P}\left(e^{eV/kT} - 1\right) \quad \text{and} \quad J_n = -eD_n\frac{dn}{dx} = \frac{eD_n p_p}{L_n}\left(e^{eV/kT} - 1\right) \quad (11.10)$$

This yields a net current flux in the form of the rectifier formula:

$$J_{\text{net}} = \left\{\frac{eD_P p_n}{L_P} + \frac{eD_n p_p}{L_n}\right\}\left(e^{eV/kT} - 1\right) \quad (11.11)$$

Finally the saturation current is given as

$$J_0 = \left\{\frac{eD_P p_n}{L_P} + \frac{eD_n n_p}{L_n}\right\} \quad (11.12)$$

Figure 11.11b displays more realistic rectifier behavior with an initially slightly changing current in the reverse bias regime. After a large reverse bias a sudden and large increase in the reverse bias current is observed. In this high-current region electrons are accelerated by the applied potential. The rapidly moving electrons cause more electrons to gain energy and contribute to the reverse current. This is sometimes called impact ionization, and it refers to the ionization caused by electron impact. Ultimately this cascading effect, called an avalanche, creates high current. In this region the high currents can cause permanent damage to the material, and thus to a breakdown. The avalanche effect that leads to breakdown in lightly doped semiconductors can be tolerated and rendered indefinitely useful in heavily doped semiconductors. This kind of rectifier is called a Zener diode or Zener rectifier. The rectifier can be used as a voltage reference because it breaks down at precise voltages that depend on the doping.

The rectifier formula for PN junctions assumes that the transfer of charge is via diffusion of carriers, electrons, and holes. However, for very heavily doped P and N semiconductors the junction formed will have a thin depletion width of the order of a few nanometers. With high applied bias potentials, the occupied electron bands can become aligned with empty bands on the other side of the junction. In this event and with a thin barrier or depletion region, electron tunneling can occur. The resulting rectifier is called a tunnel rectifier or diode, and sometimes an Esaki diode after its inventor. The operation of a tunnel diode can be understood with the use of Figure 11.12. Figure 11.2a illustrates the band structure with no bias applied to a heavily P- and N-doped junction. The heavy doping creates a narrow depletion width that acts as a tunneling barrier. With no applied bias, there is no current flow and this is the point on the I-versus-V plot corresponding to a) in Figure 11.12e. E_F is indicated by the dotted line in each frame of Figure 11.12. Note that filled electron states are not in energy proximity to empty states across
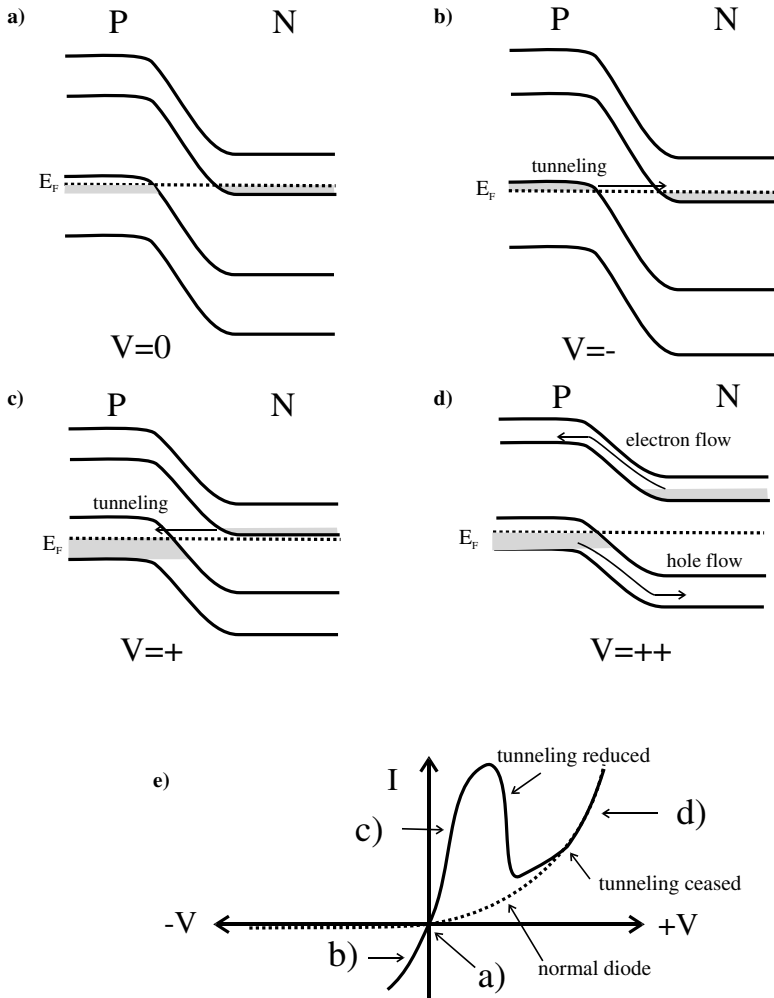
Figure 11.12 Heavily doped PN junction with different bias states illustrating tunnel diode operation: (*a*) No applied bias and no tunneling; applied reverse (*b*) or forward (*c*) bias yields filled states above empty states and tunneling occurs; (*d*) tunneling ceases for a larger forward bias; (*e*) the *I-V* characteristic for a tunnel diode with the bias conditions indicated and with an ideal diode characteristic.

the narrow depletion zone. However, when either a forward (Figure 11.12*c*) or reverse (Figure 11.12*b*) bias is applied, occupied states align with empty states and with the narrow barrier, so electron tunneling can occur in the direction indicated by the arrows in Figure 11.12*b* and 11.12*c*. The tunneling current is indicated on the *I*-versus-*V* curve (Figure 11.12*e*) as points *b* and *c* for reverse and forward bias, respectively. As more forward bias is applied, the bands reach optimum alignment, which is characterized by a maximum in the tunneling current. After the maximum, the bands begin to be mis-aligned with increasing forward bias. Consequently the tunneling decreases, and the current decreases. This downward slope of the *I*-versus-*V* characteristic is called a neg-

ative resistance region, since an application of Ohm's law to this region requires a $-R$. Finally, with the application of a large forward bias as shown in Figure 11.12d (or reverse bias but not shown in Figure 11.12), the filled states are no longer aligned in energy with empty states, and thus the only mode for current flow is diffusion of carriers as with a normal (nontunneling) PN junction. Consequently the I-versus-V characteristic returns to the normal diode case as shown by point d in Figure 11.12e.

Although not covered here, the negative resistance characteristics of the tunnel diode can be exploited in oscillator and amplifier circuits.

11.3.2.2   *Photocells*   Another important kind of PN junction diode is a photodiode. Figure 11.13a shows the cross section of a PN junction with a thin P region so as to permit light to penetrate to the depletion region. The photons with energy greater than the band gap can create electron-hole pairs by promoting valence band electrons to the conduction band leaving an equal number of holes behind. The electron-hole pairs created away from the depletion width are of little consequence and eventually merely re-combine. However, as is depicted in Figure 11.13b, the electron-hole pairs created in the depletion width are of great interest because the electrons can fall "down hill" into the CB of the N-type semiconductor while the holes flow "uphill" to the P-type side of the junction. This way current is created and photon energy is effectively transduced into electrical energy.

To optimize a photodiode, it is best to operate in the reverse bias region. In this case the downhill shape of the energy bands for both electrons and holes maximizes the current. Also the only current flowing in reverse bias is the saturation current. As was mentioned above, the thickness of the top layer needs to be minimized to allow maximum photon flux to the depletion region, and the area needs to be maximized to harvest as
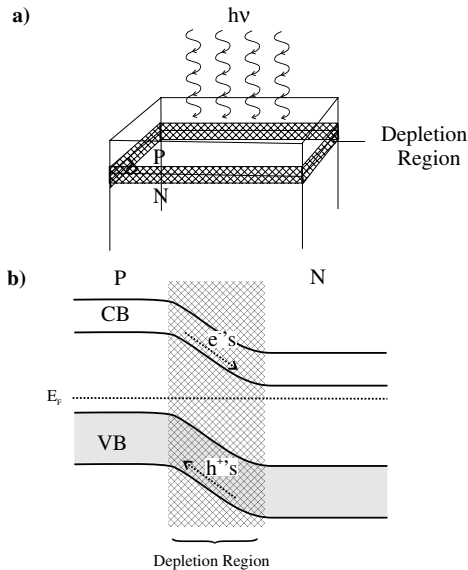
Figure 11.13   (*a*) Photocell PN junction with a thin P-type region to allow light penetration; (*b*) the junction depletion region where photo-created electron-hole pairs are separated.

much of the light as possible. The depletion region is where the useful electron-hole pairs are created by the photons. This region can be maximized by using lightly doped semiconductors and even by the use of an intrinsic semiconductor that is insulating in between P and N regions. Such a diode is called a PIN diode. The current produced can be fed to a resistor where a potential can be measured.

11.3.2.3 Transistors Transistors are also comprised of junctions. Typically transistors are solid state devices that can control the flow of current. A good analogy is a water valve as shown in Figure 11.14. Water current flows into the valve body and out, as is indicated by the arrows, and the control of the water current is by means of a valve that can turn the flow on or off or at various flow rates. The simple water valve shown in Figure 11.14 is a three-terminal device: in, out, and control. A transistor is also a three- (at least) terminal device that controls the electrical current in and out (two terminals) by means of a third terminal. Here we discuss three types of transistors: the bipolar transistor that uses two carrier types for operation, electrons and holes; the metal oxide semiconductor field effect transistor (MOSFET) that uses minority carriers and is presently the heart of the computer industry; and a novel organic transistor that appears to operate like a MOSFET but actually is different and operates using majority carriers. This organic transistor is called a thin film transistor, TFT.

Bipolar Transistor. Figure 11.15*a* shows a bipolar transistor as consisting of two PN junctions. Specifically, this is a NPN transistor consisting of two PN junctions with a common P region sandwiched in between two N regions. It is also possible to fabricate a PNP transistor with two P regions and a N region sandwiched in between. The three regions in this kind of transistor are contacted and the three contacts are called the emitter (e), the base (b), and the collector (c). As before for PN junctions, the depletion regions are shown as crosshatched areas. Figure 11.15*b* shows the energy band structure for this kind of device with no applied bias. There are two depletion zones at the two junctions, and these depletion zones cause the electron energy bands to rise from the N to P regions. These rises impede the flow of electronic charge into the device from emitter to base and from collector to base. By means of judiciously applying external bias to the junctions, various kinds of device operation can be obtained. For example, a common application is to amplify an electrical signal. We can imagine an audio signal (music) being transduced by a microphone to electrical signals of the same frequency distribution. The electrical signals, electrons, are placed onto the emitter of our NPN bipolar amplifier. The NPN amplifier is set up with a forward biased e–b junction and a reverse biased c–b junction as is shown in Figure 11.15*c*. The incoming current is largely dropped as a potential across the resistance of the e–b junction. However, after the slightly uphill journey of electrons injected into the e–b junction depletion region, those "lucky" electrons that drift near the center of the junction will "feel" the large down hill
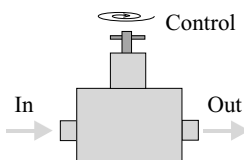
Figure 11.14   Water valve as an example of a three-terminal device: in, out, and control.
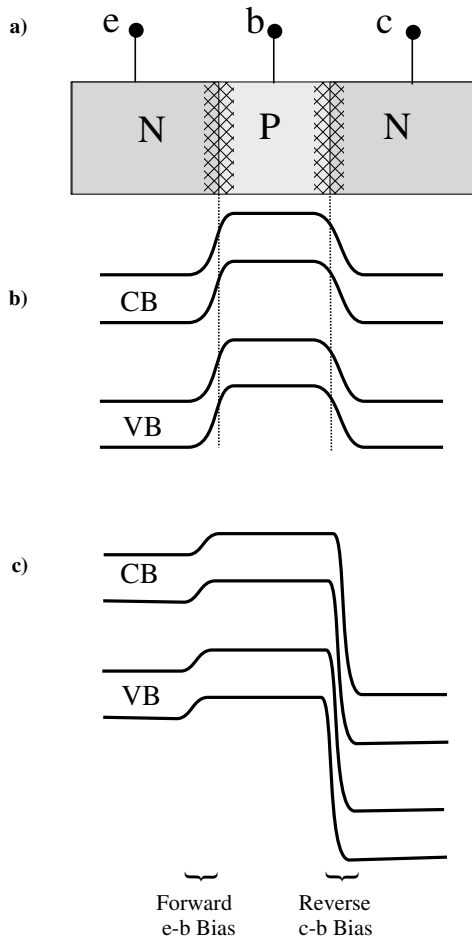
Figure 11.15   (*a*) NPN bipolar transistor showing two PN junctions and the three terminals: emitter (e), base (b), and collector (c); (*b*) the energy band structure of an unbiased NPN transistor showing the depletion regions; (*c*) the NPN transistor with forward-biased emitter-base terminals and reverse–biased collector-base terminals.

potential created by virtue of the reverse bias at the c–b junction. This is a strongly forward bias to an electron entering from the b side of the c–b junction. This "lucky" electron gains energy, and its energy is amplified, and if desired, amplified to the extent necessary to drive speakers that transduce the electrical signal to an acoustic signal. If we vary the reverse bias on the c–b junction, we can vary the volume of the final acoustic signal. The emitter region is usually heavily doped to enhance the operation of this amplifier, thereby increasing the number of carriers drifting into the base region. Also the base is made thin so as to improve the chances of an injected electron in reaching the downward potential to propel it to the collector. The gain of an amplifier is a measure of the size of the output signal potential relative to its input potential. There are other configurations of amplifiers that can increase the output signal and provide regulation.

They are all based on the two-junction bias ideas presented above. In addition it is straightforward to consider that the base potential can prevent emitter-injected carriers from reaching the collector. For example, a small positive or a negative potential on the base can reduce the possibility of electrons drifting to the collector. Thus the base potential can turn the device on by allowing current to flow to the collector, or off by not allowing current flow. This is an example of a bipolar switch. This kind of switch can be fabricated with a thin base region and can thus operate very fast, and in certain circumstances this switch is used for fast logic operations in computers. Older generations of mainframe computers used bipolar switches for its most competent and fastest switching operations. More recently much of the more costly bipolar circuitry has been replaced by less costly configurations of MOSFET devices, as discussed below.

It should be noticed that a bipolar device uses both electrons and holes for operation.

MOSFET. The MOSFET is at the heart of present-day microelectronics, and this device is based on silicon as the semiconductor material. The main reason for this choice of Si as the main semiconductor material used in microelectronics has to do with the ability to reduce surface and interface electronic states to acceptable levels via films of SiO_2, which also serves as a dielectric in the devices. Intrinsic interface electronic states arise from the termination of a crystal at its surface (more on this topic is beyond the scope of this text). The typical MOSFET is shown as Figure 11.16*a*, which specifically displays an N-channel MOSFET. A N-channel MOSFET is constructed on P-type Si.

The MOSFET device is usually a three-terminal device with the terminals named Source (S), Drain (D), and Gate (G). A back side contact is sometimes used, but it is not necessary for device operation. The MOSFET is most often used as a switch where the S to D current is switched on and off using the potential at G to effect the switching. The channel region separates the N-type S and D, and this is the region where carriers flow from S to D to turn on the device. Notice that the N-type doping in S and D is labeled n^+ where the superscript + indicates heavy doping. Initially the N-type S and D are separated by a P-type region. Thus, if one injects a majority carrier, an electron, from S into the channel for the purpose of effecting conduction, the electron will eventually re-combine with the holes in the P-type channel region before the electron ever reaches D. If a negative potential is applied to the gate, $-V_G$, the majority carriers that are holes in the Si substrate will be attracted to the channel. This condition is called accumulation, and it refers to majority carriers that will accumulate in the channel. Accumulation further prevents S–D current, and thus maintains the off state for the MOSFET. Now, as the V_G is made, more positive the majority carrier holes are repelled from the channel. When the number of holes in the channel becomes less than the equilibrium number, the channel is said to be depleted. The condition in the channel goes from accumulation with $-V_G$ to depletion with a more positive V_G. As V_G gets yet more positive, the minority carrier electrons in the P-type Si will outnumber the holes that are being depleted. The resulting condition is called inversion because the majority carrier type in the channel (and only in the channel) has been changed from holes to electrons; namely the carrier type has been inverted by virtue of changing V_G.

The important aspect of inversion is that now electrons from S injected into the inverted channel can make it to D and thus turn on the device. The *I*-versus-*V* characteristic is shown in Figure 11.16*b*, and it is seen that for the S–D current, I_{SD} is zero as V_G increases but then rises steeply when V_G is sufficiently positive in this N-channel device to effect inversion and turn-on of the device. A V_G value at which I_{SD} is sufficiently high to indicate unambiguous turn-on is called the threshold voltage and labeled as V_T. The
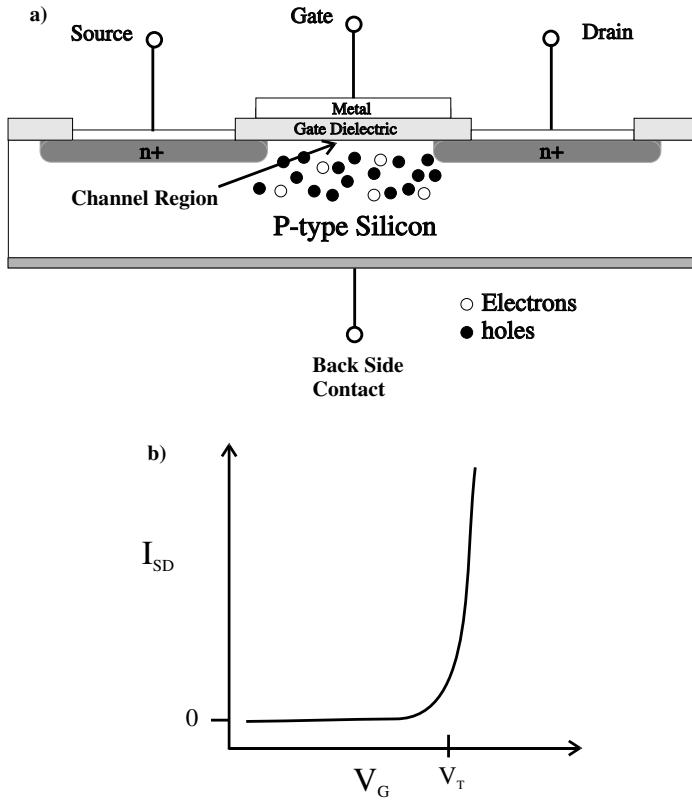
Figure 11.16   (*a*) N-Channel metal oxide semiconductor field effect transistor (MOSFET) with three terminals—source, gate, and drain—and with heavily doped source and drain regions; (*b*) the source to drain current (I_{SD}) versus gate voltage (V_G) characteristic for the MOSFET, showing on and off states at the threshold voltage (V_T).

on and off states are all that are necessary to perform computer memory operations based on boolean 1 and 0. A MOSFET can be configured for computer memory application, as shown schematically in Figure 11.17. In this figure, D is connected to a capacitor and then to ground. By the operation of the gate the device can turn on, as discussed above. With the device on, I_{SD} will flow and charge the storage capacitor. The device is turned on using the connection to the so-called word line. Current flows from the bit line to charge the storage capacitor. Whether the capacitor already has a stored charge (a stored binary bit 1) or not (a binary bit 0) can be sensed using the word line again and the bit line. If the word line turns the device on, then current will flow if the capacitor is at binary 0. But, if the storage capacitor is already charged, no current will flow from the bit line. This way boolean logic can be read and written to memory. This kind of memory is called dynamic random access memory (DRAM), and it must be refreshed by the power supply.

Of course, for a P-channel MOSFET device that is fabricated on N-type Si, the turn-on will be at negative values of V_T. Similarly a P-channel DRAM can be fabricated. Both
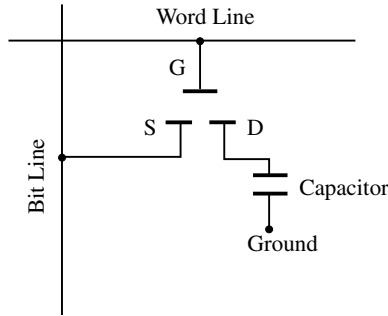
Figure 11.17 MOSFET connected to word and bit lines that operate and access charge (information) stored on the capacitor.
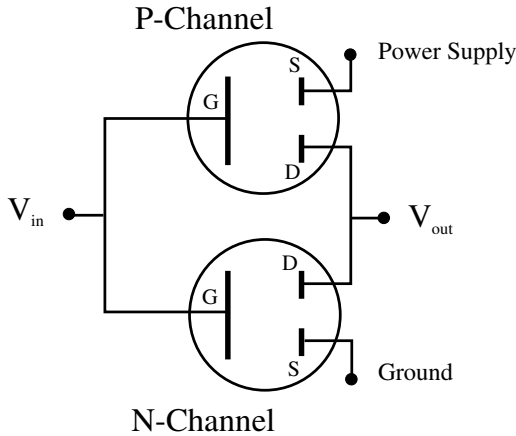
Figure 11.18 P-Channel and N-channel MOSFETs connected as a voltage inverter. The MOSFETs operate in a complimentary manner and the circuit is called a complimentary MOSFET, or CMOS.

N-channel and P-channel devices are used for microelectronics applications. The N-channel MOSFET is preferred because the mobility of electrons in Si is higher, and thus this device is faster than the corresponding P-channel MOSFET. However, a common memory device is composed of both N-channel and P-channel MOSFETs connected together so that the G's and D's are connected as shown in Figure 11.18. This configuration with both kinds of MOSFETs is called a complementary MOSFET or CMOS. Without going into specific operating conditions, this device operates by placing a potential at V_{in} and observing either an on or off state at V_{out}. With V_{in} sufficiently negative the P-channel device is on and when sufficiently positive the N-channel turns on. When one is on, the other is off. So when the P-channel is on with V_{in} negative and the N-channel is off, the output V_{out} goes toward the power supply voltage, which is usually some small positive voltage (e.g., + several volts). When the P-channel is off with a positive V_{in} but the N-channel is on, V_{out} goes to 0 (ground potential). Thus, a low V_{in} yields a high V_{out}, and a high V_{in} yields a low V_{out}. This circuit is called an inverter, and it is used as the basis for static random access memory (SRAM). SRAM does not require

refreshing like DRAM. The real advantage of CMOS is that it draws very little current, and only when the device switches from one state to another. This is crucial for high device density microelectronics chips that can be damaged from the great heat created by high current flow.

Organic Transistors. At the present time virtually all important electronic devices are made using inorganic materials as the active components. However, in these devices often an organic constituent will be used for one purpose or another. Also display devices often use a variety of organic materials for key components. There are a many good reasons for the use of organic materials to substitute for inorganic materials in electronic devices. One reason is the cost of large area devices, another is flexibility, and another is compatibility with biological systems. Thus there has been considerable activity in both research and development aimed at organic electronics. One interesting development that has led to considerable research is the ability to prepare an organic film-based device that operates like a MOSFET. This device is typically named a thin film transistor (TFT), since it does not operate in exactly the same way as a MOSFET. Figure 11.19 shows a typical TFT, but many configurations have been fabricated. A semiconductor polymer film is deposited onto a structure that has an insulator deposited onto a conducting substrate. The insulator has metal contacts deposited upon it prior to the organic film deposition. The organic semiconductor can be N- or P-type, although most stable organic semiconductors are P-type. Purely as a matter of convenience, the substrate conductor is usually a low resistivity Si wafer, and a SiO_2 film is grown on the Si wafer that serves as the gate insulator. The convenience stems from the availability of high-quality Si wafers and the ease of growth of the excellent insulator SiO_2, but any suitable insulator and conducting substrate can be used. The device operation is straightforward. A potential is applied to G in such a way to attract the majority carriers to the organic film–insulator interface. This way a highly conducting channel is created in between the conducting metal S and D contacts, thereby enabling the turn-on of I_{SD}. The essential difference in operation of the TFT as compared with the MOSFET is that the TFT uses only majority carriers. The TFT takes advantage of the low conductivity of organic semiconductors that is due to the localized state conduction rather than extended states.
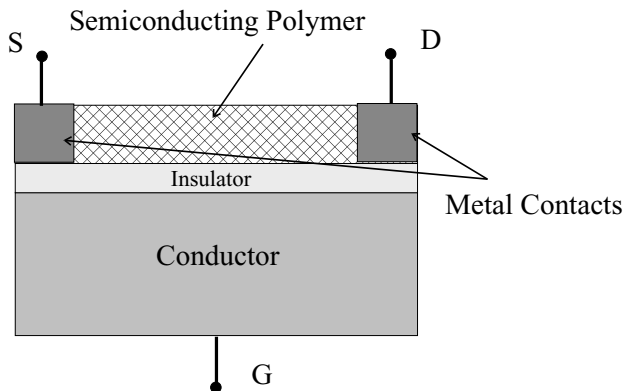
Figure 11.19 Thin film transistor (TFT) configuration using and organic semiconductor film with three terminals S, G, and D.

This low conductivity can be greatly enhanced by concentrating the carriers in a channel. Considerable improvements in TFTs will be necessary to render the use of organic materials in microelectronics.

11.4   NANOSTRUCTURES AND NANODEVICES

Size is the determining factor in the emerging science of nanostructures and nanodevices. As presently practiced in the most advanced computer chips, the smallest feature sizes are on the order of 50 nm, and the most critical thin film thickness, the gate oxide in MOSFETs is less than 5 nm. Future devices today in development will have even smaller dimensions. Therefore, in at least the field of microelectronics, nanostructures and nanodevices are part of reality. Already the advances in ultra-thin film deposition and lithography that have taken place in the field of microelectronics have enabled the creation of ultra-thin films in the nanometer range, and nanometer size structures for use in other fields. For example, nanometer size motors, pumps, and other machines are being fabricated by advanced lithography techniques. Thus, while the field of nanostructures and nanodevices has begun with electronics materials and devices, presently many other engineering objectives in medicine, optics, mechanics, sensors, and so on, are advancing into the nanotechnology arena. In keeping with the theme of this text, namely electronic materials, the nanostructure and nanodevice discussion below is limited to electronic and optoelectronic applications. However, is should be understood that virtually all modern science and engineering fields will soon have their name start with the prefix nano-.

11.4.1   Heterojunction Nanostructures

Many of the earliest and now most developed nanostructures and nanoelectronic devices are fabricated with junctions, as was the case for conventional electronic devices discussed above. The difference with the so-called nanostructures with nano-sized dimensions is that they contain a relatively small number of atoms. For a piece of material with large numbers of atoms in a periodic array, the results of Chapter 9 yielded electron energy bands that contained quasi-continuous allowed energy levels. However, for a single atom the bound electron states for a system of size l is given by equation (9.73) as follows:

$$E = \frac{\hbar^2}{2m_e} \cdot \frac{n^2 \pi^2}{l^2} \tag{9.73}$$

This 1-D formula yields discrete energy levels for the integer n and for small l. Thus for electrons confined to the nanometer thick films and in the direction normal to the film surface, the electrons will experience discrete allowed levels. This kind of structure is called a quantum well, since the discrete quantum levels result from the confinement of the electrons to l in the direction normal to the plane of the film. Specifically, a quantum well is fabricated by sandwiching an ultra-thin film (<10 nm) material that has a relatively small band gap in between layers of material with a larger gap, as shown in Figure 11.20*a* for the semiconductor GaAs sandwiched in between insulating GaAlAs layers. To maintain the highest material quality, each of the layers should be single crystalline. That means that each of the layers requires deposition in the single crystalline morphology on a different single crystal material. The deposition of a film with a single crystal morphology, and with some crystallographic relationship to the single crystalline layer
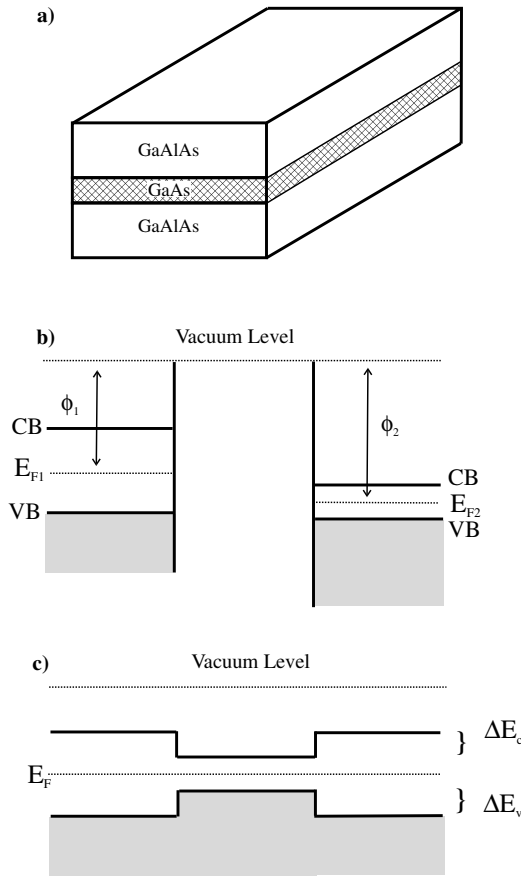
Figure 11.20   (*a*) Quantum well structure made from thin film layers; (*b*) separated wide and narrow gap semiconductors; (*c*) when the semiconductors in (*b*) are joined, a quantum well is formed with band offsets indicated ΔE_c and ΔE_v.

beneath, is called epitaxy. All the successful quantum well structures reported in the literature are epitaxial layers.

In practice, it is difficult to find materials systems that are compatible, that is, that do not react with each other, or are otherwise altered by contact, and that have lattices matching closely enough to fabricate the film sandwich structure shown in Figure 11.20*a*. Significant lattice parameter mismatch and structural dissimilarity between the narrow and wide band gap materials used for the quantum well will at best yield epitaxy with large numbers of interfacial defects, or no epitaxy at all. Recall from Chapter 7 that if materials have a different thermal expansion coefficient, a stress can develop with a change in temperature. Many thin film epitaxy processes involve temperature cycles that impose mechanical stresses on the interfaces and lead to defect formation and reduced quality. Therefore it is an enormous materials science challenge to identify suitable materials systems for obtaining the desired quantum effects and devices. Where lattice misfit

or high stresses lead to interfacial defects, the resulting quantum well devices will not be operable. One of the most important materials systems for fabricating quantum wells is GaAs as the narrow gap material ($E_g = 1.4\,eV$, $a_o = 0.565\,nm$) and Ga_xAl_yAs with a larger band gap (AlAs = 2.4 eV, $a_o = 0.566\,nm$). So the GaAlAs alloy has a gap somewhere in between 1.4 and 2.4, depending on the values for x and y. The junction resulting from these two dissimilar materials is called a heterojunction, and this term indicates different materials across the junction. Later for optical device fabrication it will be useful to use $x = 0.7$, $y = 0.3$, since this alloy has the largest GaAlAs gap (2 eV) and can effect quantum confinement while retaining a direct gap band structure that is useful for optical devices, as was mentioned in Chapter 9, Section 9.5.

Figure 11.20a shows a quantum well formed using GaAs as the narrow gap material ($E_g = 1.4\,eV$) and Ga_xAl_yAs (2 eV). Figure 11.20b shows separated semiconductors corresponding to narrow and wide gap materials. When these materials are joined as in Figure 11.20a, the result is shown in Figure 11.20c where the dissimilar band gaps lead to the band offsets, ΔE_c (barrier for electrons) and ΔE_v (barrier for holes). The offsets result when the Fermi levels equilibrate, as was discussed earlier in Chapter 11. If the original values of E_F for the materials are close, then there is little band banding, and the ideal quantum well structure depicted can be approximated. The actual energy levels are determined by the size of the well (l) and the offsets (the strength of the confinement).

The quantum well structure is fabricated from three layers in a sandwich structure. Modern film making techniques such as molecular beam epitaxy (MBE) usually use atomic beams to form elemental or compound layers on a substrate. Different layers can be produced with different atoms or the same atoms in differing proportions (alloys), and the layers can be repeated. In effect different alternating nanometer thick layers can be alternated virtually indefinitely. Such an array of repeating quantum wells is called a superlattice, and the superlattice structure can be used in many device applications. MBE systems are operated in ultra-high vacuum systems and are therefore large, complicated, expensive, and time-consuming to keep in operation. However, a variety of high-quality MBE systems are commercially available and extensively used in the scientific and engineering areas of electronic materials.

One important application of the superlattice is to enhance the performance of photodetectors. Recall Section 11.3.2.2 above and Figure 11.13 which shows that a photocell is a PN junction in which an incident photon creates electron hole pairs that are separated across the depletion width of the junction. The current flow derives from the carriers that are essentially produced by absorption of the incident light. A photodetector is the same kind of device, but it is used to sense light and measure its intensity, rather then produce a usable current or potential. Figure 11.21a shows a superlattice structure comprised of multiple quantum wells that were shown in Figure 11.20c. Figure 11.21b shows the multiple quantum well structure inserted in between a P and N semiconductor, and with an external electric field applied. From this figure it is seen that photoinduced carriers generated as a result of light incident on the device can gain sufficient energy from the applied field to overcome the barriers (ΔE_c the barrier for electrons and ΔE_v the barrier for holes) between quantum wells. The accelerated carriers can create additional carriers by impact ionization. Impact ionization is the process by which carriers are produced from the energetic collisions of already produced carriers. This creates an avalanche effect that enhances the original signal from the original photoproduced carriers. The photocell enhanced by the impact ionization via the superlattice is called an avalanche photocell (or photodiode). By judicious choice of the materials that com-
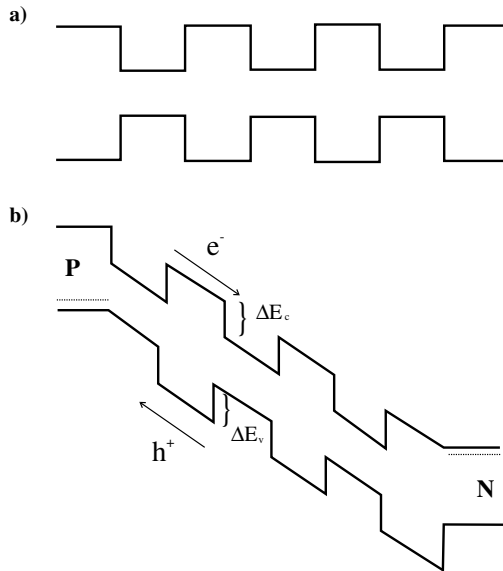
Figure 11.21   (*a*) Multiple quantum wells forming a superlattice; (*b*) the superlattice in (*a*) with an applied forward bias. The band offsets and direction for electron and hole motion are indicated.

prise the superlattice junctions, the band offset values, ΔE_c and ΔE_v, can be engineered so that one carrier amplifies via avalanche at a greater rate than the other thereby improving the signal to noise ratio of the device. Also, with the proper values of the applied field and sufficiently thin quantum (<10 nm) in the superlattice, electron tunneling can occur from well to well and such a device is called a tunneling photocell (or photodiode). The tunneling probability for holes is typically less than that of electrons so that holes accumulate or become localized relative to electrons. The overall effect is that electrons are separated from and transported in greater numbers than holes, and the gain of the device, namely the number of electrons produced as a result of the photonic process, can be enhanced with the use of the quantum well nanostructure.

The quantum well and superlattice nano structures discussed above are examples of the integration of optical and electrical devices into one structure, and they comprise one kind of optoelectronics. Solid state lasers and optical switches can also be fabricated by using similar superlattice heterojunctions that are also optoelectronic devices.

11.4.2   2-D and 3-D Nanostructures

The devices based on quantum wells and superlattices discussed above are the result of quantum effects in one direction (*l* in equation 9.73). It is also important and useful to achieve 2-D and 3-D quantum effects that lead to quantum wires and dots, respectively. Further quantization can be obtained by configuring the GaAlAs to surround GaAs so as to produce a bar of GaAs that extends in one direction, as shown in Figure 11.22*a*, and called a quantum wire or a cube of GaAs that is surrounded on all directions by the wider band gap GaAlAs, as is shown in Figure 11.22*c* and called a quantum dot. The additional quantum confinement yield quantization in the other directions and new
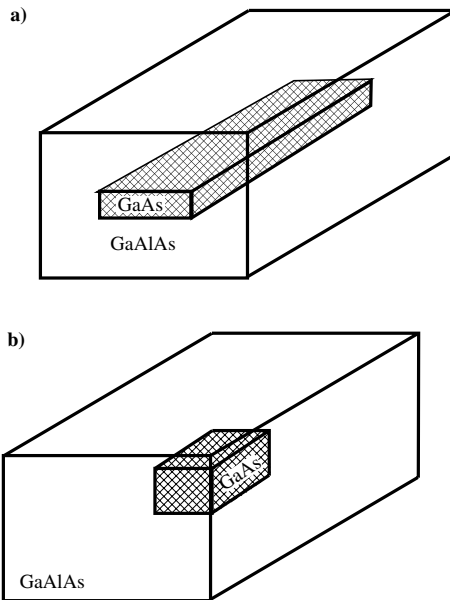
Figure 11.22 (*a*) Quantum wire resulting from 2-D confinement and (*b*) quantum dot resulting from 3-D confinement.

electronic states similar to the quantum well. Devices utilizing 2-D and 3-D quantization are now being researched. For example, arrays of quantum dots are envisioned for computer memory where charge carriers can move from one dot to another via tunneling when a suitable potential pulse is applied. Also the 3-D quantization can lead to wavelength selectivity for optical devices. While several electronics-based nanostructures and devices have been presented here, there are many other nanostructures receiving attention. For example, nano-sized motor components are being fabricated that can do various tasks such as pumping minute amounts of fluid. Such a device has potential medical applications to administer medicines in response to a nano-sized sensor that detects deficiency. The pump and sensor can be integrated with suitable nano-electronics and implanted in the patient. While this sounds like science fiction, it is rapidly becoming reality. However, with all of the emerging nanoscience and nanotechnology, significant materials challenges lie ahead in finding suitable materials and processes to produce the intricate nanostructures, and suitable measurements that operate at the nanoscale.

RELATED READING

D. A. Davies. 1978. *Waves Atoms and Solids*. Longman, London. A well-written text covering many of the topics in Chapters 9, 10, and 11 with good insights.

S. Dimitrijev. 2000. *Understanding Semiconductor Devices*. Oxford University Press, Oxford. This text is at the next level of understanding for junction devices. It is readable and well illustrated.

R. E. Hummel. 1992. *Electronic Properties of Materials*. Springer-Verlag, New York. This text provides well-written coverage of the material in Chapters 9, 10, and 11 at the appropriate level. The author has used this book as a text for the electronic materials part of the materials science course.

J. P. McKelvey. 1993. *Solid State Physics for Engineering and Materials Science*. Krieger, Higher level than Hummel, well-written, readable, and for the topics covered more complete.

M. A. Omar. 1993. *Elementary Solid State Physics*. Addison Wesley, Reading, MA. A text that covers many of the topics in Chapters 9, 10, and 11 and calso many more topics not covered in the present text. A readable text in the subject.

EXERCISES

1. Explain the sequence of events that occur when two different metals are brought into contact.

2. Repeat exercise 1 for a metal and for intrinsic N-type and P-type semiconductors.

3. Calculate the contact potential and the direction of band bending (sketch) for a junction with a metal work function of 4 eV and a semiconductor work function of 3 eV.

4. Explain why it is difficult to measure the contact potential and how it can be measured.

5. Using parallel energy band sketches, explain how to form an ohmic contact with a metal and an N- and P-type semiconductor.

6. Explain how a thermocouple and thermoelectric refrigerator works.

7. Show, using parallel band energy and *I-V* sketches, how a PN diode works with and without an applied bias.

8. Explain how a Zener diode works and what it can be used for.

9. Explain how a photocell operates and how to optimize the device.

10. Explain how bipolar and MOSFET transistors work.

11. In terms of quantum mechanical implications on electronic devices, discuss making dimensions of a solid smaller into the nm range.

12. Explain how a photocell can be enhanced by a superlattice structure.

INDEX

Note: Page numbers followed by f refer to figures, page numbers followed by t refer to tables.